Com
SIS

Computer Science and Information Systems

Vol 9, No 2, June 2012

Computer Science and Information Systems

Published by ComSIS Consortium

Volume 9, Number 2
June 2012

ComSIS is an international journal published by the ComSIS Consortium

ComSIS Consortium:

University of Belgrade:
Faculty of Organizational Science, Belgrade, Serbia
Faculty of Mathematics, Belgrade, Serbia
School of Electrical Engineering, Belgrade, Serbia
Serbian Academy of Science and Art:
Mathematical Institute, Belgrade, Serbia
Union University:
School of Computing, Belgrade, Serbia

University of Novi Sad:
Faculty of Sciences, Novi Sad, Serbia
Faculty of Technical Sciences, Novi Sad, Serbia
Faculty of Economics, Subotica, Serbia
Technical Faculty "M. Pupin", Zrenjanin, Serbia
University of Montenegro:
Faculty of Economics, Podgorica, Montenegro

ComSIS Editorial office:
University of Novi Sad, Faculty of Sciences,
Department of Mathematics and Informatics
Trg Dositeja Obradovica 4, 21000 Novi Sad, Serbia
Phone: +381 21 458 888; **Fax:** +381 21 6350 458
www.comsis.org; Email: comsis@uns.ac.rs

Volume 9, Number 2, 2012
Novi Sad

Computer Science and Information Systems

ISSN: 1820-0214

ComSIS Journal is sponsored by:

Ministry of Education and Science of Republic of Serbia - http://www.mpn.gov.rs/

Computer Science and Information Systems

AIMS AND SCOPE

Computer Science and Information Systems (ComSIS) is an international refereed journal, published in Serbia. The objective of ComSIS is to communicate important research and development results in the areas of computer science, software engineering, and information systems.

We publish original papers of lasting value covering both theoretical foundations of computer science and commercial, industrial, or educational aspects that provide new insights into design and implementation of software and information systems. ComSIS also welcomes surveys papers that contribute to the understanding of emerging and important fields of computer science. Regular columns of the journal cover reviews of newly published books, presentations of selected PhD and master theses, as well as information on forthcoming professional meetings. In addition to wide-scope regular issues, ComSIS also includes special issues covering specific topics in all areas of computer science and information systems.

ComSIS publishes invited and regular papers in English. Papers that pass a strict reviewing procedure are accepted for publishing. ComSIS is published semiannually.

Indexing Information

ComSIS is covered or selected for coverage in the following:

- Science Citation Index (also known as SciSearch) and Journal Citation Reports / Science Edition by Thomson Reuters, with 2011 two-year impact factor 0.625,
- Computer Science Bibliography, University of Trier (DBLP),
- EMBASE (Elsevier),
- Scopus (Elsevier),
- Summon (Serials Solutions),
- EBSCO bibliographic databases,
- IET bibliographic database Inspec,
- FIZ Karlsruhe bibliographic database io-port,
- Index of Information Systems Journals (Deakin University, Australia),
- Directory of Open Access Journals (DOAJ),
- Google Scholar,
- Journal Bibliometric Report of the Center for Evaluation in Education and Science (CEON/CEES) in cooperation with the National Library of Serbia, for the Serbian Ministry of Education and Science,
- Serbian Citation Index (SCIndeks),
- doiSerbia.

Information for Contributors

The Editors will be pleased to receive contributions from all parts of the world. An electronic version (MS Word or LaTeX), or three hard-copies of the manuscript written in English, intended for publication and prepared as described in "Manuscript Requirements" (which may be downloaded from http://www.comsis.org), along with a cover letter containing the corresponding author's details should be sent to official Journal e-mail.

Criteria for Acceptance

Criteria for acceptance will be appropriateness to the field of Journal, as described in the Aims and Scope, taking into account the merit of the content and presentation. The number of pages of submitted articles is limited to 30 (using the appropriate Word or LaTeX template).

Manuscripts will be refereed in the manner customary with scientific journals before being accepted for publication.

Copyright and Use Agreement

All authors are requested to sign the "Transfer of Copyright" agreement before the paper may be published. The copyright transfer covers the exclusive rights to reproduce and distribute the paper, including reprints, photographic reproductions, microform, electronic form, or any other reproductions of similar nature and translations. Authors are responsible for obtaining from the copyright holder permission to reproduce the paper or any part of it, for which copyright exists.

Computer Science and Information Systems

Volume 9, Number 2, June 2012

CONTENTS

Regular Papers

Special Section: Emerging Trends in Technology Enhanced Learning

Invited Papers

Regular Papers

EDITORIAL

Since the announcement of the two-year impact factor of our journal in the summer of last year, the number of submissions to ComSIS has increased dramatically, placing a strain on our editorial staff. In order to continue the publication of our journal in a timely and quality manner, we increased the number of editorial rejections, as well as strengthened our criteria for acceptance of articles for publication. If this trend continues, we may additionally be forced to adopt the policy of accepting only a certain portion of articles (undeserving of editorial rejection) into the reviewing process. We apologize to all prospective authors for any inconvenience this may cause, but the circumstances and our commitment to keeping and enhancing the quality standard of our journal are influencing this shift in policy.

In addition, we are happy to announce that the two-year impact factor of our journal has increased from 0.324 (2010) to 0.625 (2011).

On behalf of the Editorial Board and the ComSIS Consortium, we thank our Guest Editor, Prof. Vladan Devedžić (University of Belgrade, Serbia) for organizing the special section on emerging trends in technology-enhanced learning (TEL), featured in this issue of ComSIS. We would also like to express our gratitude to the authors and the reviewers for their high-quality contributions and effort invested into this issue of our journal.

This issue of Computer Science and Information Systems consists of fifteen regular articles and five articles in the special section on TEL.

In the first regular article, “Building XML-Driven Application Generators with Compiler Construction Tools,” Antonio Sarasa-Cabezuelo, Bryan Temprado-Battad, Daniel Rodríguez Cerezo and José-Luis Sierra describe the use of classic compiler-construction tools (i.e., parser generators), to build XML-driven application generators. Their approach consists of a document interface based on standard stream-oriented XML processing, used to build the XML scanner, characterization of the syntax of the streams by generation-specific context-free grammars, and augmentation of these grammars with suitable semantic attributes and semantic actions.

The next paper by Dragan Milićev, “Towards Understanding of Classes versus Data Types in Conceptual Modeling and UML,” provides an in-depth study of the ambiguities and discrepancies when making the distinction between two kinds of entity types – classes and data types – in traditional conceptual modeling and UML. A novel semantic interpretation is proposed for consolidation, based on the premise that populations of the two kinds of entity types are defined in two ways: by intensional (for data types) and extensional (for classes) definitions. The proposed interpretation also lends itself to the description of several semantic consequences: value-based vs. object-based semantics, associations vs. attributes, and identity vs. identification.

“Specification of Data Schema Mappings using Weaving Models”, by Nenad Aničić, Siniša Nešković, Milica Vučković and Radovan Cvetković, tackles the problem of unsuitability of weaving models for the specification of heterogeneous schema mappings in model driven engineering. The lack of mapping rules, which allows specifications that are semantically meaningless, wrong, and/or disallowed, is overcome by providing the explicit support for semantic mapping rules, based on the introduction of weaving metamodels augmented with constraints.

Vıctor M. Prieto, Manuel Alvarez, Rafael Lopez-Garcıa and Fidel Cacheda, in “A Scale for Crawler Effectiveness on the Client-Side Hidden Web,” present an evaluation framework for assessing the success of crawlers in traversing the client side of the “hidden Web.” The authors formulate a scale by grouping basic scenarios in terms of several common features, and propose evaluation methods of crawler effectiveness in terms of the scale’s levels. Through extensive evaluation of open-source and commercial crawlers, the paper highlights the ones that are most capable of handling the evaluation scenarios.

In “Wavelet trees: a survey,” Christos Makris gives an overview of various characteristics of the wavelet tree data structure, which is used for many tasks, such as text compression, text indexing, and retrieval. The article considers issues concerning the efficient maintenance of the structure, and its handling in various applications.

The article “Impact of Personnel Factors on the Recovery of Delayed Software Projects: A System Dynamics Approach,” by Mostafa Farshchi, Yusmadi Yah Jusoh and Masrah Azrifah Azmi Murad, explores the possibility for significant schedule improvement by adding new staff during delayed realization of a software project, considering new manpower’s capabilities, skills and experience. The study is conducted through formulation and evaluation of a system dynamics model which simulates the project’s progress when new members are added.

“A New Strategic Tool for Internal Audit of the Company Based on Fuzzy Logic,” by Aleksandar Pešić, Duška Pešić and Andreja Tepavčević, presents strategic management tool for the assessment of internal organizational factors. The tool addresses the limitations of traditional appraisal methods, enabling more comprehensive evaluation of a company’s internal environment. The authors propose an original method – fuzzy synthesis of internal factors (FSIF) – which extends the approach of the classical internal factor evaluation (IFE) matrix approach by incorporating fuzzy logic in order to better describe real situation.

In the paper “Experimental Investigation of the Quality and Productivity of Software Factories Based Development,” Andrej Krajnc et al. evaluate the quality and productivity benefits of the software factories (SF) approach to software development. The traditional and the SF approach were compared using the Goal – Question – Metric (GQM) methodology, on participants grouped into 32 teams, demonstrating the superiority of the SF method.

“Nearest Neighbor Voting in High Dimensional Data: Learning from Past Occurrences,” by Nenad Tomašev and Dunja Mladenić, presents a new approach for exploiting the hubness phenomenon, which is an aspect of the “curse of dimensionality” inherent to nearest-neighbor (NN) methods, in *k*-NN classification. The article proposes a novel algorithm, hubness information *k*-nearest neighbor (HIKNN), which introduces the informativeness of a point’s hubness into the *k*-NN voting framework.

In “AKNOBAS: A Knowledge-based Segmentation Recommender System based on Intelligent Data Mining Techniques,” Alejandro Rodríguez-González et al. describe design and evaluation of the recommender system AKNOBAS, which employs clustering algorithms for the tasks of marshalling information and recommending content to users.

The paper “Scope of MPI/OpenMP/CUDA Parallelization of Harmonic Coupled Finite Strip Method Applied on Large Displacement Stability Analysis of Prismatic Shell Structures,” by M. Hajduković et al., studies the effectiveness of parallelization in large displacement stability analysis of orthotropic prismatic shells. The computational overhead introduced by the semi-analytical harmonic coupled finite strip method (HCFSM) used to solve the large deflection and the post-buckling problems is tackled by application of MPI, OpenMP and CUDA approaches to parallelization.

Yordan Kalmukov, in the article “Describing Papers and Reviewers’ Competences by Taxonomy of Keywords,” deals with the problem of automatically assigning reviewers to articles by introducing a taxonomy of keywords to describe the competencies of both the articles and reviewers. The hierarchical nature of the taxonomy allows the expression of semantic proximity in addition to exact matches. A modification of Dice coefficient is proposed to achieve the given task.

“Journal Evaluation based on Bibliometric Indicators and the CERIF Data Model,” by Dragan Ivanović, Dušan Surla and Miloš Racković, proposes an application of extended CERIF data model for storing journal impact factors and journal scientific fields, and based on this data presents a journal evaluation approach. The method does not unambiguously evaluate a journal, but rather suggests possibly journal appropriate categories according to the values of the employed metric, but final decision is made by a committee.

In the paper “Ontology-Based Home Service Model,” Moji Wei, Jianliang Xu, Hongyan Yun and Linlin Xu study home service retrieval and invocation for a smart home. An ontology-based service model is employed to retrieve and invoke services according to user’s needs automatically. Two domain ontologies are constructed – function concept ontology and context concept ontology – to annotate the semantics of the home service different facets. Two scenarios involving different types of services are presented that demonstrate the usage of the model.

Finally, in “Automatic Generation of E-Courses Based on Explicit Representation of Instructional Design,” Goran Savić, Milan Segedinac and Zora Konjović present a system for automatic generation of IMS LD compliant e-courses from three components: machine readable explicit representation of instructional design, ontology of learning goals, and IMS content packaging compliant learning resources.

Editor-in-Chief
Mirjana Ivanović

Managing Editor
Miloš Radovanović

GUEST EDITORIAL

Technology Enhanced Learning (TEL) has been around for quite some time now, and is still thriving. It has the goal of "providing socio-technical innovations (also improving efficiency and cost effectiveness) for learning practices, regarding individuals and organizations, independent of time, place and pace. The field of TEL therefore focuses on the support of any learning activity through technology".[1]

Note also that the term TEL is often used synonymously with E-Learning even though there are significant differences. The main difference between the two expressions is that TEL focuses on the technological support of any pedagogical approach that utilizes technology. However, this is rarely presented as including print technology or the developments around libraries, books and journals in the centuries before computers.[2]

Still, TEL has already spread out to include a vast number of other aspects. Not long ago, the major issues in TEL were learning content, instructional design, personalization, recommendation, adaptation and adaptivity, interoperability, and natural language processing. Nowadays, all major TEL conferences and journals include dozens of new and emerging topics, including social processes associated with learning, learners' motivation, learners' and teachers' trust and reputation, and many more.

Thus, when issuing the call for papers for this special section we have tried to identify important new trends in the area of TEL and to represent such trends with selected articles. We have consulted the topics suggested by top-level TEL conferences, high-reputation TEL research journals, and several other Internet resources. Submissions of extended versions of papers previously published in conferences and workshops were also welcome, given that they were substantially expanded and improved. To this end, the call has proposed submissions including (but not limited to) the following topic areas:

- Personalization, learner modeling and adaptation
- Context-aware learning systems
- Social Semantic Web and learning

[1] Wikipedia, "Technology-Enhanced Learning". [Online]. Available: http://en.wikipedia.org/wiki/Technology-Enhanced_Learning (Last visited: May, 2012)

[2] Ibid.

- Mobile technologies for learning
- Network infrastructures and architectures for TEL
- Ubiquitous learning
- Data mining and information retrieval
- Recommender systems for TEL
- Learning analytics
- Problem- and project-based learning / Inquiry based learning
- Computer-supported collaborative learning
- Collaborative knowledge building
- Game-based and simulation-based learning
- Story-telling and reflection-based learning
- Instructional design and design approaches
- Communities of learners and communities of practice
- Teaching techniques and strategies for online learning
- Learner motivation and engagement
- Evaluation methods for TEL
- Self-regulated and self-directed learning
- Reflective learning
- Social processes in teams and communities
- Knowledge management and organizational learning
- Sustainability and TEL business models and cases
- Business-learning models
- Trust and reputation in TEL
- Applications of TEL in various domains
- Practical experiences of TEL
- Results of TEL research projects applied in practice
- Workplace learning in small, medium and large companies
- Aggregated learning at the workplace
- Distance and online learning
- Lifelong learning (cradle to grave)
- Vocational training
- Informal learning
- Accessible learning for all
- Psycho-pedagogic support for users
- Educational guidance for tutors
- Adapted learning flow, content and monitoring process

Out of all the submissions we have received, we have selected five high-quality papers containing original research results on some of the above and related topics. Out of these five, two are invited papers, and the other three are regular papers.

The invited paper "Recommending Collaboratively Generated Knowledge", by W. Chen and R. Persen, addresses two of the above listed topics: recommender systems and collaborative knowledge building. It recognizes the fact that learners (and not only teachers, tutors, and other traditional content authors) have become active *producers* of knowledge, not only consumers – by actively conducting their learning processes, learners generate a vast amount of content. This kind of content can be a valuable learning resource in both studying and assessment processes. However, it also gives rise to new challenges for indexing, sharing, retrieval and recommendation of such learning content. The paper presents a newly developed recommender system for such emerging learning content generated through collaborative knowledge building processes and studies the implications and added values of the recommendations.

Another invited paper is "A Model-based Approach for Assessment and Motivation", by J.M. Spector and C.M. Kim. It pertains to representations as a foundational aspect of learning and instruction. The paper briefly reviews the research literature about cognition and processing internal mental models, putting the emphasis is on the role that mental models play in critical reasoning and problem solving. Then it presents a theoretically-grounded rationale for taking internal mental representations into account when designing and implementing support for learning. The emphasis here is on interaction with meaningful problems. The authors have developed and presented a conceptual framework for integrating models into learning environments that includes technologies for formative assessment and motivation.

D.G. Tremblay and V. Psyché, the authors of the paper "Analysis of processes of cooperation and knowledge sharing in a community of practice with a diversity of actors", focus on communities of practice as an important sub-topic of organizational and workplace learning. Communities of practice often stem from a voluntary initiative within an organization, whose members share some knowledge or expertise they wish to improve. However, a community of practice can take the form of a network that brings together members with common interests coming from different organizations and even several countries in which they perform different types of work. To this end, the paper discusses the evolution of communities of practice supported by the use of Web 2.0 tools that promote collaboration.

The paper "Web Service Support for Collaboration between Demographers", by M. Devedzic and her co-authors, runs along a similar avenue, focusing on collaboration among demographers working in different organizations. In fact, the paper promotes the use of novel Web services in the daily work and research of demographers and illustrates how such services can facilitate learning and collaboration between a university research group in the field of demography and professionals in the field of demographic statistics. Still, since the technology used is generic, it can be easily instantiated for use by other social science researchers. The Web services used in this collaboration are developed as part of an EU research project.

Lu Xiao, the author of the paper entitled "Exploring the Use of Contextual Modules for Understanding and Supporting Collaborative Learning Activities: An Empirical Study", reports her experiences with three student groups that have collaborated in a semester-long classroom project. The project included both tasks that required students to complete in virtual group workspace and activities that could be carried out in physical world environment. The author has noticed different collaboration patterns among the groups with respect to building and maintaining social relationships, submitting individual's work to the group, and scheduling group meetings. She has then used Bereiter's two contextual modules, intentional learning and schoolwork modules, to interpret the observed patterns. The interpretation suggests that the group leader's contextual module plays a significant role in all members' group learning experiences and outcomes. Based on this, the author discusses design implications intended for encouraging learning-based (as opposed to work-based) practices in virtual group environment.

Guest Editor

Vladan Devedzic
University of Belgrade, Serbia
http://devedzic.fon.rs/
devedzic@fon.rs

DOI: 10.2298/CSIS110505002S

Building XML-Driven Application Generators with Compiler Construction Tools

Antonio Sarasa-Cabezuelo[1], Bryan Temprado-Battad[1], Daniel Rodríguez-Cerezo[1], José-Luis Sierra[1]

[1] Computer Science School,
Complutense University of Madrid
Calle Profesor José García Santesmases, s/n
28040 Madrid, Spain
{asarasa, bryan, drcerezo, jlsierra}@fdi.ucm.es

Abstract. This paper describes how to use conventional compiler construction tools, and parser generators in particular, to build XML-driven application generators. In our approach, the document interface is provided by a standard stream-oriented XML processing framework (e.g., SAX or StAX). This framework is used to program a generic, customizable XML scanner that transforms documents into streams of suitable tokens (opening and closing tags, character data, etc.). The next step is to characterize the syntactic structure of these streams in terms of generation-specific context-free grammars. By adding suitable semantic attributes and semantic actions to these grammars, developers obtain generation-oriented translation schemes: high-level specifications of the generation tasks. These specifications are then turned into working application generators by using standard parser generation technology. We illustrate the approach with <e-Subway>, an XML-driven generator of shortest-route search applications in subway networks.

Keywords: Application Generators, Compiler Construction Tools, XML Processing, Software Development Approach

1. Introduction

Application generators and generative approaches to software development are keystone technologies in enhancing productivity and ensuring the quality of final software artifacts [5][7][9]. In application generators, XML is frequently chosen as a basic encoding format for input specifications [6]. Thus, having cost-effective and efficient methods for processing XML documents is mandatory in these scenarios. For this purpose, architects of application generators have a wide range of XML-processing technologies available, ranging from task-specific (e.g., XSLT) to general-purpose ones (e.g., SAX or DOM). General-purpose XML processing frameworks (i.e., SAX, DOM, StAX, etc) [15] are particularly relevant for very specific or complex processing tasks not easily accomplished with pre-existing task-specific technology.

However, general-purpose processing frameworks are largely data-centric: they see XML documents as chunks of data. By contrast, the intrinsic nature of descriptive markup and XML is fundamentally language-oriented: to design an XML format for a particular type of document is equivalent to devising a suitable domain-specific markup language. It immediately raises an obvious question: if XML documents are structured with (formal) markup languages, why not use conventional language-processing techniques to support the processing of these documents?

The answer to this question depends on the complexity of the markup language and the processing tasks. For simple XML documents (e.g., a sequence of logs with a description and a timestamp) and simple processing tasks (e.g., producing an HTML table with the logs), the effort of designing and implementing the processing component as if it were a sort of compiler, using methods and techniques specific to the compiler construction field, may be excessive. However, for more complex documents (e.g., QTI documents describing assessments in an e-Learning system [11]) and more complex processing tasks (e.g., configuring assessment systems with the QTI documents), this effort can pay off. Actually, the latter constitute the kind of scenarios faced by developers of application generators.

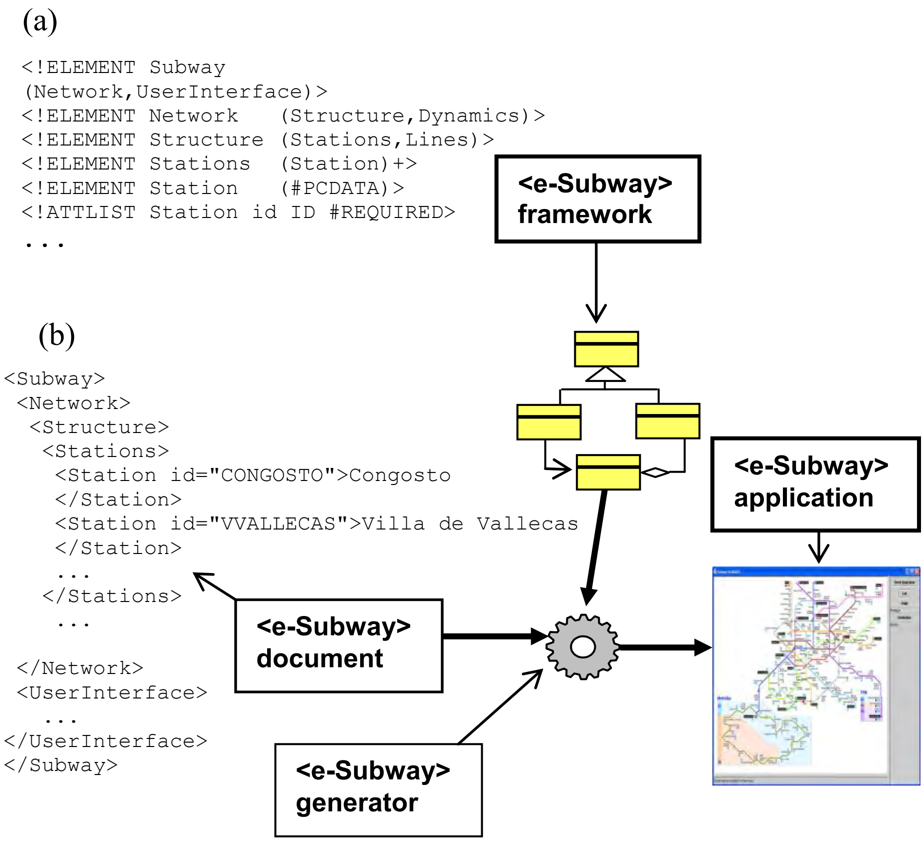
An attractive feature of the language-oriented approach is that the design and implementation of language processors (and, in particular, of translators) is mature enough to support a wide range of tools able to produce reliable and efficient implementations from high-level specifications. Of those tools, the most widely known are parser generators (i.e., YACC-like tools) [1]. These tools accept translation schemes, i.e., context-free grammars annotated with the semantic actions that actually perform the processing, as input, and produce working translators as output. Thus, by using one of these tools, it is possible to drastically reduce the development effort compared to a handcrafted implementation.

This paper shows how it is possible to build sophisticated XML processing environments by combining parser generators with general-purpose stream-oriented XML processing frameworks. For this purpose, it develops a general method that can be used with a great variety of parser generation environments or underlying XML processing frameworks. The result is a systematic approach to the language-oriented development of complex syntax-directed XML processing components, which is especially well-suited to the development of XML-driven application generators.

The rest of the paper is organized as follows: section 2 introduces <e-Subway>, the system that will be used for illustrative purposes. Section 3 outlines the approach and illustrates it with <e-Subway>. Section 4 presents some work related to ours. Finally, section 5 presents some conclusions and lines of future work.

2. Case study

The system <e-Subway> is an XML-based system for the construction of shortest-route search applications in subway networks. This system was already used as a case study in some of our previous experiences concerning the generation of applications from structured documents [38][39]. <e-Subway> integrates:



(a)

```
<!ELEMENT Subway
(Network,UserInterface)>
<!ELEMENT Network   (Structure,Dynamics)>
<!ELEMENT Structure (Stations,Lines)>
<!ELEMENT Stations  (Station)+>
<!ELEMENT Station   (#PCDATA)>
<!ATTLIST Station id ID #REQUIRED>
...
```

(b)

```
<Subway>
 <Network>
  <Structure>
   <Stations>
    <Station id="CONGOSTO">Congosto
    </Station>
    <Station id="VVALLECAS">Villa de Vallecas
    </Station>
    ...
   </Stations>
    ...

 </Network>
 <UserInterface>
   ...
</UserInterface>
</Subway>
```

Figure 1. (a) Excerpt of the <e-Subway> DTD; (b) The <e-Subway> generation process

- An XML-compliant markup language for structuring documents that describe the different aspects of route searching applications (e.g., stations, lines, connections and other aspects of the subway network, as well as selected aspects of the final application's user interface). In Fig. 1a, we outline a fragment of the DTD for this language.
- A domain-specific object-oriented framework. Applications in <e-Subway> are instantiations of this framework.

- A *generator*. This component processes documents that describe <e-Subway> applications and produces the documented applications as instantiations of the <e-Subway> framework (Fig. 1b) (i.e., it does not actually generate code, but produces in-memory instances –objects– of the <e-Subway> framework's classes, and establishes appropriate links between these instances).

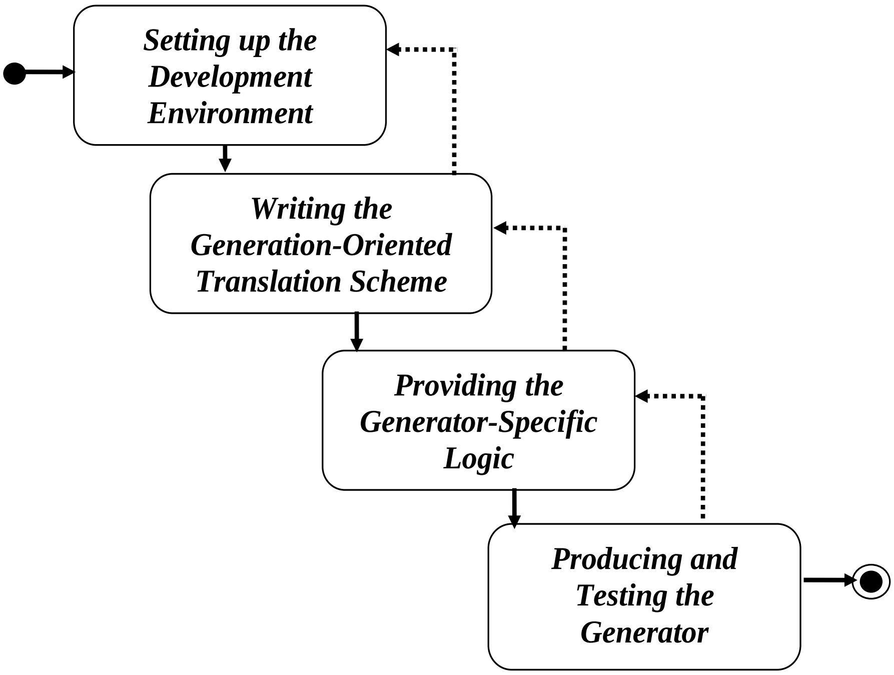

Setting up the Development Environment
Writing the Generation-Oriented Translation Scheme
Providing the Generator-Specific Logic
Producing and Testing the Generator

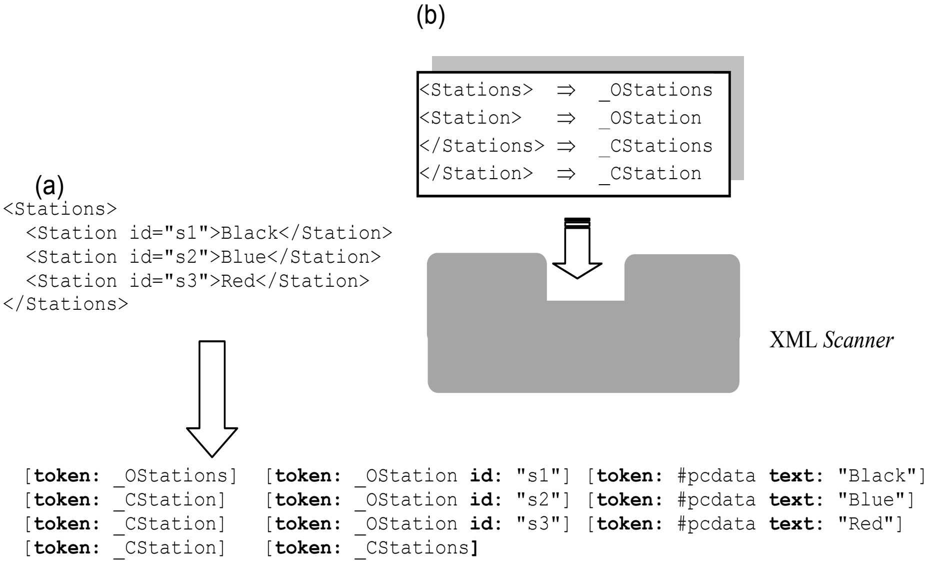
Figure 2. Activities and sequencing of activities in the development approach

3. The development approach

Fig. 2 summarizes the approach to developing XML-driven application generators with conventional parser generation tools, focusing on the main activities and on the sequencing of these activities (the backwards transitions allow an iterative/incremental production process). Notice that this workflow largely mirrors that which is usually followed by any compiler developer. Indeed, he/she must provide a suitable grammar for the source language, add semantic actions to this grammar to yield a translation scheme, generate the translator either by hand or by using a suitable generation tool, etc. This parallelism makes the language-oriented nature of the proposal described in this paper apparent. Nevertheless, it is important to notice that the goal is not to provide a full translator from scratch, but instead to put an additional language processing layer on top of an existing stream-oriented XML processing framework. In particular, the processor will operate on XML information elements (e.g., represented in the form of SAX events) instead of individual characters. As a result, it will lead to the organizing of the application-specific logic attached to a general processing framework into two well-differentiated tiers: one that operates as a syntax-directed translator, and another that provides services to this translator. The following subsections analyze each activity in this workflow.

3.1. Setting up the development environment

This activity integrates a parser generation tool with a general-purpose XML processing framework. This activity will be performed only sporadically, since the same development environment can be used in the development of many different application generators.



```
<Stations>   ⇒  _OStations
<Station>    ⇒  _OStation
</Stations>  ⇒  _CStations
</Station>   ⇒  _CStation
```

```
<Stations>
  <Station id="s1">Black</Station>
  <Station id="s2">Blue</Station>
  <Station id="s3">Red</Station>
</Stations>
```

```
[token: _OStations]  [token: _OStation id: "s1"] [token: #pcdata text: "Black"]
[token: _CStation]   [token: _OStation id: "s2"] [token: #pcdata text: "Blue"]
[token: _CStation]   [token: _OStation id: "s3"] [token: #pcdata text: "Red"]
[token: _CStation]   [token: _CStations]
```

Figure 3. (a) Example of tokenization; (b) Customization of an XML Scanner

As previously stated, a parser generation tool produces translators for formal languages from high-level specifications. These translators are driven by parsers that operate on streams of tokens provided by lexical analyzers. Thus, the key idea behind integration is to see XML documents as streams of tokens. Integration itself is focused on the logical structure: streams of tokens are produced by remapping the data structures provided by the general-purposes processing frameworks, instead of by directly operating on the actual XML files. The integration distinguishes four different lexical categories, or kinds of tokens:

- *Character data tokens*, which correspond to fragments of textual content in the processed documents.
- *Opening* and *closing* tags.
- The *end of document*.

In addition to its lexical category, each token can include additional lexical information in the form of lexical attributes:

- *Character data tokens* have the actual textual content associated with them.
- *Opening* tags have the element attributes specified in the tag, as well as namespace information, associated with them.

Fig. 3a shows an example of tokenization.

Based on these considerations, integration provides a generic and customizable XML Scanner by using the selected XML processing framework. This component can be generic, since it is only needed to indicate how to map opening and closing tags into lexical categories (e.g., by using a table, as suggested in Fig. 3b). Also, this kind of integration can be successfully carried out by using a stream-oriented framework such as SAX or StAX. Indeed, the action of the XML Scanner can be conceived of as the transformation of a stream of documental information items into a stream of tokens, as expected by the generated translators.

Concerning the technical details, since generated translators are push components (i.e., they take control, requesting tokens from the scanners when required), integration is particularly straightforward with a pull XML processing framework (e.g., StAX), since these frameworks provide each next information item on demand. On the other hand, integration with a push framework (e.g., SAX) requires inverting control (e.g., using a producer-consumer multithreaded solution). In our previous papers [33] and [34], we give examples of the two kinds of integration.

Finally, it is important to highlight the difference between the XML Scanner proposed in this section and the scanner of a conventional language processor. Indeed, the XML Scanner proposed in our approach is built on top of a full-flagged stream-oriented XML processing framework, able to support features that are common to any XML-based markup language (e.g., support for different character sets and encodings, comment recognition, entity and namespace management, etc.). On the other hand, the scanner of a conventional language processor usually works on text files or stream of characters. Therefore, although it could be possible to provide a conventional scanner for tokenizing a particular type of XML documents, it would have to deal with the aforementioned features to be fully XML-compliant. The complexity of exiting XML parsers teaches us that it is not exactly an easy task. It makes the difference between our proposal and the conventional development of a language processor apparent: if we develop a language processor for a particular type of XML documents following the standard patterns explained in any university-level compiler construction course (see for instance [1]), we will probably get a program able to process input text files with an XML syntax-like, but not a program able to deal with the features common to all XML applications (e.g., the ability to split a huge XML document in several files and to assemble these fragments using the XML entity mechanism, to deal with different character sets, to deal with namespaces, etc.).

3.2. Writing the Generation-Oriented Translation Scheme

This is the central activity of our development approach. Its purposes are to:

– Write a suitable *generation-specific* grammar that gives structure to the stream of tokens provided by the XML Scanner.
– Annotate this grammar with code (*semantic actions*) to describe the generation task. The result is the syntax-directed, *generation-oriented translation scheme* produced by this activity.

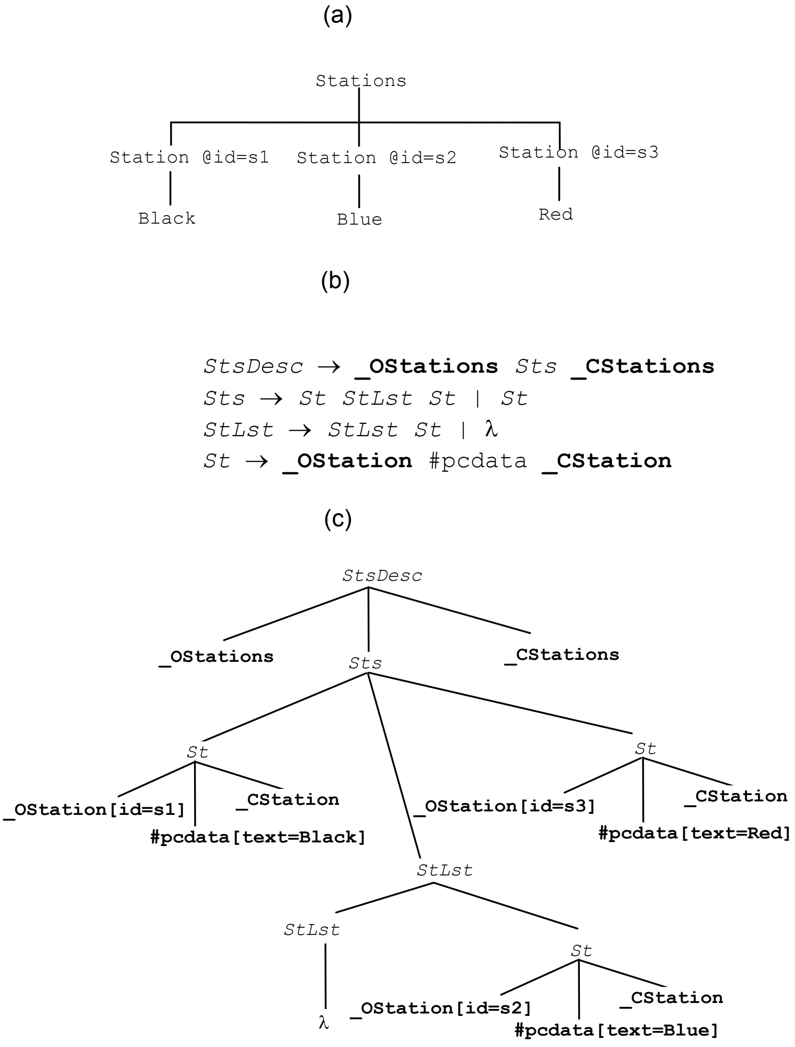
It is important not to confuse the generation-specific grammar with the document grammar (e.g., a DTD or an XML Schema) used to describe the markup language. The generation-specific grammar of this activity addresses a key aspect of the processing: to give a suitable structure to the stream of tokens in order to facilitate application generation. Indeed, this aspect must be addressed by any general-purpose XML processing solution. For instance, it is implicit in the code that deals with the children of an element node in a DOM-based processing application, in the callback methods and the state variables of a SAX event handler, or in the set of mutually recursive procedures of a StAX-based application. The main difference (and advantage) of our approach is that this structuring aspect is explicitly described at a very high abstraction level, as a context-free grammar, instead of being hand-coded in a final implementation. The structure imposed on a stream of tokens by a generation-specific grammar takes the form of a parse tree. Fig. 4b shows an example. As this example makes apparent, the parse tree is finer-grained than the usual document tree, where the element contents lack any structure outside a uniform sequence of nodes (compare Fig 4a with Fig 4c).

The conceptual processing model behind a generation-oriented translation scheme is to perform a traversal of the parse tree, executing semantic actions at significant points in this traversal. In addition, semantic actions can store and consult information in the nodes of the parse tree (typically this information is organized as an assignment of values to semantic attributes), as well as in global variables.

The exact nature of the traversal is determined by the kind of translators generated by the parser generation tool:

– *Top-down translators*, such as those generated by JavaCC and ANTLR, traverse the parse tree in *preorder* (i.e., the translator visits each node before visiting its children). The significant points are, for each node, when: (i) the translator enters the node, (ii) the translator enters a child, (iii) the translator has left a child, and (iv) the translator exits the node.

- *Bottom-up translators*, such as those generated by YACC-like tools (e.g., CUP), traverse the parse tree in *postorder* (i.e., for each node, the translator first visits the node's children and then the node itself). There is a significant point each time the translator exits a node.



(b)

```
StsDesc → _OStations Sts _CStations
Sts → St StLst St | St
StLst → StLst St | λ
St → _OStation #pcdata _CStation
```

Figure 4. (a) Document tree for the document in Fig. 3a; (b) (Part of) a generation-specific grammar; (c) Parse tree for the document in Fig. 3a according to this grammar.

It is worthwhile to note that, while it is useful to have this model in mind when writing generation-oriented translation schemes, it is only a conceptual model. In practice, the parse tree is never built, the traversal is implicitly performed during parsing, and the semantic actions are executed in a suitable order. Also, the semantic attributes are only available as parameters of recursive procedures (e.g., in recursive descent translators generated by JavaCC or ANTLR) or stored in the records of a semantic stack (e.g., in YACC-generated bottom-up translators). This behavior is a fundamental feature when dealing with huge documents (like those required by the generation of data-intensive applications) or with documents made available asynchronously in an XML stream (as required by on-line generators, which incrementally generate applications as they process their descriptions). It also constrains the kind of specifications that can be done. For instance, top-down translators do not work with *left-recursive* grammars, which are useful for characterizing left-associative structures. Also, although bottom-up translators are able to deal with left-recursion in a very efficient way, it is substantially more difficult to deal with *inherited* information (i.e., information that flows from parent to child or from sibling to sibling) than in top-down translators [1].

```
...
StsDesc → _OStations Sts _CStations {
   $$.stations = $2.stations
}
Sts → St StLst St {
   ops.addFirstStation($2.stations,$1.id,$1.name);
   ops.addLastStation($2.stations,$3.id,$3.name);
   $$.stations = $2.stations;
}
Sts → St {
   $$.stations = ops.makeStList();
   ops.addFirstStation($$.stations,$1.id,$1.name);
   ops.addLastStation($$.stations,$1.id,$1.name);
}
StLst → StLst St {
   $$.stations = addStation($1.stations,$2.id,$2.name);
}
StLst → λ {
   $$.stations = ops.makeStList();
}
St → _OStation #pcdata _CStation {
   $$.id = $1.id;
   $$.name = $2.text;
}
...
```

Figure 5. Excerpt of a translation scheme for a fragment of the <e-Subway> markup language

Knowing the traversal carried out by the translator makes it possible to place the semantic actions in the syntax rules of the generation-specific grammar. The specification formalism must also provide a way of referring to the semantic attributes (e.g., placing them as parameters of the syntax symbols, as in JavaCC, or using pseudovariables, as in YACC-like tools). Fig. 5 depicts a fragment of the syntax-directed translation scheme for a bottom-up translation model of the <e-Subway> generator using a YACC-like notation (in particular, it uses YACC-like pseudovariables: $$ to refer to the semantic record of a rule’s head, $i to refer to the semantic record of the i-esime body’s symbol). The translation scheme builds an in-memory representation of the stations in a line, following the typical generation pattern of populating a suitable semantic model [9].

Finally, it is interesting to remark that, for the sake of generalization, we have kept our approach simple enough to fit in the different parser generation tools available. For this reason, more advanced capabilities have been explicitly omitted, although they might facilitate some advanced processing tasks. For instance, one of these advanced capabilities could be the interplay between syntax and semantics, supported by tools like ANTLR [30], and which, for instance, would allow us to make parsing dependent on predicates concerning certain semantic attributes. Still, some clever behavior can be achieved without introducing these advanced features by setting the XML Scanner to produce different tokens for different occurrences of the same element type, depending of the values of some of their XML attributes.

3.3. Providing Generator-Specific Logic

The semantic actions in the translation scheme will typically use other, more conventional machinery that must also be provided to produce a fully functional application generator. This machinery constitutes the so-called *generator-specific logic*.

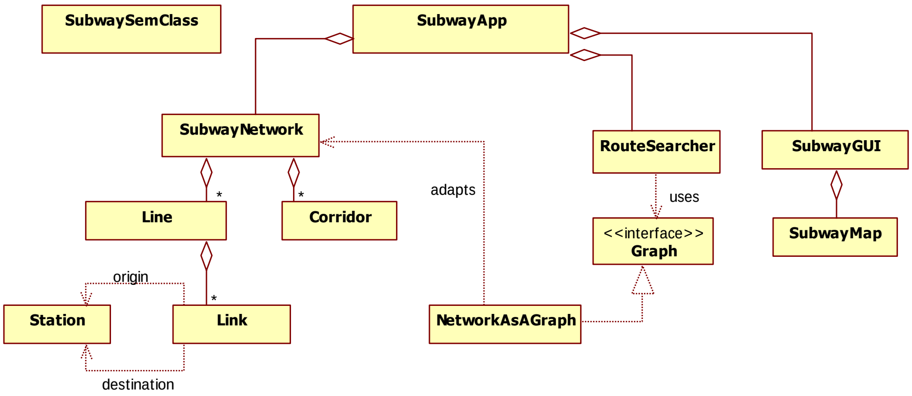


Figure 6. Main components of the <e-Subway> framework

For instance, in <e-Subway>, this generator-specific logic is formed by the <e-Subway> *framework*, which constitutes the aforementioned *semantic model* in this scenario [9]. Thus, and as indicated in section 2, the resulting generator does not generate actual code, but instantiates classes in the

<e-Subway> *framework* and links the resulting objects in appropriate ways (using the terminology introduced in [9], it *populates* the <e-Subway> *framework*, as we indicate below). Fig. 6 depicts the main components of the <e-Subway> framework.

In this way, the approach promotes a clear separation between the language-oriented processing of the XML documents and the conventional software that supports this processing. This separation can be further emphasized by providing a suitable façade for the generator-specific logic, with operations that will be invoked from the translation scheme (it indeed follows the embedment helper pattern described in [9]). The `ops` global variable in Fig. 5 illustrates this practice (in the actual <e-Subway> generator, the variable refers to an instance of such a façade, which is represented by `SubwaySemClass` in Fig. 6).

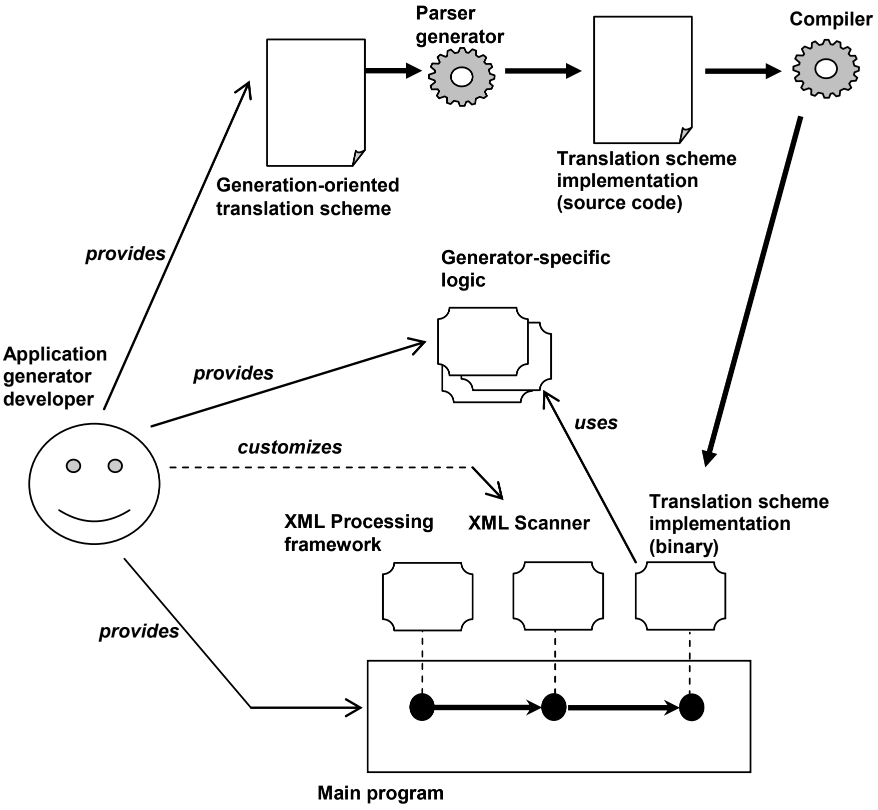

Parser generator
Compiler
Generation-oriented translation scheme
Translation scheme implementation (source code)
provides
Generator-specific logic
Application generator developer
provides
uses
customizes
Translation scheme implementation (binary)
XML Processing framework
XML Scanner
provides
Main program

Figure 7. The production process of XML-driven application generators in the development approach

3.4. Producing and Testing the Generator

Once the translation scheme and the application-specific logic are available, it is possible to get the working generator automatically by using the parser generation tool. The production process is detailed in Fig. 7. Indeed:

- The translation scheme is used as input to the parser generation tool in order to obtain the implementation of a translator written in the target language of the parser generation tool (e.g., Java for JavaCC or CUP). Notice that this way, the parser generation tool becomes a kind of meta-generator [6] in our proposal.
- In turn, this implementation can be turned onto a working binary component by using a compiler for such a target language (e.g., a Java compiler, assuming JavaCC or CUP was used).
- The customized XML Scanner must also be provided. Usually it can involve writing the mapping table (see section 3.1) using a customization file, or directly writing this table in the target programming language (e.g., Java).
- Finally, the developer must provide a small main program gluing all this together. This program will properly connect all the components required to constitute the generation pipeline. This pipeline will be made of: (i) a standard XML processing framework able to turn XML documents into information elements (e.g., represented by SAX events) suitable for the XML Scanner, (ii) the customized XML Scanner used to turn these elements into tokens accepted by the translator generated, and (iii) the translator itself, which makes use of the generator-specific logic.

The resulting generator can be tested in order to resolve possible defects and/or malfunctions. This activity therefore completes the development process.

4. Related Work

In this section we compare our work to conventional XML processing approaches (subsection 4.1), to other approaches to language-driven XML processing (subsection 4.2), and to approaches to XML processing based on attribute grammars (subsection 4.3)

4.1. Conventional XML processing approaches

As indicated in section 1, conventional approaches to XML processing range from task-specific ones (e.g., XSLT [41] for document transformation or XQuery [43] for expressing queries to XML structured documents) to general-purpose frameworks (e.g., tree-oriented ones, like DOM [19], or stream-oriented ones, like SAX [3][21], StAX [21] or XML-Pull [21]; see also [15] for a survey of this kind of general-purpose XML processing frameworks). While both types of these traditional approaches (task-specific and general purpose ones) share a data-centric orientation (i.e., these approaches see XML documents as chunks of data instead of sentences in a formal language), task-specific approaches tend to be of a higher level and of a more declarative nature than general-purpose ones. Indeed, task-specific approaches promote the use of domain-specific languages specifically tailored to the task at hand (e.g., transformation specifications in XSLT, FLWOR expressions in XQuery), while general-purpose processing frameworks are usually expressed in a general-purpose programming language (e.g., Java) and their use demands programming skills in this kind of general-purpose programming languages. As a consequence, task-specific approaches are usually more usable than general-purpose ones. However, the applicability of task-specific approaches is reduced to concrete processing tasks; for other tasks, either another task-specific approach or a general-purpose one will need to be used.

The language-oriented approach presented in this paper tries to bring together the best of the two aforementioned XML processing worlds (task-specific and general-purpose ones). Indeed, it clearly splits the processing task into two well-differentiated layers: (i) a *linguistic* layer, explicitly governed by an underlying formal grammar, which deals with the syntax-directed processing of the stream of basic components in an XML document, and (ii) an additional *specific logic* layer, which is understood as a set of additional services required by the linguistic layer. While the second layer must be provided by using general-purpose programming languages, the first layer can rely on domain-specific languages to describe syntax-directed language processing tasks, like those provided by the parser generation tools alluded to in this paper. As a consequence, the advantages of the approach from the development and maintenance perspective become apparent. On one hand, the linguistic layer can be expressed in domain-specific, high-level and largely declarative ways, using translation schemes, which can contribute to facilitating its conception, development and maintenance. On the other hand, since the approach does not constrain the nature of the specific logic layer, it is as general as any of the aforementioned general-purpose approaches. However, as a disadvantage, developers must face an increment in complexity due to the explicit organization of processing applications in these two well-differentiated layers. Of course, and as indicated in section 1, whether this complexity pays out or not will depend on the nature of the XML-based markup language: the more complex the language is, the more convenient the adoption of this proposal will be. Indeed, the non-trivial complexity of the markup languages that can arise in the domain of application generators makes this approach very convenient for this domain.

Our proposal can also be compared to traditional approaches from the point of view of efficiency, although, concerning the domain of application generators, where the documents involved will usually be small, this factor is less critical than ease of development and maintenance. Still, since our approach is intrinsically stream-oriented, it can usually give performances comparable to pure stream-oriented approaches based, for instance, on SAX or StAX. Indeed, another advantage to our approach, which is a direct consequence of using syntax-directed translation specifications built on top of underlying context-free grammars, arises: *when we develop XML processing applications we can think of trees, but the final applications will be executed as stream-oriented ones*. Therefore, the approach can achieve (and even beat) the usability of tree-oriented processing solutions, as well as the efficiency of stream-oriented ones.

Since it promotes a generative strategy to derive the actual implementation of the linguistic layer from a high-level specification based on the input language of a parser generation tool, our proposal has some points in common with XML *data binding* proposals [20]. A typical data-binding framework incorporates generators that are able to generate an application-specific representation by processing the *document grammar* (i.e., DTD or XML Schema) for the application's document type. As with the other conventional approaches mentioned, this representation is typically *data-centric*, as it consists of a set of application-specific classes, which are instantiated during parsing. Nevertheless, data-binding proposals are not exempt from disadvantages. Indeed, these proposals are tightly coupled with the document grammar, which is turned into application-specific classes using a more or less rigid set of pre-established rules. Although the proposals usually support *binding specifications*, which let developers modulate the classes generated and the *bindings* for the documents, the transformational capabilities of these specifications are usually limited to simple mapping facilities for elements and attributes. While these capabilities are sufficient for simple data-oriented XML applications, they fail when facing complex and/or mixed-element content models arising in non-trivial XML-based markup languages (such as those used in the domain of application generators). Our proposal, in turn, makes it possible to base the processing on generation-specific grammars, which are specific, not only to each language, but also to the processing task at hand.

4.2. Language-driven processing of XML documents

The conception of applications that process XML (or, more generally speaking, structured) documents as a sort of *compiler* or *translator* for a computer language has a long tradition in the document engineering context, such that it is highlighted, for instance, in [16]. Indeed, as it made apparent in [15], the internals of general-purpose XML processing frameworks can be explained from the point of view of conventional computer language processing workflows. However, as discussed in section 1, the application-specific processing of the documents usually operates on the data structures representing the documents provided by these frameworks. As a consequence, this application-specific processing is usually viewed as the processing of conventional data structures (e.g., traversing DOM trees, responding to SAX events, …) and the connection with language processing methods, techniques and tools is definitively missed. In order to restage this connection, some proposals (which are typically used for educational purposes) suggest undertaking the development of XML-based applications by building a parser for each particular XML-based markup language with the help of a parse-generation tool (see, for instance, [29], pages 351-352). As we indicated in section 3.1, this straightforward approach, however, supposes that we ignore general features common to any XML processing application (e.g., entity processing, comment recognition, namespace support, etc.). In this paper, we have shown how it is possible to use conventional parse generation tools in combination with standard XML processing frameworks to achieve the benefits of both approaches: on one hand, using standard and well-proven general-purpose XML processing frameworks to take advantage of general-purpose features common to any XML application, and, on the other hand, being able to organize application-specific processing in linguistic terms, as promoted by parse-generation tools.

The idea of parser generators have inspired several proposals for the construction of XML processing applications (e.g., ANTXR [40], which is built on top of the ANTLR parser generator tool, and RelaxNGCC [27], an extension of the RelaxNG [42] schema language for the specification of translation schemes). While these proposals usually rely on specialized tools supporting dedicated specification languages, in this paper we have shown how it is possible (and reasonable) to use conventional and well-proven parser generation tools without requiring dedicated languages for the description of the translation schemes. As indicated above, this fact is confirmed in our previous works [33][34], where we have shown how it is possible to build sophisticated XML processing environments by combining parser generators (JavaCC [14] and CUP [2]) with general-purpose stream-oriented XML processing frameworks (SAX and StAX).

4.3. XML Processing and Attribute Grammars

Although the tendency in formal models for processing XML documents is to emphasize tree automata and related formalisms [36], there are several works on using *attribute grammars*, a well-known formalism for describing the syntax and semantics of context-free languages [12][28], for the language-oriented implementation of XML processing tasks. Many of these works are typically focused on amalgamating attribute grammar concepts with the EBNF syntax that usually underlies an XML DTD, and which is reflected in unranked tree representations for the XML documents. The approach adopted in [31] to cope with EBNF is to decouple semantic rules and productions. Indeed, their semantic rules are associated in terms of parent-child relationships, instead of being associated with productions. This problem was addressed early by the work reported in [8] regarding a transformation system for structured documents supporting different document models (e.g., SGML, LaTEX, etc.). In [24][25], this kind of extended attribute grammar is used for querying structured documents, and it constrains the type of semantic expressions allowed to regular expressions in the alphabets of attribute occurrences. In the work described in [17][18], which reports on an application in the domain of information retrieval, documents are represented using abstract attribute grammars, where each non-terminal corresponds to an element type. In this work, a set of pre-established rules is used to derive such grammars from the DTDs, using a similar approach to that described in [16] (see [18] for an explicit enumeration of these rules). In [13], l-attributed grammars defined from EBNF syntaxes are used to support the efficient processing of XML streams. Similarly, in the works reported in [23][26] the unranked nature of the XML document trees is managed by promoting binary encodings of these document trees. Finally, in [32][35] we describe XLOP (XML *Language-oriented Processing*), an attribute grammar–based front-end to the proposal described in this paper. Indeed, by using encoding patterns similar to those described in [4] to implement attribute grammars by using conventional compiler construction tools, XLOP is able to turn the attribute grammar-based specifications of XML processing tasks into translation schemes for the CUP parser generation tool.

Our work in XLOP makes the relationships between the proposal described in this paper and attribute grammar-based approaches to XML processing apparent. Indeed, since the designer who writes an attribute grammar does not need to specify the evaluation order for the semantic equations, attribute grammars are of a higher level than translation schemes, where designers must make the execution order of the semantic actions explicit. However, many times it can burden the applicability of the approach, since average developers, who do not necessarily have deep knowledge of specialized formal semantic specification techniques, usually find it hard to work with non-standard computation models [9], like the dependency-driven one that underlies attribute grammars. For this purpose, the plain use of parser generation tools presented in this paper can provide an intermediate approach that can be more easily accepted by developers of XML processing applications. Also, sometimes parser generation tools can lead to more efficient / more straightforward implementations than those directly generated from attribute grammars. In addition, the use of patterns like the one described in [4] can enable hybrid approaches: indeed, it is possible to start with an attribute grammar-based specification, to encode it as a translation scheme using the patterns given in [4], and then to evolve it into a more efficient / more conventional implementation. These ideas have been partially applied in [34] by including dependency-driven translation capabilities in the application of the approach to the CUP + STaX marriage. Finally, based on our experiences, we have realized that one of the key aspects of the approach described in this paper is to perform the explicit provision of (plain BNF) context-free grammars (e.g., the generation-specific grammar) instead of relying on direct EBNF counterparts to the DTDs / document schemas as in [13][24][25], on pre-established rules to convert (EBNF-based) document grammars into BNF grammars as in [17][18], on the explicit decoupling of syntax and semantics as in [31], or on pre-established encodings of the document trees as in [23][26].

5. Conclusions and future work

In this paper, we have proposed a metalinguistic conception of the development of XML-driven application generators. According to this approach, these generators are treated as a sort of language processor. This treatment allows us to use compiler construction tools, and in particular parser generators, as adequate tools to orchestrate the development. It enables the automatic production of application generators from high-level specifications based on generation-oriented translation schemes. In addition, these application generators can be smoothly integrated with general-purpose standard XML-processing frameworks by using a generic and customizable XML Scanner. The approach facilitates the development and maintenance of application generators driven by complex XML-based markup languages, as well as by huge data-intensive XML documents and/or by documents that are provided asynchronously in an XML data stream.

Currently we are working on more flexible configuration mechanisms for the XML Scanner. We are also investigating mechanisms to improve the efficiency of the final generators. We are also planning to test the approach on the development of other application generators in the e-Learning domain, such as was reported in [22][37], as well as in the domain of multi-agent systems [10].

Acknowledgements. Thanks are due to project grants TIN2010-21288-C02-01 and Santander-UCM GR 42/10, group reference 962022. Also, Daniel Rodriguez-Cerezo was supported by the Spanish University Teacher Training Program (EDU/3445/2011).

References

1. Aho A.V., Lam M.S., Sethi R., Ullman J.D.; Compilers: principles, techniques and tools (2nd edition). Addison-Wesley. (2006)
2. Appel, A.W.: Modern Compiler Implementation in Java. Cambridge University Press. (1997)
3. Brownell, D. SAX2. O'Reilly. (2002)
4. Cerezo, D., Sarasa, A., Sierra, J.L. Implementing Attribute Grammars Using Conventional Compiler Construction Tools. 3rd Workshop on Advances in Programming Languages (WAPL'11), Szczezin, Poland. (2011)
5. Cleaveland, J.C.: Building Application Generators. IEEE Software, Vol. 5, No. 4, 25-33. (1988)
6. Cleaveland, J.C.: Program Generators with XML and Java. Prentice Hall. (2001)
7. Czarnecki, K.: Generative Programming: Methods, tools and Applications. Addison-Wesley. (2000)

8. Feng, A., Wakayama, T.: SIMON: A Grammar-based Transformation System for Structured Documents. Electronic Publishing, Vol. 6, No. 4, 361-372. (1993)
9. Fowler, M.: Domain Specific Languages. Addison-Wesley. (2010)
10. Fuentes-Fernández R., Gómez-Sanz J., Pavón J.: Requirements Elicitation and Analysis of Multiagent Systems Using Activity Theory. IEEE Transactions on Systems, Man, and Cybernetics, Part A, Vol. 39, No. 2, 282-298. (2009)
11. IMS. IMS Question and Test Interoperability 2.1. www.imsglobal.org/question/
12. Knuth, D.E: Semantics of Context-free Languages. Mathematical System Theory Vol. 2, No. 2, 127-145. (1968)
13. Koch, C., Scherzinger, S.: Attribute Grammars for Scalable Query Processing on XML Streams. The VLDB Journal, Vol. 16, No. 3, 317-342. (2007)
14. Kodaganallur, V.: Incorporating Language Processing into Java Applications: A JavaCC Tutorial. IEEE Software, Vol. 21, No. 4, 70-77. (2004)
15. Lam, T.C., Ding, J.J., Liu, J.C.: XML Document Parsing: Operational and Performance Characteristics. IEEE Computer, Vol. 41, No. 9, 30-37. (2008)
16. Leite-Ramalho, J.C.: Anotação Estrutural de Documentos e sua Semântica -- Especificação da Sintaxe, Semântica e Estilo para Documentos. Ph.D. Thesis. Braga, Portugal. (2000)
17. Lopes-Gançarski, A. L., Doucet, A., Rangel-Henriques, P.: Grammar-based Interactive System to Retrieve Information from XML Documents. IEE Proceedings-Software, Vol. 153, No. 2, 51-60. (2006)
18. Lopes-Gançarski, A., Rangel-Henriques, P.: Information Retrieval from Structured Documents Represented by Attribute Grammars. International Conference on Information Systems Modeling, Rep. Cheque. (2002)
19. Marini, J.: Document Object Model: Processing Structured Documents. McGraw-Hill. (2002)
20. McLaughlin, B. Java & XML Data Binding. O'Reilly. (2002)
21. McLaughlin, B. Java & XML. O'Reilly. (2006)
22. Moreno-Ger P, Sierra J.L., Martínez-Ortiz I., Fernández-Manjón B.: A Documental Approach to Adventure Game Development. Science of Computer Programming, Vol. 67, No. 1, 3-31. (2007)
23. Nakano, K.: An Implementation Scheme for XML Transformation Languages Through Derivation of Stream Processors. Second Asian Symposium of Programming Languages and Systems (APLAS'04), Taipei, Taiwan. (2004).
24. Neven, F.: Attribute Grammars for Unranked Trees as a Query Language for Structured Documents. Journal of Computer and System Sciences, Vol. 70, No. 2, 221-257. (2005)
25. Neven, F.: Extensions of Attribute Grammars for Structured Document Queries. 7th International Workshop on Database Programming Languages (DBLP'99), Kinloch Rannoch, Scotland, UK. (1999)
26. Nishimura, S., Nakano, K.: XML Stream Transformer Generation through Program Composition and Dependency Analysis. Science of Computer Programming, Vol. 54, No. 2-3, 257-290. (2005)
27. Okajima,D.: RelaxNGCC -- Bridging the Gap Between Schemas and Programs. XML.com, 8. (2002)
28. Paakki, J.: Attribute Grammar Paradigms - A High-Level Methodology in Language Implementation. ACM Computer Surveys, Vol. 27, No. 2, 196-255. (1995)
29. Parr, T.: Language Implementation Patterns. Pragmatic Bookshelf. (2010)
30. Parr, T.: The Definitive ANTLR Reference: Building Domain-Specific Languages. Pragmatic Bookshelf. (2007).

31. Psaila, G., Crespi-Reghizzi, S.: Adding Semantics to XML. 2nd International Workshop on Attribute Grammars and their Applications (WAGA'99), Amsterdam, The Netherlands. (1999)
32. Sarasa, A., Martinez-Aviles, A., Sierra, J.L., Fernández-Valmayor, A.: A Generative Approach to the Construction of Application-Specific XML Processing Components. 35th Euromicro Conference on Software Engineering and Advanced Applications, Patras, Greece. (2009)
33. Sarasa, A., Navarro, I., Sierra, J.L, Fernández-Valmayor, A.: Building a Syntax Directed Processing Environment for XML Documents by Combining SAX and JavaCC. 3rd Int. Workshop on XML Data Management Tools & Techniques. DEXA'08. September 1-5, Turin, Italy. (2008)
34. Sarasa, A., Temprado, B., Martínez, A., Sierra, J.L., Fernández-Valmayor, A.: Building an Enhanced Syntax-Directed Processing Environment for XML Documents by Combining StAX and CUP. Fourth Int. Workshop on Flexible Database and Information Systems. DEXA'09. August 31 – September 4, Linz, Austria. (2009)
35. Sarasa, A., Temprado-Battad, B., Sierra, J.L, Fernández-Valmayor, A.: XML Language-Oriented Processing with XLOP. 5th International Symposium on Web and Mobile Information Services, Bradford, UK. (2009)
36. Schwentick, T.: Automata for XML - A Survey. Journal of Computer and System Sciences, Vol. 73, No. 3, 289-315. (2007)
37. Sierra, J.L, Fernández-Valmayor, A., Fernández-Manjón, B.: From Documents to Applications Using Markup Languages, IEEE Software, Vol. 25, No. 2, 68-76. (2008)
38. Sierra, J.L., Fernández-Valmayor, A., Fernández-Manjón, B., Navarro, A.: ADDS--A Document-Oriented Approach for Application Development, Journal of Universal Computer Science, Vol. 10, No. 9, 1302-1324. (2004)
39. Sierra, J.L., Fernández-Valmayor, A., Fernández-Manjón, B.: A Document-Oriented Paradigm for the Construction of Content-Intensive Applications. The Computer Journal, Vol. 49, No. 5, 562-584. (2006)
40. Stanchfield, S. ANTXR: Easy XML Parsing, based on The ANLR Parser Generator. javadude.com/tools/antxr/index.html (current October 2011)
41. Tidwell, D.: XSLT, 2nd Edition. O'Reilly. (2008)
42. Vlist, E. Relax NG. O'Relly. (2003)
43. Wallmsley, P. XQuery, O'Reilly. (2007)

Antonio Sarasa-Cabezuelo is a full-time Lecturer in the Computer Science School at Complutense University of Madrid, Spain (UCM). His research is focused on the language-oriented development of XML-processing applications, and on the development of applications in the fields of digital humanities and e-Learning. He was one of the developers of the Agrega project on digital repositories (a pioneer project in this field in Spain). He is a member of the research group ILSA (Implementation of Language-Driven Software and Applications: http://ilsa.fdi.ucm.es). He has participated in several research projects in the fields of software language engineering, digital humanities and e-learning, and he has published over 50 research papers in national and international conferences.

Bryan Temprado-Battad is a PhD. Student in the Computer Science School at UCM and a member of the ILSA Research Group. His research is focused on language oriented development and on attribute grammars applications, being one of the principal contributors to the development of the XLOP System. The results of his works have been published in several research papers in international journals and conferences.

Daniel Rodríguez-Cerezo is a PhD student in the Computer Science School at UCM, and a member of ILSA. His research is focused on the use of several e-Learning techniques (simulations, interactive prototyping tools, recommendation systems for learning object repositories, etc.) to improve teaching and learning of the Software Language Engineering discipline. Besides, he is interested in the development and improvement of software language engineering techniques.

José-Luis Sierra is an Associate Professor at the UCM's Computer Science School, where he leads the ILSA Research Group. His research is focused on the development and practical uses of computer language description tools and on the language-oriented development of interactive and web applications in the fields of digital humanities and e-Learning. Prof. Sierra has leaded and participated in several research projects in the fields of digital humanities, e-learning and software language engineering, the results of which have been published in over 100 research papers in international journals, conferences and book chapters. He serves regularly as reviewer / PC Member for several international reputed journals and conferences.

Received: May 05, 2011; Accepted: March 21, 2012.

DOI: 10.2298/CSIS110716006M

Towards Understanding of Classes versus Data Types in Conceptual Modeling and UML

Dragan Milićev

University of Belgrade
Faculty of Electrical Engineering, Department of Computing
P.O. Box 35-54, 11120 Belgrade, Serbia
dmilicev@etf.rs

Abstract. Traditional conceptual modeling and UML take different vague, ambiguous, and apparently incompatible approaches to making a distinction between two different entity types – classes and data types. In this paper, an in-depth theoretical study of these ambiguities and discrepancies is given and a new semantic interpretation is proposed for consolidation. The interpretation is founded on the premise that populations of the two kinds of entity types are defined in two substantially different ways: by intensional (for data types) and extensional (for classes) definitions. The notion of a generative relationship set is introduced to explain the role of specific relationship types that are used to define populations of structured data types by cross-combinations of populations of the related entity types. Finally, some important semantic consequences are described through the proposed interpretation: value-based vs. object-based semantics, associations vs. attributes, and identity vs. identification. The given interpretation is based on runtime semantics and allows for fully unambiguous discrimination of the related concepts, yet it fits into intuitive understanding and common practical usage of these concepts.

Keywords: conceptual modeling, Unified Modeling Language (UML), formal semantics, class, data type, entity, relationship, object identity, identification, association, attribute.

1. Introduction

Ever since their introduction with Entity-Relationship (ER) modeling [4], the notion of *entity type*, along with its more recent and object-oriented successor concept of *class*, have had a central role in conceptual modeling and programming in general. In addition, classical conceptual modeling recognizes the concept of *data type* as a special kind of entity type, and fairly clearly and seemingly unambiguously describes the distinction between data types and the other entity types [17, 6]. Unfortunately, as it will be shown in this paper, this distinction does not seem to be precise enough to allow clear separation and proper use of the two kinds of entity types. The lack of unambiguous criteria for discriminating between data types and the other entity types inevitably causes uncertainty to modelers about which of the two concepts should be used in each particular case, or about what the original semantic intent was behind a particular use of one or the other in a model.

On the other hand, the Unified Modeling Language (UML) has adopted the analogous separation of *classes* and *data types* in a yet more explicit way, clearly emphasizing a few distinctive characteristics of classes versus data types [16]. However, as it will be discussed in this paper, the definitions of the distinction between classes and data types appear to be completely different from those widely used in classical conceptual modeling. Such a situation unnecessarily increases the confusion in understanding of these notions and their use in practice. This is yet another of many existing parts of the UML's definition that contributes to its main drawback – the lack of a complete semantic formalization. Many concepts of UML, especially in its early versions, have not had definitions precise enough to be interpreted unambiguously. In other words, many parts of a syntactically correct UML model can still be interpreted in different ways by different readers. This has been recognized as the main obstacle for UML to become a machine-interpretable, i.e., executable language. Instead, UML served predominantly for recording ideas, sketches, and design decisions in early phases of software analysis and design, without an ambition of encouraging unambiguously interpretable models.

Recent model-driven development (MDD) trends in software engineering [22, 23, 14] have dramatically increased the importance of formalizing the semantics of UML, as the key prerequisite for its switch from a solely descriptive to an unambiguously interpretable and executable language. As a result, a major revision of UML has emerged in its version 2 [16], making a significant step towards precise semantics of concepts. Much effort has been made to clarify the meaning of the widely adopted concepts in the very UML 2 specification, as well as in other attempts before and around it [1, 3, 5, 7, 8, 9, 14, 15, 17, 18, 19, 24]. However, the distinction between classes and data types has not yet been fully clarified and consolidated with the conception in classical conceptual modeling.

In this paper, we describe apparent inconsistencies in and incompatibilities between the definitions given in classical conceptual modeling and in UML, address these issues, and propose consolidation, attempting to get closer to a fully clear understanding of these classical and widely used notions. We also show how the understanding of these notions affects the understanding of associations versus attributes in UML. We believe that the improvements in the common understanding of these notions can lead to better communication between software designers and developers, proper and certain use of these concepts in conceptual modeling, as well as their coherent implementation in environments for building and running executable models.

The rest of the paper is organized as follows. The next section gives basic definitions and assumptions in the field as the foundations for the discussions that follow. Section 3 brings an overview of the treatment of the concepts relevant for the subject matter in classical conceptual modeling and in UML, and reveals some ambiguities, discrepancies, or seeming contradictions. Section 4 gives an informal analysis of the subject matter that leads to the proposed solution in an intuitive way. Section 5 formalizes the proposed interpretations through precise definitions. Section 6 summarizes some semantic and practical consequences of the presented interpretations. The paper ends with conclusions.

2. Basic Concepts and Assumptions

2.1. Entity Types

To avoid terminological confusion and ambiguity, we will use the terms *class* and *data type* adopted from UML to refer to the two different kinds of classifiers of relevance in this discussion, and the term *entity type* from classical conceptual modeling and ER [4, 17] to refer to both (or any). In other words, entity type will be used to refer to a generalization of the concepts of class and data type, while the division of entity types into classes and data types is assumed to be a partition, i.e., a covering and disjoint specialization.[1]

As a starting definition, we can say that an entity type is a concept whose instances at a given time are individual entities that exist in the domain at that time. This is a reasonably precise and, with some variations, widely adopted definition. For example, this is a variation of the definition from classical conceptual modeling [17], with one intentional modification: the definition in [17] requires that instances of every entity type be identifiable, while we will revisit this aspect in this paper. Put another way, entity types represent sets of individual entities. These sets are generally variable in time. Variability of entity types over time is one of the fundamental assumptions in conceptual modeling. Entities can begin or cease to exist, or they can be reclassified to other entity types [17, 6, 14]. A more formal definition of entity types as variable sets can be found in [6].

[1] We will not use the UML term *classifier* for such a generalizing concept, because there are many other kinds of classifiers in UML, such as collaborations or use cases, which are out of scope of this discussion.

In order to make this definition slightly more practical and allow some more precise semantic interpretations that will be given in the paper, we will refine this definition and say that an entity type represents a set of entities that exist in the *system* or in the *domain* in *runtime*. By referring to runtime, we imply an inherent dichotomy between *design time* (or modeling time) and *runtime* that exists in practice. The exact practical meaning of these two terms certainly depends on a concrete framework or tooling in case, but traditionally, they refer to the activities of data schema design or modeling (for design time), and data manipulation or model execution (for runtime). It should be noted that the distinction between the two is mostly ontological and does not necessarily imply different physical times; in reality, these two may refer to the same physical time, meaning that the two activities (of system design and its execution) may take place simultaneously. For example, many database management systems and other frameworks allow modifications of the database schema or model during the exploitation of the system; others may need to interrupt the exploitation of the system in order to recompile and redeploy the schema/model, and to restart the execution environment, but without affecting the existing entities that survived the interruption. Anyhow, we consider the runtime as an absolute temporal scope of all entities that may exist in a certain system, and outside which the entities cannot exist due to the very nature of the considered system or the technology used. That might be, for example, the lifetime of a certain installation of a database system, or an execution of a program in the classical sense.

A *constant entity type* is an entity type whose set of instances is constant, i.e., invariable (immutable) in runtime [17, 6].

2.2. Relationship Types

For similar reasons as for entity types, before we discuss and clarify the semantics of attributes and associations in UML, we will use a generic term *relationship type* taken from classical conceptual modeling and ER. A relationship type is a concept whose instances at a given time are individual relationships between entities that are considered to exist in the domain at that time [17]. Again, we have omitted the requirement in the definition given in [17] that individual instances of a relationship type be identifiable in order to allow interpretations like the one given in [14, 15] and to align it with the definition given in UML 2 [16], where instances of relationship types (i.e., associations in UML) cannot be identified. We do not see any significant semantic or practical impact of this modification.

A relationship type of degree $n \geq 2$ consists of an ordered set of n participants, whereby a participant is an entity type that plays a role in the relationship type [17]. We will write $R(p_1:E_1, ..., p_n:E_n)$ to denote a relationship type named R, with participant entity types E_1, ..., E_n, playing roles p_1, ..., p_n respectively. Note that E_1, ..., E_n do not have to be different entity types, because the same entity type can play several roles in the same relationship type. Roles must be, naturally, pairwise different. We may omit the role p_i played by participant p_i:E_i either because it is obvious or it is the same as the name of E_i; then it is assumed that p_i is the same as E_i. For example, *Reads*(*reader*:*Person*, *Book*) is the same as *Reads*(*reader*:*Person*, *book*:*Book*).

A relationship of type R has a form of a set {<p_1:e_1>, ..., <p_n:e_n>}, sometimes also referred to as a *tuple*, where e_1, ..., e_n are instances of their corresponding entity types E_1, ..., E_n. In classical conceptual modeling, all relationships of a certain relationship type that exist at a certain moment in runtime form a set of distinct tuples [17]. A more recent interpretation [14, 15] defines relationship types as concepts whose instances at a certain moment in runtime form a bag of tuples (i.e., a multiset), in order to support advanced notions of uniqueness of roles in UML 2. The discussions in this paper are independent of the interpretation used.

A binary relationship type R(p_1:E_1, p_2:E_2) defines two inverse mappings [5, 6, 14, 15]:

- p_1 that maps an instance e_2 of E_2 to a set (or bag in [14, 15]) of all those and only those instances e_1 of E_1 for which {<p_1:e_1>, <p_2:e_2>} is in R at any particular moment in runtime;
- p_2 that maps an instance e_1 of E_1 to a set (or bag) of all those and only those instances e_2 of E_2 for which {<p_1:e_1>, <p_2:e_2>} is in R at any particular moment in runtime.

It is easy to see that e_1 is in p_1(e_2) if and only if e_2 is in p_2(e_1). The dynamicity of the sets of instances of entity and relationship types during runtime implies that these mappings are generally also variable in time [5, 6].

A relationship type R(p_1:E_1, ..., p_n:E_n) is *constant* with respect to a particular participant p_i if the instances of R in which an instance e_i of E_i participates are the same during the temporal interval in which e_i exists, for each e_i of E_i. A relationship type is constant if it is constant with respect to all its participants [17]. Obviously, when a binary relationship R(p_1:E_1, p_2:E_2) is constant with respect to e.g. p_1, it means that the mapping p_2(e_1) is constant during the lifetime of e_1 for each e_1 of E_1.

2.3. Populations and Actions

We will refer to the set of instances of an entity type E that exist at a certain moment in runtime as the *population* of E.[2] Similarly, the set or bag of instances of a certain relationship type R that exist at a moment in runtime will be referred to as the *population* of the relationship type.

[2] Sometimes also called the *extent* of E. However, we will use the term *extent* for a slightly different concept.

Populations of entity and relationship types are assumed to be dynamically changed during runtime by means of *actions*. Actions are atomic units of behavior that affect populations of entity and relationship types. In this context, we are focused on a generic set of elementary actions on entity and relationship types, without going deeper into their formal semantics as they are not directly relevant to the conclusions of this paper:

- *Create a new instance of one or more given entity types*, which adds a new entity *e* to the system and to the populations of the given entity types. (In general, an entity can be an instance of more related or unrelated entity types.)
- *Delete an existing entity e*, which removes *e* from the populations of its entity types and from the entire system. Deletion of an entity implicitly removes all relationships in which that entity participates.
- *Reclassify an existing entity e*, by removing the entity from the population of zero or more given entity types and adding it to the population of zero or more entity types. Removing an entity from the population of an entity type implicitly removes all relationships in which that entity participates with a role of that entity type.
- *Create a new instance of the given relationship type R*, which adds a new relationship {<p_1:e_1>, ..., < p_n:e_n>} to the population of *R*.
- *Delete an existing relationship*, which removes a given relationship {<p_1:e_1>, ..., < p_n:e_n>} from the population of its relationship type.

While the system's overall state is defined in terms of populations of its entity and relationship types, the system's behavior is defined in terms of executed actions. More detailed and formal treatments of actions and their semantics can be found in [6, 14].

Constant entity and relationship types have immutable populations. This means that actions on constant entity or relationship types that would change their populations are illegal.

2.4. Identification

An *identifier* of an entity *e* is an expression, written in some language, that unambiguously denotes *e* [17].[3] In this paper, most relevant are identifiers built from binary relationship types that have particular properties, referred to as *reference relationship types*, as defined in [17]. A more general approach to building identifiers through so-called *observation terms* is presented in [10].

[3] In general, identifiers may have temporal or other kinds of scopes, meaning that the same identifier (being an expression) may denote different entities at different times or in different scopes. We will not consider this case, as it is orthogonal to the discussions in this paper.

We will here provide a semi-formal definition of observation terms, which is sufficient for the discussions given in this paper; for a fully formal definition, the reader is referred to [10].

Definition (Observation term). *Observation term* of E (denoted with $\tau(E)$) is an injective function τ: $\pi(E)\rightarrow S_\tau$, where $\pi(E)$ is the population of an entity type E at any moment in runtime, and S_τ is a certain fixed set:

$$(\forall e_1, e_2\in\pi(E))(\tau(e_1)=\tau(e_2) \Rightarrow e_1=e_2).$$[4]

Even less formally, $\tau(E)$ is a function that maps an entity $e\in\pi(E)$ to an atomic or structured value (formally bound by a certain set S_τ) so that $\tau(e)$ uniquely identifies e. Observation terms are more general than simpler identifiers normally used in practice, such as reference relationship types, because $\tau(E)$ may map an e to an arbitrarily deeply structured value whose components may be obtained by more complex mappings than simple relationships, like queries. Of course, reference relationship types (both simple and compound) are special cases of observation terms. In [10], Gogolla gives a detailed definition of how $\tau(E)$ can be constructed, i.e., how S_τ can be defined.

By its definition, $\tau(e)$ represents an identifier of e, meaning that it implies an inverse function f_τ: $S_\tau \rightarrow\pi(E)$ that maps a value from S_τ to an entity e, for a subset of S_τ for which it is defined. It is important to note that $\tau(E)$ is an injection that does not have to be a surjection, meaning that there does not have to be an $e\in\pi(E)$ for each element s of S_τ so that $\tau(e)=s$. This is trivially true because the set $\pi(E)$ can be dynamically changed in runtime, while S_τ is fixed. Therefore, in general, f_τ is a partial function, defined for a subset of S_τ. The function f_τ is the *identification function* that can be used for identifying entities in $\pi(E)$.

For a comprehensive study of identification and formalization of conceptual modeling in general, the reader is referred to [17, 10].

[4] The formal semantics of equality is a very subtle and difficult topic; for example, see a very interesting perspective and conclusions described in [13]. This is why we do not go deeper into that matter. We may simply note that the operation "=" may mean equality (up to isomorphism) or identity in different contexts. It seems that our observations do not require a fully formal clarification of this issue and that the interpretation of $e_1=e_2$, where e_1 and e_2 are (references to) entities, is that e_1 and e_2 represent the same entity (identity).

3. Overview of the Matter

3.1. Classes vs. Data Types in Classical Conceptual Modeling

Classical conceptual modeling defines *data type* as a constant entity type whose instances, called *values*, can be identified by literals, whereby literals are strings (i.e., sequences) of some symbols (e.g., characters) [17, 6]. The same value of a certain data type can often be identified by several different literals; for example, both 1.0 and 0.1E+1 denote the same value one of a floating point data type. Although this definition sounds intuitively clear and reasonably practical, it is still formally vague and conceals several important issues.

First, isn't it possible that an instance of a class (i.e., of an entity type that is not a data type) be identifiable by a literal? For example, one can imagine a system where each instance of a constant class is assigned (e.g., at design time) a literal that identifies it, so identifiability through literals is not a sound discriminator between the two notions. (It will be shown later that this is exactly the case with enumerations.) On the other hand, is it really necessary that each instance of every possible data type (e.g., a structured one) be identifiable by a literal? What is the real difference between data types and constant classes then?

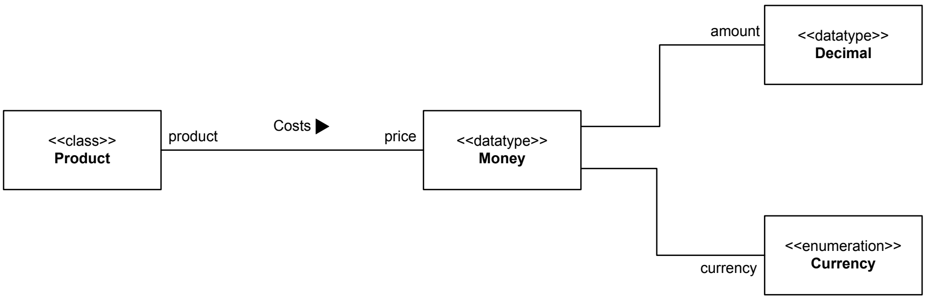


Figure 1. A constant entity type *Money* with a non-constant relationship type with *Product* and constant relationship types with *Decimal* and *Currency*

Just being a constant entity type is not a sufficient discriminator for data types, simply because classes can also be constant. Data types often imply constancy of some (but not all) relationship types. Figure 1 gives an example. The data type *Money* is a structured entity type that represents a certain amount of money expressed in a given currency; these two properties of *Money* are modeled with the two respective relationship types with *Decimal* and *Currency* data types.[5] Obviously, these two relationship types are constant with respect to *Money*. One should not change the pair of instances of *Decimal* and *Currency* related to a certain instance of *Money*, because it would change the very identity of that instance of *Money*. Indeed, if it were the opposite, one could relate two instances of *Money* with the same pairs of *Decimal* and *Currency*, thus making two undistinguishable and fundamentally the same instances of *Money*. In fact, these two relationship types form a compound reference [17] of *Money*, so this situation should not occur. On the other hand, however, there are usually non-constant relationship types in which a data type participates. The relationship type *Costs* is obviously non-constant with respect to *Money*. However, to the best of the author's knowledge, there is not a clear explanation and understanding of these aspects in the open literature, especially of which relationship types of a data type are or are not constant (with respect to that data type).

Beeri [2] gives one of the most profound analyses on the difference between classes (i.e., objects as their instances) and data types (i.e., values as their instances) in the context of a formal foundation of object-oriented databases. However, his discussion is mostly restricted to primitive data types. Using the example of numbers, Beeri gives a few criteria for differentiating data types from classes (referred to as abstractions in his discussions); we first quote each criterion and then give our comments on it:

"1. Numbers are a universally known abstraction, and have the same meaning to all people. Abstractions such as employees are application specific."

This is correct for the case of most general and application-independent data types. However, applications very often introduce quite specific, derived data types, typically structured ones.

"2. Numbers have names in the name space. In contrast, application specific abstractions normally do not have names."

We agree that objects of application-specific abstractions (classes) normally do not have names as their identifiers, but as discussed above, there is not any practical or formal reason against having names, as also clearly indicated in the same Beeri's paper. There are many examples of such cases, and this paper will later argue that enumerations are actually constant classes with named objects.

[5] One can notice that *Money* is actually a reified ternary relationship type between *Product*, *Decimal*, and *Currency*. It might be regarded so, but the reification (i.e., the promotion of the ternary relationship type into the entity type with the three binary relationship types) has been done for the purpose of reusability of the data type *Money* in other places in a broader model, such as for properties of other entity types.

"3. Numbers (and their names) are built into the system; they need not be defined. Object abstractions need to be introduced to the system by definitions."
This is actually a distinction between built-in and user-defined (application-specific) entity types, primarily determined by the construction of a particular language or model, but not a distinction between data types and classes. We believe that this distinction is orthogonal to the distinction between classes and data types, as both classes and data types can be built-in as well as user-defined. For example, a structured derived data type such as *Money* in Figure 1 can be either built-in or user-defined, depending on the concrete implementation of the language or framework.

"4. The information carried by a number is in itself. ... In contrast, the interesting information about application specific abstractions is carried in relationships they have with other objects and values."
We argue that this is a correct distinction between primitive data types on one hand and both classes and structured data types on the other, but not between classes and data types. For example, a structured derived data type such as *Money* in Figure 1 is an example of a data type whose interesting information is truly in its relationships with *Decimal* and *Currency* and not in itself.

Beeri also concludes that "none of these distinctions is absolute" and that "both objects and values are objects in the general intuitive sense." In this paper, we will stay aligned with this conclusion, but will try to provide a more precise and unambiguous distinction between classes and data types.

3.2. Classes vs. Data Types in UML

The confusion becomes even worse when the definitions of data types in UML get involved [16]. In UML, instances of classes are entities (called *objects*) with inherent identity, while instances of data types are "pure values" that have no identity: "A data type is a special kind of classifier, similar to a class. It differs from a class in that instances of a data type are identified only by their value. All copies of an instance of a data type and any instances of that data type with the same value are considered to be the same instance. Instances of a data type that have attributes (i.e., is a structured data type) are considered to be the same if the structure is the same and the values of the corresponding attributes are the same. If a data type has attributes, then instances of that data type will contain attribute values matching the attributes." [16, p. 60]

This "definition" is extremely vague. First, the construction that "all copies of an instance of a data type and any instances of that data type with the same value are considered to be the same instance" sounds logically problematic indeed. Second, it is unclear what "values" really are. Then, in case of structured data types, the definition partially relies on the notion of attributes, which is even more unclear, as it will be discussed soon. Finally, even if the notion of attribute is taken to be just a relationship type as in [17], it is unclear which relationship types have to relate the given two instances of a data type to the same sets of other instances in order to be treated as "having the same value." For example, for the data type *Money* in Figure 1, if two instances of *Money* have relationships with the same instances of *Decimal* and *Currency*, respectively, they obviously have to be interpreted as having the same "value". However, these two same instances of *Money* obviously may be related to different instances of *Product*, stating simply that these two products have the same price, but the two instances of *Money* are still having the same value, although they do not have the same relationships (structure). This observation indicates a certain hidden difference between the meaning and purpose of different relationship types relating the data type *Money*.

An elaborate discussion and clarification of the meaning of these statements and their practical implications can be found in [14]. In brief, every object of a class has its own identity, which is an inherent characteristic of that object that distinguishes it from any other object of the same or any other class. The identity of an object is inherent, meaning that it is ensured by the very existence of the object, and does not require any special activity by the modeler or user in design or runtime. For example, one instance of class *Person* is, by default, different and can be distinguished from any other instance of the same class or any other class. How it can be distinguished is a matter of a concrete technique, notation, and implementation. It is only important to emphasize that two objects are distinguishable simply because they are two separate identities by their nature and existence (that is, because they have been created by two separate executions of the Create Entity action). The identity of an object is its inherent characteristic, and is independent of any relationships (e.g., properties of the object or any other part of its state). In other words, two objects can have exactly the same relationships and exactly the same state, but they are still two different identities. For example, two instances of the class *Person* may have the same name, home address, date of birth, and all other properties, but they are still treated as two different and independent entities. In addition, the identity of an object is not affected by any modification of its properties.

Unlike class instances, instances of data types do not have their identity in UML's interpretation — that is, they are pure values. Two instances of the same data type with all properties having equal values cannot be distinguished. Being pure values without identity, instances of data types in UML are immutable, meaning that operations of data types do not have side effects and do not change values of their owner instances, but are pure functions that can only produce new instances and not change the existing ones.

To conclude, the two approaches taken in classical conceptual modeling and in UML obviously differ significantly. (The only apparent commonality is that operations of data types cannot modify their owner instances.) Both rely on the notion of value, but what is really "a value"? In UML, one can create many "equal" instances of a data type by Create Entity actions, while data types are constant in classical conceptual modeling and such actions should be treated as illegal.

3.3. Associations vs. Attributes

The confusion is similar with distinguishing associations and attributes as different subkinds of relationship types.

Olivé [17] clearly points out that attributes are very similar to binary relationship types and are not really needed at the conceptual level. In fact, an attribute is a binary relationship type $R(p_1{:}E_1,\ p_2{:}E_2)$ in which participant p_2 is considered to be a *characteristic* (or *property*) of E_1 (with E_2 being the type of the attribute), or p_1 a characteristic of E_2 (with E_1 being the type of the attribute). Therefore, an attribute is like a binary relationship, except that users and designers add the interpretation that one participant is a characteristic of the other. In associations, on the other hand, the order of the participants does not imply any priority or subordination between them.

Obviously, except from the rather subjective interpretation by the humans, there is not any formal distinction between attributes and associations. “Thus it is not clear whether attributes should be used or when,” concludes Olivé [17, p. 76].

Quite similarly, standard UML [16] does not impose any significant semantic difference between binary associations and attributes either. It simply provides them as different kinds of modeling elements, without giving any clear and formal difference in their runtime semantics. For example, one can use either or even both at the same time to model a specific binary relationship type between certain entity types. In other words, mostly similarly to classical conceptual modeling, standard UML treats attributes and associations just as two syntactic variations of the same background concept of a binary relationship type. Obviously, having such a situation does not help modelers and model readers too much, because it increases confusion in which modeling concept to apply or how to interpret the intention of the model designer when reading a model.

Instead of sound criteria on the use of attributes, some guidelines exist only. One guideline, well known in conceptual modeling and also suggested by standard UML [21], is that types of attributes should be entity types defined outside the scope of the considered model, while associations should be used to relate entity types that are defined within the considered model. This is, however, not a strict guideline; it just helps modelers make models easier to understand.

A formal and executable profile of UML named OOIS UML and described in [14] gives more precise and strict discriminating characteristics between the two notions: associations are relationships that may involve only classes as their participants, while attributes are binary relationship types whose one participant is always a data type and represents the type of the attribute, while the other can be either a class or a data type and represents the owner of the attribute. In other words, the type of an attribute may only be a data type. The profile also describes some additional semantic implications, briefly summarized as follows:

- Objects of classes are compared for equality by reference, while instances of data types are compared for equality by value.
- As they can have only attributes, instances of data types have implicit copy semantics that deeply (recursively) copies values of their attributes. Objects of classes do not have implicit copy semantics.
- Objects of classes have an explicit lifetime – they are explicitly created and destroyed by executing actions. Instances of data types have an implicit lifetime – they are created by executing actions, but are destroyed implicitly, when they are not referred to any more.
- Operations of classes may modify the objects' state. Operations of data types must not modify values of their attributes; they must be pure functions that possibly produce new values.

The semantics given in [14] is strict, clear, and unambiguous. The author has applied the profile and the rules in many industrial projects and has found the interpretation practically useful. However, it is not obvious whether the described two approaches are compatible and how they relate to each other. Although there are practical and intuitive rationales behind the definition given in [14], its formal background was still unclear.

All the described approaches agree that data types can have attributes too. Such data types are called *structured* in UML. Attributes of data types are considered to be immutable characteristics of their owner instances.

3.4. Summary

The presented discussions are summarized in Table 1 that gives an overview of the criteria used to discriminate classes vs. data types in classical conceptual modeling (including Beeri's guidelines), standard UML, and OOIS UML.

It is useful to mention briefly that a similar confusion exists in programming in classical object-oriented programming languages, such as C++ or Java. There, objects of classes (as the only kinds of supported entity types) have their inherent identity that is based on their very existence in time and space – the computer memory. Unrelated to their state, objects can be distinguished by their technical identifiers – references (pointers) that conceptualize objects' physical location in computer memory. However, programmers usually feel a need for what is clearly recognized in conceptual modeling and UML as a data type, so they introduce "special kinds of classes" that they usually call "immutable," rely on their "value-based equality comparison, copying, argument passing, etc.," and must deal with implicit lifetimes of data type instances or explicit lifetimes of objects (although the language sometimes supports only one of them).

As a conclusion, these two approaches (in classical conceptual modeling and in UML) to distinguishing between classes and data types on one hand, as well as between associations and attributes on the other, may seem unrelated or even contradictory, or at least vague and confusing at first sight. We will try to consolidate them and form a unified and consistent understanding of the notions. In fact, we will show that both approaches are actually correct and compatible, although not complete and fully formal, and will provide complete formalizations of all related notions.

Table 1. Summary of the discriminators of classes vs. data types used in classical conceptual modeling and UML

	Class	*Data Type*	*Issue*
Conceptual modeling	Variable?	Constant	Why cannot a class be constant as well?
	Not identifiable by a literal?	Identifiable by a literal	Why cannot a class be identifiable by a literal?
	Domain/application-specific	Universally known abstraction	A user-defined, domain-specific, structured data type is not a universally known abstraction.
	User-defined	Built-in	Both classes and data types can be built-in and user-defined. This is an orthogonal distinction.
	Interesting information carried by relationships	Interesting information carried by the value	Not true for structured data types.
Standard UML	Instances have identity	Instances do not have identity, but are pure values	What do “values” exactly mean?
	Equality by identity	Equality by structure/value	Definition for structured data types based on the unclear notion of attribute. The isomorphism of relationships/attributes unclear.
	Instances are generally mutable	Instances are immutable	No issues.
OOIS UML	Instances have identity	Instances do not have identity, but are values	What do “values” exactly mean?
	May take part in associations	May have attributes only	No issues.
	May not be types of attributes	May be types of attributes	No issues.
	Equality by identity (reference)	Equality by attribute values (recursively)	No issues.
	No implicit copy semantics	Implicit copy semantics	No issues.
	Explicit lifetime of instances	Implicit lifetime of instances	No issues.
	Instances are generally mutable	Instances are immutable	No issues.

4. Analysis

Before reaching precise definitions, we will carefully analyze the practical use and intents behind of what are traditionally referred to as classes and data types (including primitive and derived data types, structured data types, and enumerations).

4.1. Populations of Classes and Data Types

We deem that the central difference between different subkinds of entity types is the way how the population of an entity type is defined.

For entity types that we traditionally treat as data types, the population is typically defined *intensionally*. That is, the set of instances of a data type is defined by a formula or a set of rules or constraints that are necessary and sufficient conditions for belonging to the set. For example, the population of a primitive type *Integer* can be defined as the set of integer numbers that can be represented with a certain binary format, e.g., a 32-bit two's complement, which defines the set of integers in the range from -2^{31} to $+2^{31}-1$. Similarly, the population of the data type *String* is the set of all logically unlimited arrays of characters, including the empty array.

On the other hand, the population of a typical class is defined *extensionally*: for each particular instance of a class in its population, it is directly and explicitly stated that the instance belongs to the population, basically by executing a Create Entity or Reclassify Entity action; the instance belongs to the population of the class until it is destroyed or reclassified by another action.

4.2. Structured Data Types

The given simple example of data type *Integer* addresses only a subset of simple data types that are traditionally called *primitive* data types. These are types that are not defined in terms of other data types; they exist ab initio. *Derived* data types are those that are defined in terms of other data types. The way how (populations of) derived data types are defined predominantly depends on the language constructs and features, but we can list some of the most general approaches:

1) Defining a subset of another data type or of a union of other data types. The subset can be defined intensionally, by a rule, or extensionally, by enumerating instances that form the subset.

2) Through a set and functional calculus, by defining a data type whose instances are sets or functions of other data types, or with algebraic specifications of abstract data types [11, 12]. Such types include different kinds of collections, sets, bags, maps, and other non-atomic types used in different languages.

3) Defining structured data types by composition of other entity types.

In this paper, we will focus on and confine to the semantics of the third subcategory – structured data types only, as they are most typical derived data types used in conceptual modeling. There, a derived data type is defined in terms of relationship types with other entity types. The data type *Money* in Figure 1 represents one such example. Another simple and intuitive example is a data type *Point* with two relationship types (*Point*, *x*:*Decimal*) and (*Point*, *y*:*Decimal*) that define the coordinates of a point in a two-dimensional space.

Let us carefully consider the semantics of structured data types, actually, the intent of the modeler who defined them and the meaning she obviously had in mind. The populations of such entity types are defined intensionally, in terms of a set of some other entity types and relationships with them. For the two given examples, this looks like the following:

- For each pair of existing instances of *Decimal* and *Currency*, there exists exactly one (one and only one) instance of *Money* that is related to the two instances of *Decimal* and *Currency*.
- For each pair of two existing instances (possibly the same) of *Decimal*, there exists exactly one (one and only one) instance of *Point* that is related to the two instances of *Decimal* that play the *x* and *y* roles.

Optionally, in a more general case, such a definition may include a constraint that reduces the set of instances obtained by a simple Cartesian product of the related populations.

Consequently, unlike other relationship types in which a certain data type participates, such relationship types have a special meaning: they are used to define the population of the entity type by the (optionally reduced) Cartesian product of populations of the related entity types. Due to their special role in the definition of an entity type *E*, we will call such a set of relationship types the *generative set of E* and denote it with $\gamma(E)$. Consequently, the relationship types in $\gamma(E)$ are always constant with respect to *E*. For the example shown in Figure 1, the relationship types of *Money* with *Decimal* and *Currency* form the generative set of *Money*, because they are used to “generate” and identify instances of *Money*, and are therefore significantly different from other relationship types of *Money*, such as the one with *Product* whose instances do not affect the population of *Money*.

It can be easily concluded that if all other entity types that participate in the relationship types in $\gamma(E)$ are constant, *E* is also constant. However, one can imagine a more general case where there is a relationship type *R* in $\gamma(E)$ with a non-constant class *C* whose instances are dynamically created and destroyed by actions during runtime. The definition of *E*’s population remains the same, but the population is not constant any more. Namely, when an instance *c* of *C* is removed from *C*’s population, all instances *e* of *E* related to *c* by a relationship type from $\gamma(E)$ should also cease to exist, as their existence is defined in terms of their relationships with instances of *C*. Thus, the population of *E* changes over time, as the population of *C* changes. As a conclusion, an entity type *E* is constant if and only if $\gamma(E)$ relates *E* with constant entity types only.

Consequently, if the notion of the generative set is to be taken as the basis for defining populations of structured data types, then the previous conclusion calls for:

- either a restriction that $\gamma(E)$ for a structured data type *E* cannot relate *E* with a non-constant entity type; in that case, structured data types are constant entity types;
- or for a revision of the claim that data types have to be constant entity types; as just shown, if $\gamma(E)$ for a data type *E* relates *E* with a non-constant entity type, then *E* is also non-constant.

We will take the first approach in this paper in order to stay aligned with a commonly accepted conception that data types are constant entity types.

4.3. Constant Classes

In a traditional sense, as well as in UML, an ordinary, non-constant class specified in a conceptual model designates a set of objects that can potentially exist in the system's object space in runtime. This (typically infinite) set of potential objects is basically defined by intensional semantics, in terms of the features that objects of the class will have; all objects and only objects that have the features of a specified class are potential instances of that class. On the other hand, the actual set of objects that exist in the system's object space at a particular point in runtime is defined by extensional semantics, in terms of explicit Create, Reclassify, and Destroy Entity actions that have been executed up to that point. Such set of objects is always finite. In order to ensure clear disambiguation, we will refer to the former (the usually infinite set of potential objects defined intensionally) as the *extent* of the class or entity type in general, and to the latter (the set of actually existing objects at a certain point in runtime) as the *population*. One important corollary of this definition is that the population of a data type is always equal to its extent, that is, the population covers the entire extent and is constant. This is a fundamental observation and assumption that underpins the further ideas presented in this paper.

As opposed to a non-constant class, the extent and population of a constant class are the same, and are defined in design time rather than in runtime. In fact, this observation conceals a subtle potential inconsistence. Namely, objects of classes are created in an explicit, extensional manner, by executing Create Entity actions. Execution of actions is naturally bound to runtime. Hence, objects of constant classes cannot be "created in design time," but can only be "specified" (defined) in an extensional manner in design time, for example by so-called instance specifications of UML [16] (see, for example, creational object specifications in [14] as an advanced example). Anyhow, these specifications are manifested as actions being executed in runtime. Every successfully executed Create Entity action in turn modifies the population of the entity type. However, if a class is constant, it does not allow such actions. Therefore, totally constant classes, in this purely theoretical sense, actually cannot exist.

It seems that this inconsistence, however, is purely conceptual and does not have a significant practical impact. Constant classes may be defined as those whose instances are created in some strictly defined and controlled periods of runtime, e.g., at system's startup or by certain restricted administrative, initialization, or configuration procedures only.

An alternative resolution could be a relaxation of the notion of extensional population definition: an instance specification, made in a model and in design time, does not have to result in an action of creating an instance in runtime, but may be conceptually treated as a formal statement that such an instance (with its optionally specified relationships) merely exists, with the lifetime that has not its "creational beginning nor ending during runtime" but spans over the entire system's lifetime, and is a member of the population of its entity type ab initio.

This is just the case with enumerations: in essence, an enumeration is nothing more than a constant class whose instances can be referred to by literals specified in design time. Indeed, if an enumeration can take part in relationship types (either constant or non-constant with respect to itself) and can have other features (such as operations), there is not any semantic difference between a constant class and an enumeration, as both must have populations defined extensionally, in design time. This is really the case in some languages, such as Java, where an enumeration is nothing but a constant class with objects enumerated in the program and identified by literals. UML does not prevent enumerations from having relationship types and other features as well. However, in UML, "enumeration is a kind of data type, whose instances may be any of a number of user-defined enumeration literals" [16, p. 67]. Except from being influenced from some older programming languages, such as C/C++, and the semantics and implementation of enumerations in them, there seems to be no rationale behind the decision that enumerations are data types. We argue that this is yet another design flaw of UML that makes its semantics vague. From our perspective, the extensional nature of the definition of an enumeration's population is a much stronger qualifier that classifies it into classes, as opposed to the identifiability of its constant set of instances through literals, which is a characteristic applicable to constant classes as well.

In general, a constant class can take part in constant as well as in non-constant relationship types (with respect to itself); it seems that there is not a valid reason for any restriction in this respect.

5. Definitions

In this section, we provide a summary of the conclusions from the previous analysis in a form of precise definitions. Our definitions are based on the following observations described informally so far:

- Extents of both kinds of entity types are defined intensionally. Populations, on the other hand, are defined differently: extensionally for classes and intensionally for data types. This reflects our basic orientation to providing clear runtime semantics.
- Data types are constant entity types and their populations are equal to their extents.
- Populations of structured data types are defined as all allowed cross-combinations of different relationships of the relationship types from a generative set.

In the discussions that follow, we use the following notation:

- $\varepsilon(T)$ denotes the extent of an entity or relationship type T, that is, the (possibly infinite) set of all possible instances of T allowed by the model and constraints in it.
- $\pi(T)$ denotes the population of an entity or relationship type T, that is, the set of all existing instances of T at a certain moment in runtime; $\pi(T)$ obeys all implicit and explicit constraints defined by the model.
- $\pi(S)$ denotes the entire system's population, that is, populations of all entity and relationship types from the considered conceptual model at a certain (but the same) moment in time; the populations altogether obey all implicit and explicit constraints defined by the model.
- In symbols for entities, subscripts differentiate types of those entities, while superscripts differentiate instances within the same entity type.

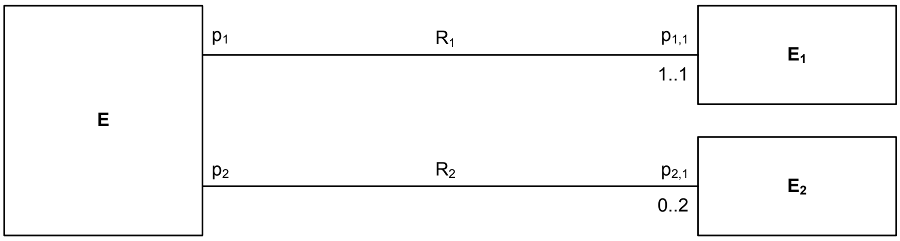


Figure 2. An example of an entity type E and its generative set $\gamma(E)=\{R_1, R_2\}$

Prior to giving a general definition of generative set, we first develop it for a simpler case with binary relationship types shown in Figure 2. The example shows an entity type E and its generative set $\gamma(E)$ consisting of two binary relationship types R_1 and R_2. There are also some cardinality constraints (called multiplicities in UML) associated with the relationship types, so that R_1 is total and functional with respect to E ($p_{1,1}$: $\pi(E)\rightarrow\pi(E_1)$), while R_2 is neither total nor functional ($p_{2,1}$ is multivalued and partial because of the 0..2

constraint). In addition, we may assume a set C of domain-specific constraints defined in the model, in a form of predicates that take entities as parameters.

The idea of the generative set $\gamma(E)=\{R_1, R_2\}$ is that there is exactly one $e \in \pi(E)$ for each valid combination of relationships with all existing entities of E_1 and E_2 so that all cardinality constraints and all predicates in C are also satisfied. For example, if there is only one existing instance e_1^1 of E_1 and two existing instances e_2^1 and e_2^2 of E_2, there are exactly four instances of E, related to instances of E_1 and E_2 as given in Table 2, assuming that C holds for all of them.[6]

Table 2. The population of entity type E defined by its generative set $\gamma(E)$ in Fig. 2

Entity of E	*Relationships of R_1*	*Relationships of R_2*
e^1	$\{<p_1{:}e^1>,<p_{1,1}{:}e_1^1>\}$	None
e^2	$\{<p_1{:}e^2>,<p_{1,1}{:}e_1^1>\}$	$\{<p_2{:}e^2>,<p_{2,1}{:}e_2^1>\}$
e^3	$\{<p_1{:}e^3>,<p_{1,1}{:}e_1^1>\}$	$\{<p_2{:}e^3>,<p_{2,1}{:}e_2^2>\}$
e^4	$\{<p_1{:}e^4>,<p_{1,1}{:}e_1^1>\}$	$\{<p_2{:}e^4>,<p_{2,1}{:}e_2^1>\}$, $\{<p_2{:}e^4>,<p_{2,1}{:}e_2^2>\}$

Just as another example, if the constraint at $p_{1,1}$ were 0..1, the number of instances of E would be eight: four of them would be related as shown in the given table, and the other four would be unrelated to any instance of E_1 and related to the instances of E_2 as shown in the table.

In our formal definitions, we introduce the notion of *relationship projection*: informally, this is a tuple obtained from a relationship by omitting a certain role. For example, the projection of the relationship $\{<p_1{:}e^1>,<p_{1,1}{:}e_1^1>\}$ from Table 2, for the role p_1, denoted with $\Pi(\{<p_1{:}e^1>,<p_{1,1}{:}e_1^1>\},p_1)$, is $\{<p_{1,1}{:}e_1^1>\}$, while $\Pi(\{<p_2{:}e^4>,<p_{2,1}{:}e_2^2>\}, p_2)= \{<p_{2,1}{:}e_2^2>\}$.

Similarly, the projection of an entity e for the given role p, denoted with $\Pi(e,p)$, is the set of projections of all relationships in which e participates with the role p. For the same example in Table 2, $\Pi(e^3,p_2)=\{\{<p_{2,1}{:}e_2^2>\}\}$, while $\Pi(e^4, p_2)=\{\{<p_{2,1}{:}e_2^1>\}, \{<p_{2,1}{:}e_2^2>\}\}$.

Finally, the projection of the given entity e for the entire ordered set of given roles $p_1, p_2, ..., p_n$, is an n-tuple consisting of projections for each role: $\Pi(e, <p_1, p_2,..., p_n>)=<\Pi(e,p_1), \Pi(e,p_2), ..., \Pi(e,p_n)>$. For the same example in Table 2, $\Pi(e^4,<p_1,p_2>)=<\{\{<p_{1,1}{:}e_1^1>\}\}, \{\{<p_{2,1}{:}e_2^1>\},\{<p_{2,1}{:}e_2^2>\}\}>$.

The purpose of the notion of projections is to allow for considering and comparing different possible and well-formed configurations of relationships

[6] We assume that the roles have the semantics of unique association ends in UML as discussed in [14]. If a role is non-unique, the result is slightly different, but the core idea remains the same.

for the generative set of a certain entity type E, independently of concrete entities of E that take part in them.

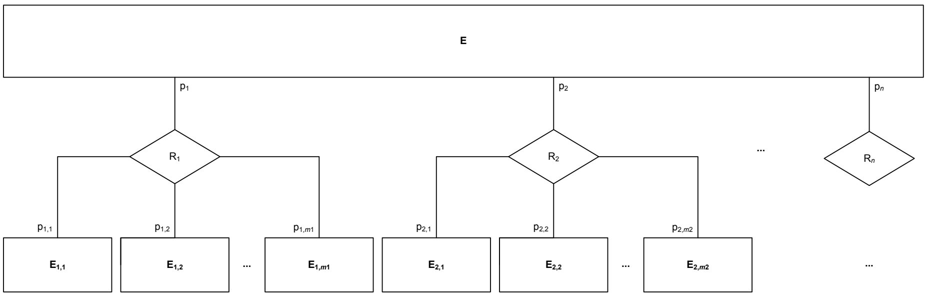


Figure 3. A general case of an entity type E and its generative set $\gamma(E)=\{R_1, R_2, ..., R_n\}$

Now we can get to general definitions. In the following definitions, these assumptions and notations are used (Figure 3):

- Let E be an entity type.
- Let $\gamma=\{R_1, R_2, ..., R_n\}$ be a finite set of relationship types such that E participates in each, where $R_i(p_i{:}E, p_{i,1}{:}E_{i,1}, p_{i,2}{:}E_{i,2}, ..., p_{i,mi}{:}E_{i,mi})$, $m_i \geq 1$, $1 \leq i \leq n$.

Definition (Projections). (See Figure 3 for assumptions and notation.)

- Let $r_i \in \mathcal{E}(R_i)$ be a relationship from the extent of R_i, $r_i=\{<p_i{:}e>, <p_{i,1}{:}e_{i,1}>, <p_{i,2}{:}e_{i,2}>, ..., <p_{i,mi}{:}e_{i,mi}>\}$, $e \in \mathcal{E}(E)$, $e_{i,j} \in \mathcal{E}(E_{i,j})$, $1 \leq i \leq n$, $j=1,...,m_i$. The *projection* of r_i for p_i, denoted with $\Pi(r_i, p_i)$, is the set:

  $\Pi(r_i, p_i)=\{<p_{i,1}{:}e_{i,1}>, <p_{i,2}{:}e_{i,2}>, ..., <p_{i,mi}{:}e_{i,mi}>\}$.

- Let $e \in \mathcal{E}(E)$ be an entity from the extent of E and p_i the role played by E in a relationship R_i, $1 \leq i \leq n$. The *projection* of e for p_i, denoted with $\Pi(e, p_i)$, is the set of projections of all relationships with existing entities of $E_{i,j}$ in which e plays the role p_i:

  $\Pi(e, p_i)=\{\Pi(r_i, p_i) \mid r_i=\{<p_i{:}e>, <p_{i,1}{:}e_{i,1}>, <p_{i,2}{:}e_{i,2}>, ..., <p_{i,mi}{:}e_{i,mi}>\},$
  $r_i \in \mathcal{E}(R_i), e_{i,j} \in \pi(E_{i,j}), j=1,...,m_i\}$.

- Let $e \in \mathcal{E}(E)$ be an entity from the extent of E. The *projection* of e for the roles p_1, p_2, ..., p_n played by E is an n-tuple consisting of projections for each role:

  $\Pi(e, <p_1, p_2, ..., p_n>)=<\Pi(e, p_1), \Pi(e, p_2), ..., \Pi(e, p_n)>$.

Definition (Generative set). (See Figure 3 for assumptions and notation.) γ is a *generative set* of E, denoted with $\gamma(E)$, if and only if (by definition):

i) In any system's population, there are no two different entities of E with the same projections for the roles of E in the relationship types from γ:

$$(\forall\pi(S))(\forall e^1, e^2\in\pi(E))(\Pi(e^1, <p_1, p_2,..., p_n>)=\Pi(e^2, <p_1, p_2,..., p_n>) \Rightarrow e^1=e^2).$$

ii) For any system's population, and for any possible well-formed configuration of relationships for γ, that is, for any legal projection, there is an existing entity of E in that population with that projection:

$(\forall\pi(S))(\forall e^1\in\varepsilon(E))$(relationships of e^1 satisfy all model constraints $\Rightarrow$
$(\exists e^2\in\pi(E))(\Pi(e^1,<p_1, p_2, ..., p_n>)=\Pi(e^2,<p_1, p_2, ..., p_n>)))$.

These definitions imply a function Π_γ that maps $\pi(E)$ on all possible projections of relationships allowed by the model and its constraints, defined as $\Pi_\gamma(e)=\Pi(e, <p_1, p_2,..., p_n>)$. The clause *i*) of the latter definition means that Π_γ is injective, while the clause *ii*) means that Π_γ is surjective.

Being both injective and a surjective, Π_γ is a bijection. Being injective, Π_γ is an observation term. As it is also a surjection, its corresponding identification function f_Π is total and is equal to Π_γ^{-1}. It can be shown that the statement *ii*) of the given definition implies that there is not any potential true extension of the population of E and of the populations of some or all of the relationship types from $\gamma(E)$ that satisfies all model constraints and preserves injectivity of Π_γ.

It should be noted that surjectivity differentiates Π_γ from an observation term in general. For the example in Figure 2, if $\gamma(E)$ were a simple observation term that is not also a generative set, then instances of E would not have to exist for each combination given in Table 2; for a generative set, this is mandatory (i.e., f_Π is total for a generative set). In addition, the relationship types from a generative set do not have to be reference types as defined in [17], because the participation of E in reference relationship types has to be total, while this is not necessary for a generative set; an example is the participation of E in R_2 in Figure 2 with the lower multiplicity bound equal to 0 at the opposite end.

Now we can formulate the definitions of different kinds of data types and classes.

Definition (Primitive data type). A *primitive data type* is a constant entity type without a generative set, whose population definition is intensional and whose instances can be identified by literals.

Definition (Structured data type). A *structured data type* is an entity type *E* whose population definition is intensional and has a generative set $\gamma(E)$, where any $R\in\gamma(E)$ has all participating entity types other than *E* constant.

Definition (Class). A *class* is an entity type without a generative set, whose population definition is extensional.

Definition (Enumeration). An *enumeration* is a constant class whose instances can be identified by literals.

We will illustrate the meaning of the given definitions, especially of the one of a generative set, on two examples.

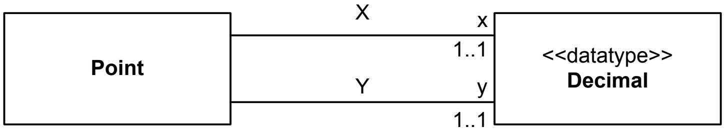
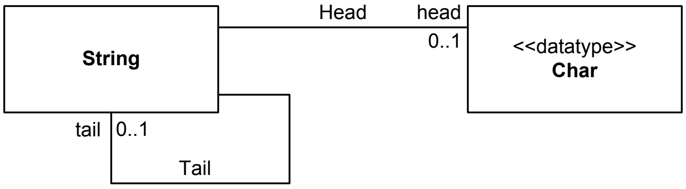


Figure 4. Example of entity type *Point*

Example (*Point*). The example in Figure 4 shows the entity type *Point* with two relationship types: *X(Point, x:Decimal)* and *Y(Point, y:Decimal)* with 1..1 cardinality constraints on *x* and *y*. If these two relationship types form the generative set for *Point*, then *Point* is a data type. In that case, the population of *Point* is defined intensionally and consists of one and only one instance of *Point* for each pair of existing instances of *Decimal* (it is basically isomorphic to the Cartesian product of *Decimal* with itself). On the other hand, if these two relationships do not have the semantics of a generative set, then *Point* is a class and its population is managed explicitly, by creating and deleting objects with actions executed in runtime. In practice, the modeler considers the case the opposite way: if she wants to have the semantics of an intensionally defined and constant population of *Point*, without having to manage its population in runtime by actions, she will declare *Point* as a data type, and the two relationship types will form the generative set of *Point*; otherwise, she will declare it as a class.



```
{context String inv:
    self.tail<>self and
    self.tail->size()+self.head->size()<>1}
```

Figure 5. Example of entity type *String*

Example (*String*). The example in Figure 5 illustrates another, more interesting case, where instances of entity type *String* can be recursively constructed from other instances of the same type. The constraint given below the diagram, along with the implicit constraints given in the multiplicities of the roles *head* and *tail*, specifies that only the following cases are allowed:

i) A *String* can have an empty head and an empty tail. This is the case of an empty string.

ii) A *String* can have one character as its head and another string as its tail. (Note that the tail can be an empty string.) This defines the recursive nature of constructing strings.

If the two relationship types are treated as a generative set and *String* as a data type consequently, there will be a logically infinite population of *String*, defined conceptually as follows:

- there is one and only one instance with no head and no tail, the "empty string;"
- for each existing instance of *Char* and each existing instance of *String*, there is another instance of *String* that is constructed by concatenating the character and the former string.

As before, in the opposite case when the two relationships do not have the semantics of a generative set, *String* is a class and its population has to be managed explicitly, by creating and destroying instances with actions. In that case, there can be two instances of *String* that are different identities, but which can be treated as equal strings with respect to their isomorphic structure, when they have the same character as their heads and (recursively) equal tails.

This example also suggests a general approach of defining other recursive structured data types, such as lists, trees, and others. It also shows the dual nature of those classical structures, since they can be treated as either classes or data types, depending on the intent of the modeler.

There are several interesting implications and properties that can be observed from the presented definitions:

- The extent and the population of a data type do not have to be finite. Since the definition of a primitive data type's population is intensional, it can be logically infinite.[7] The described example of the data type *String* is an example of a structured data type with an infinite population. The similar holds for a data type *BinaryLargeObject* (BLOB), which can be treated as primitive, and whose instances are finite but unlimited strings of bits.
- The population of a structured data type can be either finite or infinite, depending on the finiteness of the populations of other entity types that take part in its generative set, and on the way the generative set is defined.
- Due to the limitation that all entity types other than the considered structured entity type that take part in its generative set are constant, the extent and the population of the structured data type are equal, i.e., it is also constant.
- If all other entity types that take part in the generative set of a structured data type E are identifiable, then E is also identifiable through the observation term Π_γ. For example, in a typical (but not necessary) case when the relationship types in $\gamma(E)$ relate E only with primitive types, other identifiable structured data types, or enumerations, E is identifiable. E is not identifiable if, for example, it depends on a non-identifiable constant class.
- Due to its extensional definition, the population of a class is always finite, but generally unlimited.

Table 3. Summary of the main characteristics of different kinds of entity types

Entity type kind	*Population definition*	*Generative set*	*Population*	*Identifiable*	*Population finite*
Primitive data type	Intensional	No	Constant	Yes	Yes or No
Structured data type	Intensional	Yes	Constant	Yes or No	Yes or No
Constant class	Extensional	No	Constant	Yes or No	Yes
Enumeration	Extensional	No	Constant	Yes	Yes
Non-constant class	Extensional	No	Variable	Yes or No	Yes

[7] Typically, all such populations are physically finite, because they are limited by the amount of physically addressable computer memory or storage, but there is often no logical limit.

All these definitions and conclusions can be summarized in Table 3 and Figure 6.

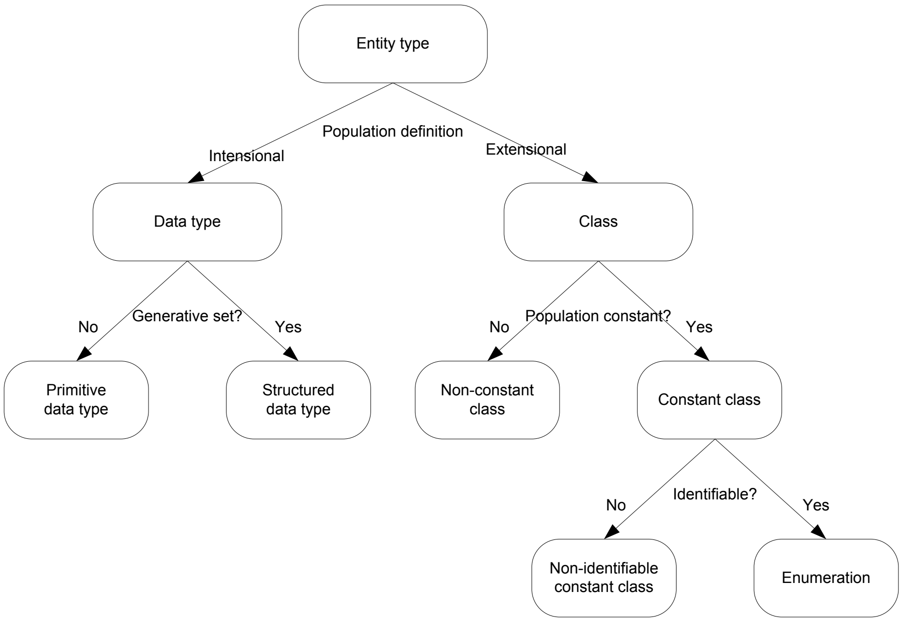


Figure 6. Classification of different kinds of entity types

6. Consequences

In this section, we analyze some consequences of the given semantic interpretations: how they relate to the definitions in UML, the value-based semantics, and the distinction between attributes and associations.

6.1. Classes vs. Data Types in UML

Populations of data types considered here are constant and defined intensionally. Therefore, in a certain sense, instances of data types (generally, of constant, intensionally defined entity types) can be considered as residing *outside* the scope of the running system, because the behavior of the system does not affect these populations and does not even define them explicitly. On the other hand, these populations are typically very large and sometimes unlimited. These are reasons why systems most often do not physically store instances of data types as individual physical entities referenced in relationships. Instead, a system internally uses *values* that are actually *identifiers* or *references* to those external instances, in all places such as relationships. For example, a 32-bit two's complement binary representation with all digits equal to one is a value that refers to the conceptual notion of minus one. Such values are then stored and manipulated wherever needed. Of course, this interpretation does not prevent an implementation to store each physical instance of a data type and use compact references to it in all other places, if this is cost-effective. The decision, of course, depends on different engineering trade offs but is certainly an implementation, not a conceptual concern. For example, if the space required for storing a reference or a key to the instance in a map is comparable to the space of storing the actual instance, the system may opt not to store physical instances. If, on the other hand, the system opts to store particular instances of a data type, when the population of a data type is huge or unlimited, the system may also opt to store only instances that are referred to from relationships with other entities. For example, if an implementation estimates that only a tiny subset of the entire extent of 64-bit integers will be referred to in the system, it can store only the used instances in a central repository and provide short integer keys to those values in a map; it is then much more cost-effective to manipulate with those shorter keys instead of "original" 64-bit values throughout the system.

These observations bring us to the rationale behind defining data types in standard UML as well as in OOIS UML [14] as described in Section 3. There, instances of data types are actually only *typed references* to external entities, but not real entities. The remaining semantic implications follow directly from that observation:

- "Instances of data types in UML do not have their identity, they are pure values; equal values represent the same instance. Instances of data types are compared for equality on value base." Indeed, values are typed references to entities and do not have identity themselves; instead, they refer to external entities with identity. Obviously, equal references refer to the same entity, because references are inherently unique identifiers.
- "Instances of data types can be arbitrarily copied and passed by value (e.g., as parameters)," again because they are simple typed references.
- "Instances of data types are immutable, and operations of data types cannot change any instance; instead, they are pure functions that can create and return new values." Really, values in UML serve as references to and proxies of external entities, while data types are constant entity types. Therefore, operations on such values are operations called on proxies. They cannot create new entities of constant entity types, but can create and return references to external, already existing entities.
- "As opposed to objects of classes, instances of data types have implicit lifetime: they disappear as soon as they are not needed any more" (this is a rule specific for the OOIS UML profile). Indeed, a reference (as an implementation thing) is just a handle that is used to access an entity; it can be discarded implicitly, as soon as it is not needed in its scope or

when its temporal scope expires. On the other hand, in UML and OOIS UML, instances of data types are created by a Create Instance action. This is, formally, a minor semantic flaw. As it implies from this discussion, when a "data type instance is created" in UML, only a reference to an external entity is created and obtained; that external entity has to be identified then. More concretely, in a textual notation, instead of using a statement e.g.:

```
Integer i = new Integer(5);
```

a more appropriate notation would be:

```
Integer i = Integer(5);
```

because this is actually not an action that creates a new instance of an entity type, but a request for a reference to an external entity identified by the parameters of the request ("constructor").

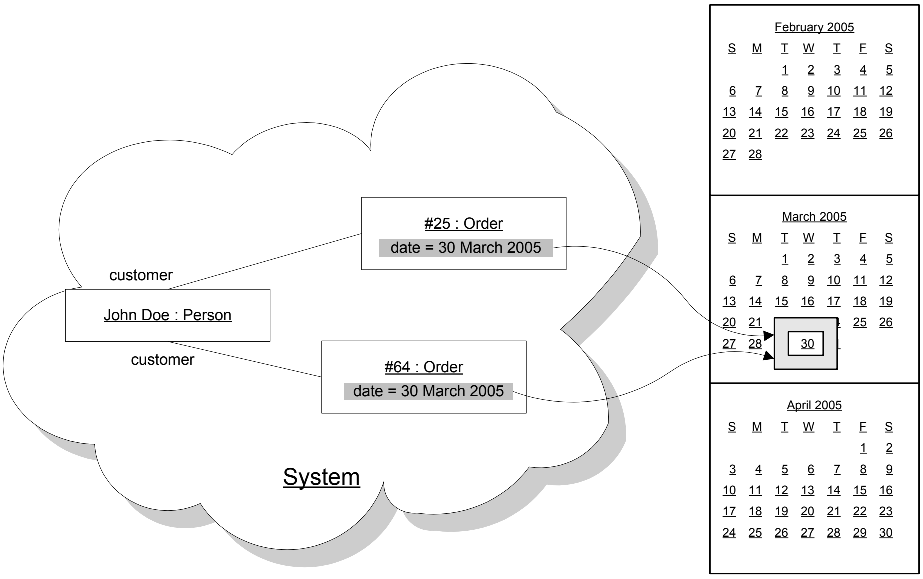


Figure 7: Instances of data types, such as *Date*, as defined in UML, represent values that refer to external entities that reside outside the system's boundaries

The decision whether an entity should be modeled with a class or with a data type is, thus, basically a matter of defining the boundaries of a system under consideration, as shown in Figure 7. If the system does not want to manage the population explicitly, but the population is defined outside the system's boundaries, then the entity type is a data type, while the values manipulated by the system refer to these external entities. For the example in Figure 7, objects of class *Order* are related with external entities of type *Date* that reside outside the system boundaries and abstract the common notion of a day. If, however, the system wants to define the population explicitly, by enumerating them or by creating and deleting them with actions, then the entity type is a class. For example, in an event management system, *Day* can be a class that conceptualizes the explicitly defined limited set of days of an event, possibly related to the external notion of *Date* in order to be properly positioned in the calendar and described additionally for human readability.

6.2. Value-Based vs. Object-Based Semantics

Following the same deduction in a more general context, value-based systems, such as relational data models, found their semantics on the fundamental assumption that real entities and relationships between them reside outside the running system and its data space, while the system manipulates data values that refer to those entities and relationships. More precisely, by storing and manipulating values and tuples of values, the system actually manipulates *facts* that certain external entities exist, are instances of their types, and are related with certain relationships. The existence of many occurrences of the same value or the same tuple of values in the system has no particular meaning, as they all represent replicas of the same fact that holds in the external domain. In some sense, the system does not affect the external domain with its execution, nor does it create or delete entities and relationships, but just tracks the facts from the external system and changes its scope of concern: at some time, some facts are relevant for the system and some are simply not. An absence of a fact (a value or a tuple) in the system does not necessarily mean that the thing referred to by that fact does not exist in the domain: it may not exist, or it may exist, but it is simply not relevant for the system or not captured by it at that time. This is why values and tuples do not have identities in such systems, and why modeling languages based on such semantics require that all entities be identifiable, as those systems have to maintain references (identifiers) to external entities.

On the other hand, object-based systems maintain entities that have their inherent identity and lifetimes internal to the system. Although such entities (objects) often *model* external concrete or abstract entities from the problem domain, like persons, things, or logical concepts, they also often represent pure abstractions invented in the implementation domain. The system maintains such entities by creating and deleting them, and by changing their states and properties, i.e., by creating and deleting relationships between entities. In such systems, entities have their inherent identities and do not have to be identifiable as in value-based systems. More detailed explanations on this aspect with examples can be found in [14].

It seems that the recent understandings and trends in the design of languages and systems for conceptual modeling and model execution, as well as in designing data models in practice even with the relational paradigm take the second approach. Indeed, it is a common practice in designing relational databases to introduce so called technical IDs as internally generated, outwardly invisible primary keys of records to ensure uniqueness of records in a certain table that cannot be identified by other means or whose "natural" (domain) identifiers are bulky. Such engineering technique is, actually, a reification ("objectization") of a tuple.

6.3. Attributes vs. Associations

The previous conclusions also justify the clear and precise discriminating characteristics of associations vs. attributes as defined in OOIS UML [14]:

- Associations are relationship types that can relate classes only; types of attributes can only be data types or enumerations.[8] This restriction clearly indicates that instances of attributes, i.e., attribute values, are references to external entities that reside outside the boundaries of the running system or to internal instances of enumerations whose populations are constant, as illustrated in Figure 4. Associations, on the other hand, model relationships (called links in UML) that are structural connections between objects, whereby the lifetime of both objects and links is managed by the system, i.e., by executed actions.
- Put the other way, (structured) data types can only have attributes, and cannot participate in associations. This rule clearly distinguishes the relationship types that form the generative set of a structured data type: they are modeled as its attributes, which fully fits into intuitive understanding of their purpose and meaning. On the other hand, the other relationship types with that same data type (that are not in its generative set) are modeled as attributes of other entity types: classes, when those attributes represent relationships that cross the system's object space boundaries, or attributes, when those attributes represent the generative set of another structured attribute whose definition is dependent on the former data type.
- The lifetime of an attribute value is bound to the lifetime of its owner object or another data type instance, while the lifetimes of objects are managed by the system and are generally independent (unless constrained by other specific model elements, such as propagated destruction [14]). Really, in this interpretation, attribute values are only references used within their owner objects to refer to external entities, i.e., relationships that cross the system's boundaries. This is why they are not needed when their scope ceases to exist. Attributes of data types actually model relationships that reside completely outside the system boundaries, so the system does not have to manage them through actions.

[8] In UML, as discussed previously, enumerations are classified as data types.

- Data types have implicit copy semantics; for an instance of a structured data type, a clone is a deep copy obtained by cloning the original's attribute values recursively. Again, this makes sense because attribute values of data types are just references to external relationships from the generative set and references can be copied.

All these rules explain that the support for classes and data types, as well as for associations and attributes in systems based on UML semantics, actually allows for mixtures of object-based and value-based semantics in the same system. The first given rule is the most important one. It seems to be a sounder rule for distinguishing associations vs. attributes than the rule that types of attributes should be entity types defined outside the model mentioned in Section 3. While in the latter rule it only matters whether an entity type is owned by the model of the considered system or is defined elsewhere, which is purely a matter of (static) model organization, packaging, and model compilation, it has nothing to do with the runtime semantics of the concepts. On the other hand, our rule is based completely on runtime semantics, which is, we believe, more sensible and appropriate for practical use, as it reduces confusion. Our experience shows that a modeling concept that has impact on the execution of the system is clear, easy to adopt, and unambiguous in use only if it has clear runtime semantics. We find this fact as the main advantage of programming languages over yet immature modeling techniques (without fully formal and executable semantics) and the main reason of a rather poor scale of adoption of model-driven development methods in industry.

7. Conclusions

We have provided a new interpretation of the semantic difference between classes and data types in conceptual modeling and UML. It is based on the difference between how populations are defined: extensionally for classes and intensionally for data types. Extensional definitions assume explicit statements that certain instances belong to the population, by enumerating instances or by executing actions in runtime. On the other hand, populations of data types are defined intensionally, by stating the conditions for belonging to the population, or describing the population implicitly. In the special case of structured data types, we have proposed the notion of a generative set, consisting of the relationship types that are used in defining the population of a structured data type: informally, the population is defined as a total and unique coverage of all allowed configurations of relationships from the generative set and for all existing entities of other entity types taking part in the generative set. We have also suggested that, according to this distinction, enumerations belong to classes rather than to data types. It has been suggested that in practice, the decision whether an entity should be modeled with a class or with a data type is, basically, a matter of defining the boundaries of a system under consideration, and whether the responsibility of managing populations of entity types is inside or outside the system's boundaries.

The proposed interpretation allows for unambiguous discrimination of the related concepts, yet it fits into intuitive understanding and common practical usage of these concepts. In addition, our interpretation provides a clear explanation of apparent discrepancies between definitions taken in conceptual modeling and in UML.

We have also described some semantic implications of the given interpretation that clearly distinguish value-based vs. object-based semantics, associations vs. attributes, and identity vs. identification. We strongly believe that such precise definitions, based on formal runtime semantics of the concepts, are crucial for practical use and wider adoption of model-driven software development techniques in practice. In fact, the described interpretations emerged from and have been verified in our practical experience in applying the described interpretations and model-driven development with an executable UML profile and its implementation in the SOLoist framework (www.soloist4uml.com), in many industrial projects.

One topic of interest that has not been covered in this paper is the formal semantic alignment of the notion of generalization/specialization (i.e., inheritance, subsumption) relationship between entity types with the interpretations given in this paper, in particular, the coherence of intensional definitions of populations via generative sets in presence of specializations. Our internal analysis shows that the notion of specialization fully accords with the given interpretations. However, a deeper study is left for some future work and publications.

Acknowledgments. The author is grateful to the anonymous reviewers for their very thorough and qualified reviews that have inspired him to improve the paper in several aspects, and clarify or correct it in many places.

References

1. Barbier, F., Henderson-Sellers, B., Le Parc-Lacayrelle, A.,Bruel, J.-M.: Formalization of the Whole-Part Relationship in the Unified Modeling Language. IEEE Trans. Software Engineering, Vol. 29, No. 5, 459-470. (2003)
2. Beeri, C.: A Formal Approach to Object-Oriented Databases. Data and Knowledge Engineering, Vol. 5, No. 4, 353-382. (1990)
3. Bourdeau, R. H., Cheng, B. H. C.: A Formal Semantics for Object Model Diagrams. IEEE Trans. Software Engineering, Vol. 21, No. 10, 799-821. (1995)
4. Chen, P. P.: The Entity-Relationship Model. ACM Trans. on Database Systems, Vol. 1, No. 1, 9-36. (1976)
5. Diskin, Z., Dingel, J.: Mappings, Maps and Tables: Towards Formal Semantics for Associations in UML2. In O. Nierstrasz et al. (eds.): Proc. of MoDELS 2006, Lecture Notes in Computer Science 4199, Springer-Verlag, 230–244. (2006)
6. Diskin, Z., Kadish, B.: Variable Set Semantics for Keyed Generalized Sketches: Formal Semantics for Object Identity and Abstract Syntax for Conceptual Modeling. Data & Knowledge Engineering, Vol. 47, No. 1, 1-59. (2003)

7. France, R. B.: A Problem-Oriented Analysis of Basic UML Static Requirements Modeling Concepts. In Proc. 1999 ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA '99), ACM SIGPLAN Notices, Vol. 34, No. 10, 57-69. (1999)
8. Génova, G., Llorens, J., Fuentes, J. M.: UML Associations: A Structural and Contextual View. Journal of Object Technology, Vol. 3, No. 7, 83-100. (2004)
9. Génova, G., Llorens, J., Martínez, P.: The Meaning of Multiplicity of N-ary Associations in UML. Software and Systems Modeling, Vol. 1, No. 2, 86-97. (2002)
10. Gogolla, M.: Identifying Objects by Declarative Queries. In Advances in Object-Oriented Data modeling, Papazoglou, M. P., Tari, Z., eds., MIT Press, 255-277. (2000)
11. Guttag, J.V.: The Specification and Application to Programming of Abstract Data Types. Ph.D. Thesis, Dept. of Computer Science, University of Toronto. (1975)
12. Guttag, J.V.: Algebraic Specification of Abstract Data Types. In Broy, M., Denert. E., (eds.): Software Pioneers, Springer (2002)
13. Mazur, B.: When is one thing equal to some other thing?. http://www.math.harvard.edu/~mazur/preprints/when_is_one.pdf, June 2007. (Retrieved December 2011)
14. Milicev, D. Model-Driven Development with Executable UML. John Wiley & Sons. (2009)
15. Milicev, D.: On the Semantics of Associations and Association Ends in UML. IEEE Trans. Software Engineering, Vol. 33, No. 4, 238-251 . (2007)
16. Object Management Group: UML Superstructure Specification, Version 2.4.1, http://www.omg.org. (2011)
17. Olivé, A.: Conceptual Modeling of Information Systems. Springer. (2007)
18. Övergaard, G.: A Formal Approach to Relationships in the Unified Modeling Language. In Proc. PSMT'98 Workshop on Precise Semantics for Modeling Techniques, Technische Universität München, TUM-I9803. (1998)
19. Övergaard, G.: Formal Specification of Object-Oriented Modelling Concepts. PhD Thesis, Department of Teleinformatics, Royal Institute of Technology, Stockholm, Sweden. (2000)
20. Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., Lorensen, W.: Object-Oriented Modeling and Design. Prentice-Hall International. (1991)
21. Rumbaugh, J., Jacobson, I., Booch, G.: The UML Reference Manual. Addison-Wesley. (2005)
22. Selic, B.: The Pragmatics of Model-Driven Development. IEEE Software, Vol. 20, No. 5, 19-25. (2003)
23. Selic, B., Ramackers, G., Kobryn, C.: Evolution, Not Revolution. Communications of the ACM, Vol. 45, No. 11, 70-72. (2002)
24. Stevens, P.: On the Interpretation of Binary Associations in the Unified Modeling Language. Software and Systems Modeling, Vol. 1, No. 1, 68-79. (2002)

Dr. Dragan Milićev is Associate Professor and Chairman of the Software Engineering Department at the University of Belgrade, Faculty of Electrical Engineering (www.etf.rs). He is specialized in software engineering, model-driven development, UML, object-oriented programming, software architecture and design, information systems, and real-time systems. He has published papers in some of the most prestigious scientific and professional journals and magazines, contributing to the theory and practice of model-driven development and UML. He is also a member of the program committees of two premier international conferences on model-based engineering (MODELS and ECMFA). He is the author of three previous bestselling books on object-oriented programming and UML, published in Serbian, and a recent book in English, published by Wiley/Wrox, “Model-Driven Development with Executable UML” (also publish in Chinese by Tsinghua University Press, Beijing, China). With more than 25 years of extensive industrial experience in building complex commercial software systems, he has been serving as the chief software architect, project manager, or consultant in a few dozen international projects, some of them having national scope and importance. He is the founder of SOL Software (www.sol.rs), a software research and development company specialized in designing software development tools using model-driven approach, as well as in building custom applications and systems. He was chief software architect and project manager for most of SOL’s projects and all its products.

Received: July 16, 2011; Accepted: February 27, 2012.

DOI: 10.2298/CSIS110823010A

Specification of Data Schema Mappings using Weaving Models

Nenad Aničić[1], Siniša Nešković[1], Milica Vučković[1] and Radovan Cvetković[2]

[1]Faculty of Organizational Sciences,
Jove Ilića 154, 11000 Belgrade, Serbia
{nenad.anicic, sinisa.neskovic, milica.vuckovic}@fon.bg.ac.rs
[2]Telekom Srbije A.D., Technical Affairs Division
Bulevar umetnosti 16a, 11000 Belgrade, Serbia
radovan.cvetkovic@telekom.rs

Abstract. Weaving models are used in the model driven engineering (MDE) community for various application scenarios related to model mappings. However, an analysis of its suitability for specification of heterogeneous schema mappings reveals that weaving models lack support for mapping rules and, therefore, cannot prevent mapping specifications which are semantically meaningless, wrong or disallowed. This paper proposes a solution which overcomes the identified open issue by providing the explicit support for semantic mapping rules. It is based on introduction of weaving metamodels augmented with constraints written in OCL. The role of OCL constraints is to restrict mapping specifications to only those which are semantically meaningful. Using well known MDE technologies, such as EMF and QVT, an existing tool is used to validate the presented solution. This solution is also successfully evaluated in practice.

Keywords: schema mappings, weaving models, model transformations

1. Introduction

Specification of mappings among heterogeneous schemas[1] has been studied in many different research areas, such as distributed databases [1], data warehouses [2], ontologies [3], model driven development [4], [5], [6], etc. According to specific needs and characteristics of a particular problem domain, researchers have proposed different approaches and techniques that can be used to specify schema mappings.

Without diving into details of each particular approach, it can be generally concluded that most of them rely on a mapping specification formalism,

[1] The term schema is used in a broader sense and includes database schemas, ontologies, or generic models.

embodied in the form of either a special language or metamodel, which enables expressing and capturing the semantics of correspondences among schema concepts. In this paper we deal with the problem of finding a suitable formalism for the specification of schema mappings.

In accordance to the motto that "models are everywhere", the corresponding mapping specification formalism in the context of the model driven engineering (MDE) utilizes (meta)models. One particular solution which has been proposed is based on so called weaving models [5], [6], [7]. A weaving model is a separate model on its own consisting of elements which represent individual links (i.e. correspondences) among elements of other models (called woven models). A weaving model conforms to a weaving metamodel, which provides the semantics of links specified in a weaving model. A weaving metamodel defines types of links that can occur among elements of woven models, i.e. links specified in a weaving model can be instances only of types defined in the weaving metamodel. A special core weaving metamodel with generic link types suitable for a range of different application scenarios is also proposed. For each application scenario, the core weaving metamodel is extended with specialized link types that are more suitable for the particular application. Supported by the ATLAS Model Weaver (AMW) toolkit [8] within the Eclipse Modeling Framework (EMF) environment, the proposed approach has gained a lot of attention in the MDE community lately. It has been reported that weaving models are successfully applied to several MDE related problems, including schema and data mapping problems [5].

However, our experience in the application of weaving models to the problem of schema mappings reveals that there exist some open issues. Namely, the definition of a weaving metamodel is based only on concepts of metameta model (i.e. Ecore metameta model in EMF). It does not rely on concepts of corresponding metamodels of models intended to be woven. Hence, it is completely unaware of any semantic rules regarding mappings among concepts of the metamodels in question. As a consequence, link types defined in a weaving metamodel cannot prevent links between elements of woven models which are semantically meaningless, wrong or disallowed. For example, when mapping concepts between an entity-relationship (ER) data schema S1 and a relational schema S2, it is possible to link an entity from the S1 schema with a column from the S2 schema. In other words, the weaving model lacks the semantics of mapping rules between ER and relational schemas, i.e. that entities can be mapped to relational tables only.

This paper proposes a possible solution which overcomes the identified open issues by providing explicit support for mapping rules. The solution is based on a weaving model which serves for definition of mapping rules between schema metamodels. This weaving model is then transformed to a weaving metamodel augmented with Object Constraint Language (OCL) constraints. The role of OCL constraints is to restrict links in weaving models to establish only those relationships between schema concepts which are meaningful.

The proposed solution is successfully tested in practice on several different schema mapping problems. This is supported by the set of software tools consisting of open-source plug-ins for model weaving (AMW [8]), model transformation (M2M QVT Operational) and model validation (Eclipse OCL).

The paper is organized as follows. The next section briefly presents the related work. Section 3 introduces a general conceptual framework, which is used to identify types of models that occur in the context of data schema mappings and to define their roles and mutual mappings. Section 4 analyses the existing practice in utilizing weaving models and discusses shortcomings and open issues. In Section 5 the proposed solution is explained and discussed. In the same chapter, a technical realization of the suggested approach is provided as well as an illustrative example. Conclusion is given in Section 6.

2. Related Work

Schema mappings are high-level specifications which express correspondences between elements in heterogeneous schemas. Although schema mappings have been studied independently in different contexts, there are two main issues involved:

- The first one is to discover the correspondences between schema elements that are semantically related in the source and target representations. This process of discovering correspondences is called schema matching;
- After the correspondences have been established, the second issue is to produce operational mappings that can be executed to perform the data exchange between source and target.

Many techniques have been proposed to discover the correspondences between schemas. One particular approach is implemented in Clio [9], which generates semi-automatic mappings between schemas based on value correspondences obtained from the user or by a machine learning technique. Furthermore, Clio supports the definition of nested mappings, but this definition cannot be extended. This makes it difficult to create complex mappings. Another method, presented in [10], [11] is called Similarity Flooding (SF). Schemas are presented as directed labeled graphs and this structural method propagates the similarity of a pair of nodes through their neighbors. However, this approach cannot be used to define a correspondence between schemas (models) conforming to different schema languages (metamodels). There are a number of other different approaches for automatic schema matching. Their thorough survey is given in [12], but they are not discussed here as our focus in this paper is on manual schema specifications.

Schema operational mappings typically incorporate the data transformation semantics required to specify how data from one source representation (e.g. a relational schema) can be translated to a target representation (e.g. an XML schema). They are also used for virtual information unification where users can pose queries over distributed heterogeneous data sources in a uniform and transparent manner.

In the case of information integration with heterogeneous schema languages, several approaches have been proposed [13], [14], [15]. These approaches are mostly based on a generic metamodel that abstracts concrete schema metamodels. Schema mappings in this case are specified in a language which depends on the chosen generic metamodel. For instance, in [13] a universal metamodel based on the supermodel is used as a generic metamodel and DATALOG is used for representing schema mappings. In [15], the GeRoMe model and corresponding specification language are used.

In the context of MDE, specifications of schema mappings and data mappings can be viewed as a special case of model mappings. Another special case of model mappings are model transformations, which represent a crucial notion in MDE. The MDE community has proposed several model transformation specification languages. For instance, OMG has defined the QVT language [16], ATL is used in the EMF environment [17], etc. Although transformation specification languages could be used for data schema mappings, they are not designed for such a task. Transformation specification languages are used to specify mappings between metamodels (M2 level models) and, consequently, are inconvenient for the specification of schema mappings, which are M1 level models.

Model weaving is an approach used for establishing fine-grained correspondences between model elements. It is supported by AMW, a component of the larger Atlas Model Management Architecture toolkit [8]. This approach is conceived with a goal to facilitate a range of different application scenarios, such as tool interoperability, model composition operations, traceability, model alignment, etc [8]. One group of supported application scenarios is related to schema mappings. The work from [5] presents the application of weaving model to discover model mappings and production of executable operational mappings (including model transformations) which translate from source models to targets. These results are extended by the work in [7], which utilize successive schema matching transformations to generate and refine a sequence of weaving models until a final one is generated, out of which data transformations are produced. Weaving models are created using different methods. In this paper we not deal with matching heuristics or algorithms which can be used to automatically create weaving models. We created weaving models manually by graphical user interfaces. The created weaving model may be later used by a model transformation language to translate source data model(s) into target data model(s).

3. Conceptual Data Integration Framework

In order to analyze suitability of the weaving modeling approach for the specification of schema mappings, we introduce a conceptual data integration framework, shown in Figure 1. The main purpose of the introduced framework is to identify kinds of models that occur in the context of data schema mappings and to precisely define their roles and mutual mappings.

The framework has 4 abstract levels, which correspond to the levels of OMG MDA standard [4], but are named here in the manner that is more appropriate for data integration purposes. As it is typical for metamodeling architectures, each level accommodates models which serve as metamodels for other models from the lower abstract level, whilst they must conform to their metamodels from the upper level. The exception is the model at the most abstract level which conforms to itself.

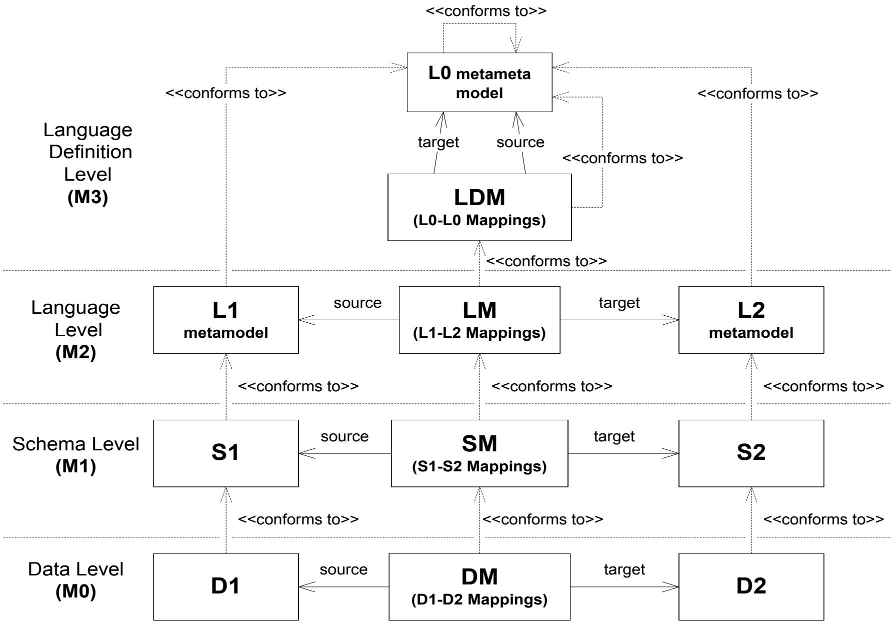


Fig. 1. Conceptual data integration framework

The framework identifies two types of models: (1) ordinary models which are used to describe domain concepts, and (2) mapping models which links elements from other ordinary models[2]. Depending on the abstract level where they reside, the following four kinds of mapping models can be identified:

– *Data Mapping Models* (DM) which specify links at the Data Level, i.e. between data instances stored in different possibly heterogeneous data sources. This level coresponds to M0 level of MDA.

[2] Mappings between different mapping models are also possible, but their consideration is beyond the scope of this paper.

- *Schema Mapping Models* (SM) which specify links at the Schema Level, i.e. between schema concepts that are possibly expressed in different schema languages. This level coresponds to M1 level of MDA.
- *Language Mapping Models* (LM) which specify links at the Language Level, i.e. specify mappings between concepts of different schema languages. This level coresponds to M2 level of MDA.
- *Language Definition Mapping Models* (LDM) which specify links at the Language Definition Level, i.e. between concepts of a metameta model used to describe schema languages. This level coresponds to M3 level of MDA.

It is important to note that mapping models must conform to their corresponding metamodels, which are also mapping models. This means that links specified in one mapping model must be instances of links specified in its mapping metamodel. In other words, links specified in the upper abstract level constrain links in the lower level to relate only certain types of model elements. Thus, links serve as mapping rules for the lower level allowing only meaningful links and preventing invalid ones.

The example shown in Figure 2 illustrates this for the case of mappings between a relational schema and an XML schema. As it is depicted in the figure, table Publication is mapped to XML element Book by a link which is an instance of the rule mapping relational tables to XML elements, whereas column ISDN is mapped to XML element BookID by a link which is an instance of the rule mapping columns to XML elements.

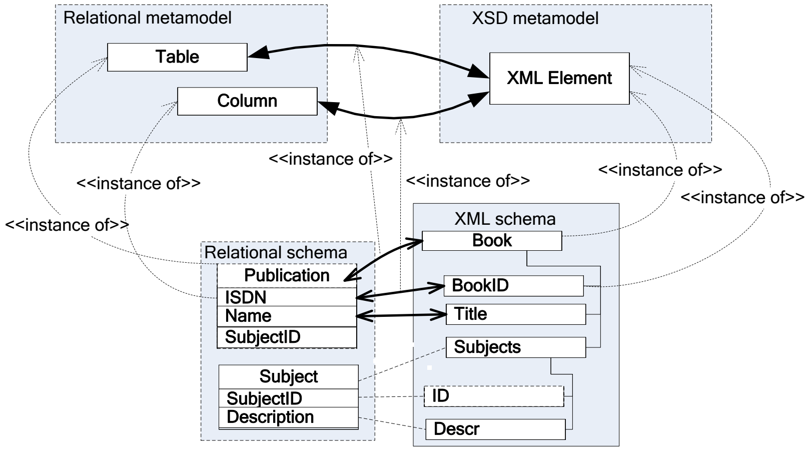


Fig. 2. Links at two different abstract levels

A special case is LDM, which is used to define rules for mappings between schema languages. Since it represents the most abstract mapping model in the framework, it is defined in terms of concepts of a corresponding metameta model *L0*.

4. Open Issues

Model mappings based on weaving models represent an approach in MDE which is supported by the AMW toolkit [17], [18]. AMW is a tool for establishing relationships (i.e., links) between model or metamodel elements and it is aimed to support a range of application scenarios where model mappings are involved. Here, we will discuss the approach from the schema mappings perspective only.

AMW supports an extensional mechanism based on the core weaving metamodel that encompasses a set of features (i.e. generic concepts) common in majority of application scenarios. Using extensions one defines a new weaving metamodel based on the core weaving metamodel. A new weaving metamodel typically defines new link types which are specific to a particular application scenario. Defined link types are used in weaving models for relating elements of two woven models.

Using the conceptual data integration framework given in Section 3 as a reference model, the typical application scenario employed by the AMW approach is shown in Figure 3. Here, the Ecore metamodel of EMF plays its usual role of the most abstract models *L0*. The role of LDM model in the conceptual framework is played by the core weaving metamodel, which is defined in terms of Ecore concepts. The role of LM, used to define mapping rules between different schema languages, is played by a weaving metamodel. It is defined as an extension of the core weaving metamodel. Note that this is different in comparison to our conceptual framework where LM is defined as an instance of a metamodel. In addition, a weaving metamodel in AMW does not specify mappings between concepts of a language, but simply defines a new link type. In other words, the semantics is provided only by giving a new name, without specifying schema concept types allowed to be related by this link type.

Specification of the mapping rules is given as an extension of the core weaving metamodel (Figure 3), which is an abstract metamodel. Therefore, the specification of the mapping rules is given at the language level and schema level separately, using LDM1 and LDM2. LDM1 specifies mapping rules (types of links) between elements of metamodels (i.e. language concepts) which are usualy simple equality links (e.g. *ElementEqual, AttributeEqual*). LDM2 specifies mapping rules for links typically expressing equality or set-inclusion relationships (e.g. *EqualLink, NestedLink, ChildLink*) which disregard types of linked schema elements. In this case, SM mappings between two concrete schemas conform to LDM2 mapping rules and do not conform to the defined LM mapping rules between the corresponding metamodels.

Of course, one can use the same weaving metamodel for specification of both LM and SM mappings. In this case LDM1=LDM2, i.e. they are the same weaving metamodel describing common link types (such as equality links) which are generally applicable to elements of different model types and at several abstract levels of our conceptual framework.

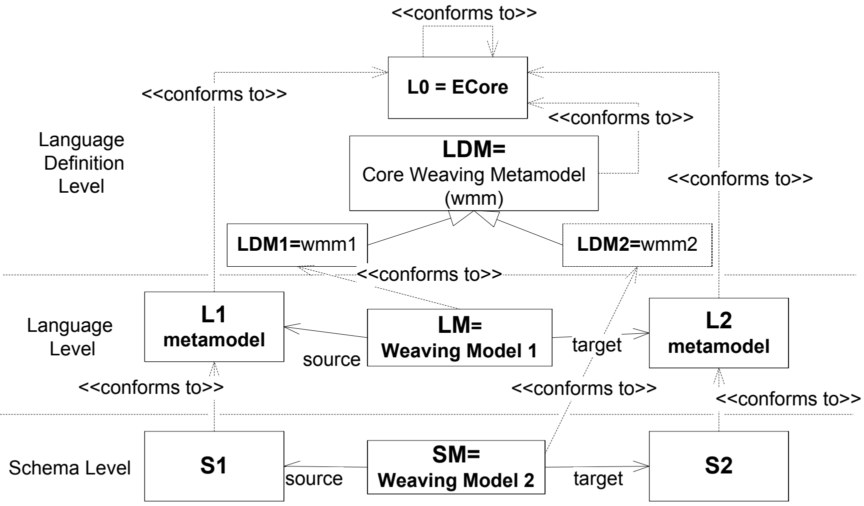


Fig. 3. The AMW approach

The role of SM, used to specify schema mappings, is played in AMW by a weaving model. It is defined as an instance of its corresponding weaving metamodel, which is in accord with the conceptual framework. However, links specified by a weaving metamodel can relate any elements from woven models. It is up to a modeler to take care whether such links are meaningful.

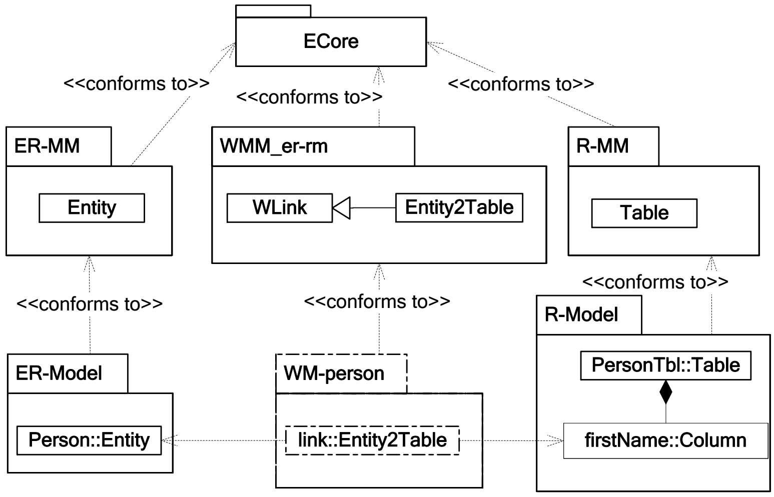


Fig. 4. A semantically invalid link in a weaving model

Figure 4 illustrates the situation that can happen when a modeler is careless or unaware of semantic mapping rules. The rule *Entity2Table* is specified in the weaving metamodel meaning that entities from ER models are translated to tables in relational schemas. However, as it is shown in the figure, it is possible to create a link which is an instance of the defined rules, but maps an entity to a column.

The Data level from the conceptual framework is not actually supported by AMW. However, application scenarios involving data instances still can be supported by lifting of models from the Data and Schema levels for one abstract level up, i.e. by expressing and treating data schemas as metamodels and data instances as M1 models. Such technique is employed in [5]. However, it leads to weaving models that must make up for a lack of models from the Language level, which is lost due to the level lifting. This technique introduces the accidental complexity to the definition of weaving models. Due to limited space, the further detailed discussion of this technique is beyond the scope of this paper.

To summarize the discussion, the following shortcomings and open issues exist:

- Weaving models may contain an invalid specification of schema mappings.
- Weaving models are not adequately constrained by their corresponding weaving metamodels.
- Link ends, which are part of link type definitions, cannot be typed, i.e. specified to which concept types they may relate.
- Data modelers are supposed to know the semantics of mapping rules and must take care about links they create.

5. Solution

The open issues discussed above can be resolved by a better alignment of the AMW approach with the conceptual data integration framework defined in Section 3. Better alignment in the context of schema mappings primarily means that link types defined in weaving metamodels have to constrain links in weaving models to relate those schema concepts that are meaningful. In other words, AMW approach needs to be extended with support for specifications of mapping rules in weaving metamodels which would enable typed end points of links in weaving models.

This paper proposes a solution to this problem by introducing several special weaving metamodels and models which have specific roles. The proposed solution is given in Figure 5.

One special weaving model having LM role is used to specify mapping rules by establishing semantically meaningful links between concepts from two schema languages. Due to constraints of EMF environment and AMW tool, this weaving model cannot be simultaneously used as a metamodel for weaving models from the Schema level. Hence, this weaving model is transformed by a model transformation into another special weaving metamodel augmented with OCL constraints. OCL constraints are integral parts of link type definitions and they specify types of data schema concepts that can be related by a particular link type. In this way, end points of links in weaving models are allowed to reference only instances of the specified types.

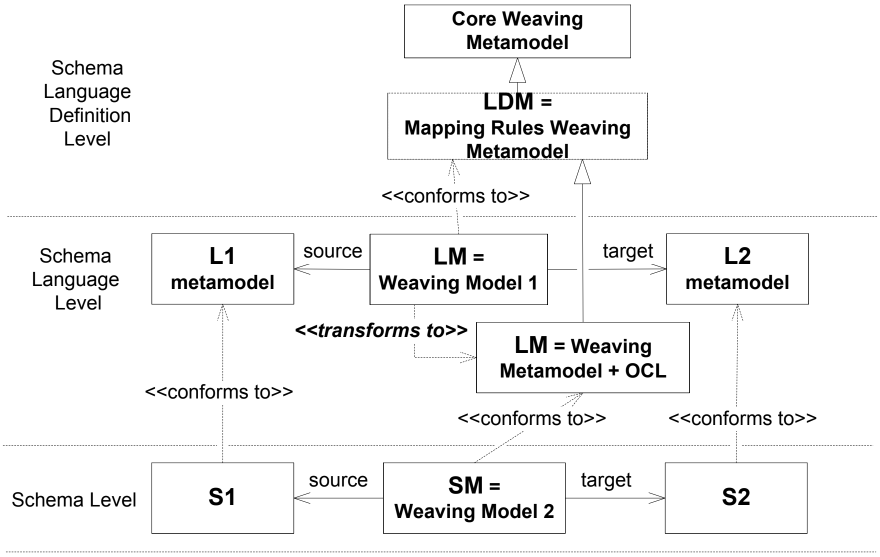


Fig. 5. Proposed solution

It is worth of observing that both the special weaving model and the generated weaving metamodel contain same information about mapping rules. However, the mapping rules are expressed in different ways in these models. In the weaving model they are expressed as structural constrains via (weaving) links, whilst in the weaving metamodel they are expressed via OCL as value based constraints. Therefore, both models have the same LM role defined in the conceptual framework in Section 3. In addition, both models are related to the same weaving metamodel. But unlike the AMW approach where they are related to the core weaving metamodel, here a new special weaving metamodel playing the LDM role is introduced. Also, they are related in a different way. The LM weaving model conforms to the LDM mapping rules weaving metamodel, whilst LM weaving metamodel is defined as its extension.

In order to describe and illustrate the proposed solution in more detail, in the rest of this section we first present LDM mapping weaving metamodel. Then, we illustrate the definition of mapping rules between concepts of ER and relational model using LM weaving model. Afterwards, we describe the model transformation from the LM weaving model into the LM metamodel augmented with OCL constraints. In the second section, we provide an example of specifying mappings between two concrete data schemas expressed in ER and relational model using the SM weaving model adhering to the mapping rules given in the generated LM metamodel. At the end of the section we summarize our apporach and describe its technical realization.

5.1. LDM Mapping Rules Waving Metamodel

LDM mapping rules weaving metamodel is defined as an extension of abstract core weaving metamodel (WMM) within AMW toolkit [8]. Both models are shown in figure 6 using UML class diagrams.

WMM model, which is shown in *mwcore* UML package in figure 6, defines abstract concepts used to specify mappings (class *Wlink*) between elements of some models (class *WlinkEnd*), as well as for referencing these models and their elements (classes *WModelRef* and *WElementRef*).

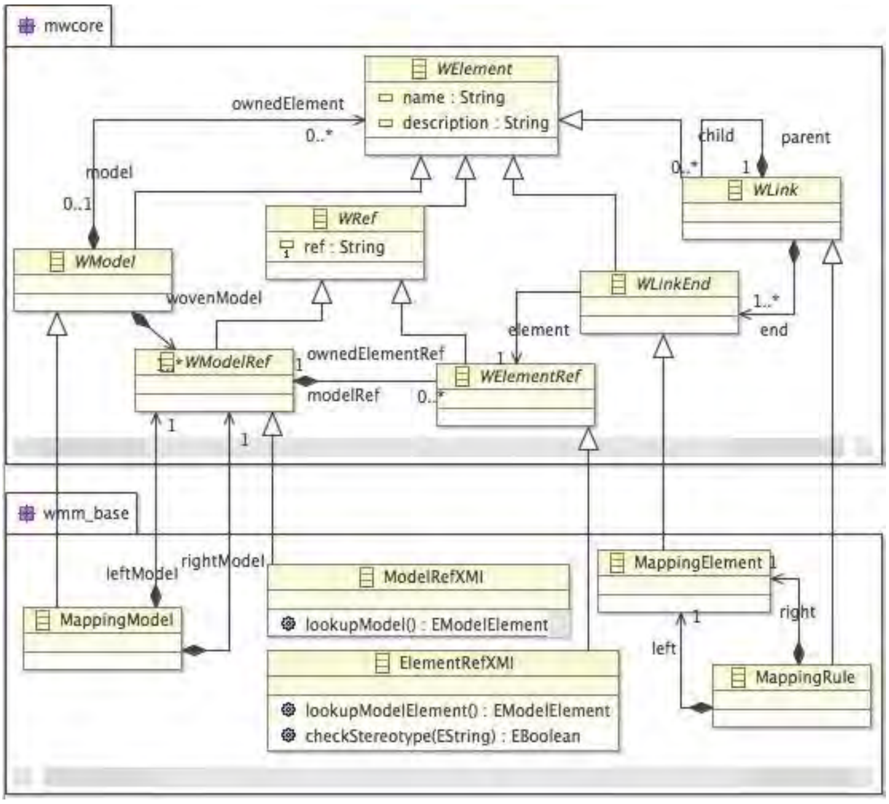


Fig. 6. LDM mapping metamodel

The definition of LDM mapping rules metamodel is given in the package *wmm_base*. It is assumed that mapping rules relates concepts of two different schema languages are represented as metamodels via standard UML class diagrams. *MappingRule* and *MappingElement,* modeled as subclasses of the concepts *Wlink* and *WlinkEnd* respectively*,* are used to define mapping rules between elements of these two different metamodels. References *left* and *right* enable to define a one-to-one mapping of elements, (reference left points to this) from one metamodel to an element (reference right points to this) of the other metamodel. *ModelRefXMI* and *ElementRefXMI,* modeled as subclasses of the concepts *WModelRef* and *WElementRef* respectively, represent references to UML models and elements serialized in XMI format. *ModelRefXMI has an* operation named *lookupModel*, which returns some UML2 model. (UML2 is an EMF-based implementation of the Unified Modeling Language 2.x OMG metamodel for the Eclipse platform). *ElementRefXMI* has two operations: operation *lookupModelElement* which returns an element of the UML2 model based on a reference, and operation *checkStereotype* which checks the applied stereotype of the UML2 model element to which the given reference points. These operations are introduced to simplify the definition of OCL constraints and queries in concrete weaving models.

5.2. An example of mapping rules

This subsection presents an example of mapping rules between concepts of ER and relational schema languages. Simplified metamodels for these two schema languages are shown in figures 7 and 8 respectively.

ER metamodel (fig. 7) includes the concepts of *Entity*, *Attribute*, *Relationship* and *Role*. *Entity* has an arbitrary number of *Attributes*. *Relationship* represents a binary association between two *Entities.* Each Entity participates in a *Relationship* with some *Role*.

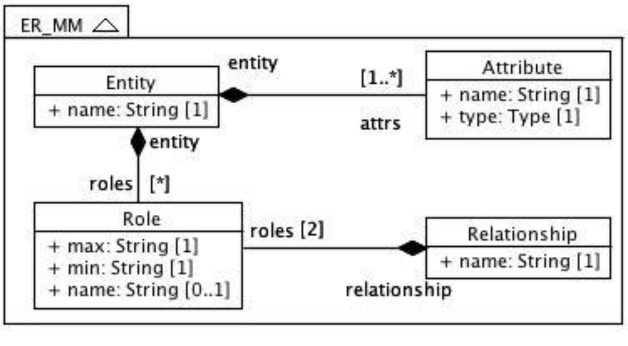


Fig. 7. Simplified ER metamodel

Relational metamodel (fig. 8) includes the concepts: *Table*, *Column* and *Key*. Each *Table* contains *Columns.* One or more columns of some Table represent its *Key*. *ForeignKey* and *PrimaryKey* are modeled as specializations of the concept *Key.* We use the reference refFK to connect a value of the foreign key in a table with a value of the primary key in some other table.

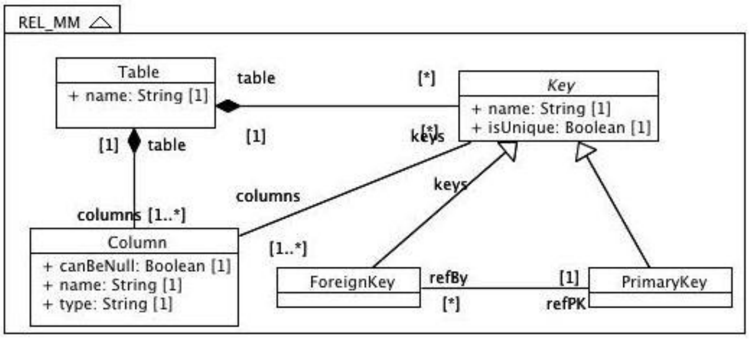


Fig. 8. Simplified relational metamodel

To define mapping rules between the ER and relational metamodel, we create a weaving model with LM role which is an instance of LDM mapping rules metamodel defined in the previous subsection. This weaving model is represented here as an UML object diagram shown in figure 9. It is an instance of the class diagram form figure 6. The LM weaving model in our example defines several mapping rules between concepts of data schemas, such as *Entity2Table*, *Attribute2Column* and *Relationship2FK.* These mapping rules constrain which concepts of ER data model can map to which concepts of relational data model. They are defined as instances of *MappingRule* class from LDM mapping rules metamodel.

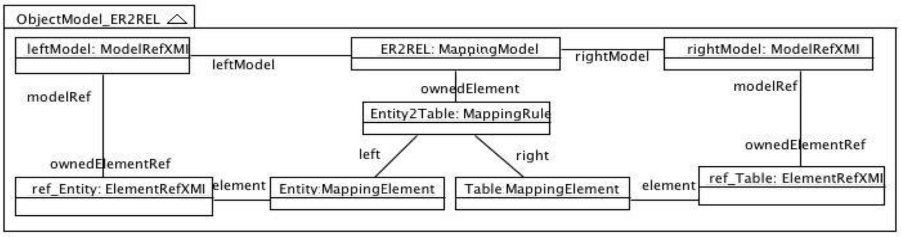
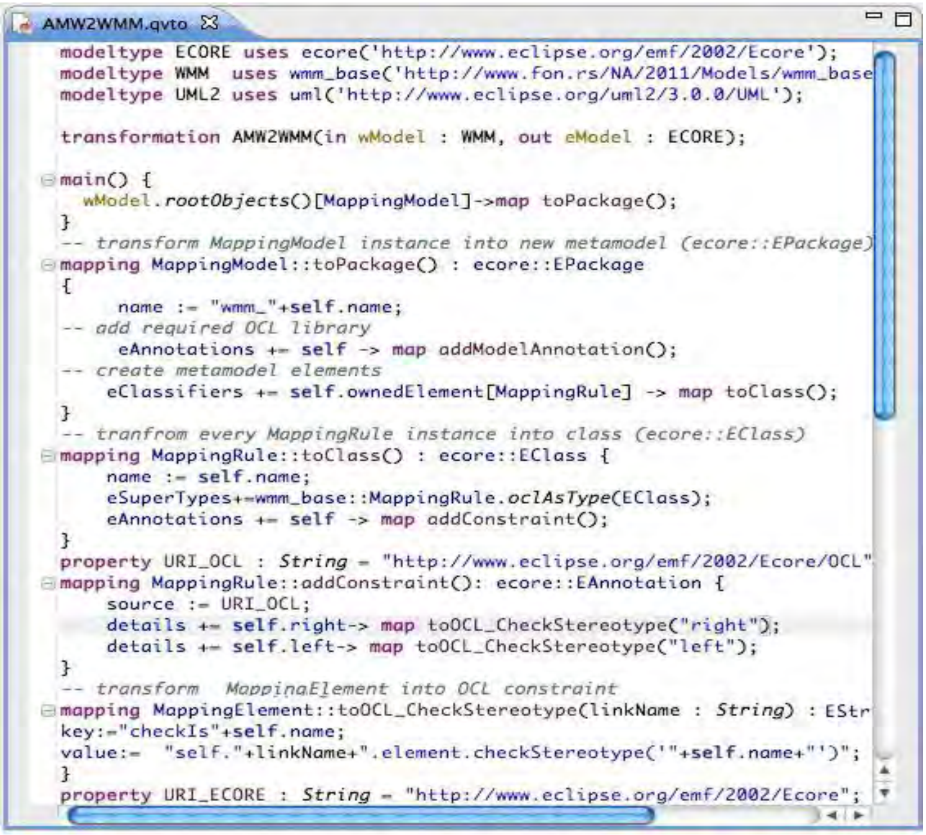


Fig. 9. An example of LM mapping model

5.3. Model transformation

The model transformation from LM weaving model to LM weaving metamodel is expressed using OMG's standard QVT Operational (QVTo) model transformation language [16]. Its implementation in EMF environment is depicted in figure 10.



```
modeltype ECORE uses ecore('http://www.eclipse.org/emf/2002/Ecore');
modeltype WMM  uses wmm_base('http://www.fon.rs/NA/2011/Models/wmm_base
modeltype UML2 uses uml('http://www.eclipse.org/uml2/3.0.0/UML');

transformation AMW2WMM(in wModel : WMM, out eModel : ECORE);

main() {
  wModel.rootObjects()[MappingModel]->map toPackage();
}
-- transform MappingModel instance into new metamodel (ecore::EPackage)
mapping MappingModel::toPackage() : ecore::EPackage
{
     name := "wmm_"+self.name;
-- add required OCL library
     eAnnotations += self -> map addModelAnnotation();
-- create metamodel elements
    eClassifiers += self.ownedElement[MappingRule] -> map toClass();
}
-- tranfrom every MappingRule instance into class (ecore::EClass)
mapping MappingRule::toClass() : ecore::EClass {
    name := self.name;
    eSuperTypes+=wmm_base::MappingRule.oclAsType(EClass);
    eAnnotations += self -> map addConstraint();
}
property URI_OCL : String = "http://www.eclipse.org/emf/2002/Ecore/OCL"
mapping MappingRule::addConstraint(): ecore::EAnnotation {
    source := URI_OCL;
    details += self.right-> map toOCL_CheckStereotype("right");
    details += self.left-> map toOCL_CheckStereotype("left");
}
-- transform  MappingElement into OCL constraint
mapping MappingElement::toOCL_CheckStereotype(linkName : String) : EStr
key:="checkIs"+self.name;
value:=  "self."+linkName+".element.checkStereotype('"+self.name+"')";
}
property URI_ECORE : String = "http://www.eclipse.org/emf/2002/Ecore";
```

Fig. 10. Model transformation in QVTo

This transformation consists of a series of QVTo transformation mappings, which are presented in table 1.

Table 1. QVT transformation rules

WMM model (source)		Ecore Model (target)
MappingModel	=>	EPackage which extends wmm_base
MappingRule	=>	EClass which extends wmm_base::MappingRule
MappingElement	=>	EAnnotation which resprensts OCL constraint

According to the transformation mappings, each source LM weaving model is transformed into a new target LM weaving metamodel (represented as an EPackage meta class in EMF). Each instance of *MappingRule* within the source LM weaving model becomes a new meta class (EClass in EMF) within the target LM weaving metamodel. This new meta class is generated as a subclass of *wmm_base::MappingRule,* indicating that it represents a specific mapping rule. Every *MappingElement* of some *MappingRule* is transformed into corresponding OCL constraints represented in the target LM weaving metamodel as EAnnotation meta classes. Each OCL constraint restricts types that some concrete mapping between two concrete data schemas is allowed to reference.

Figure 11 shows the result of the defined model transformation when it is applied to the LM weaving model from our example given in figure 9.

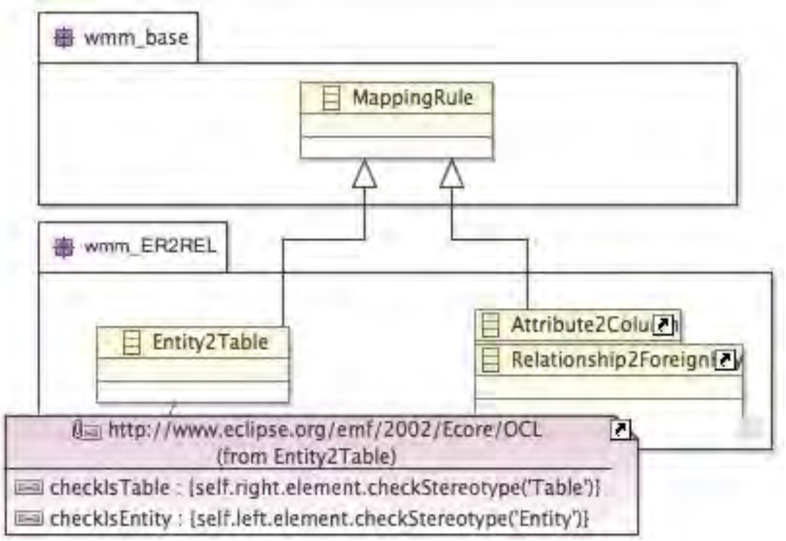
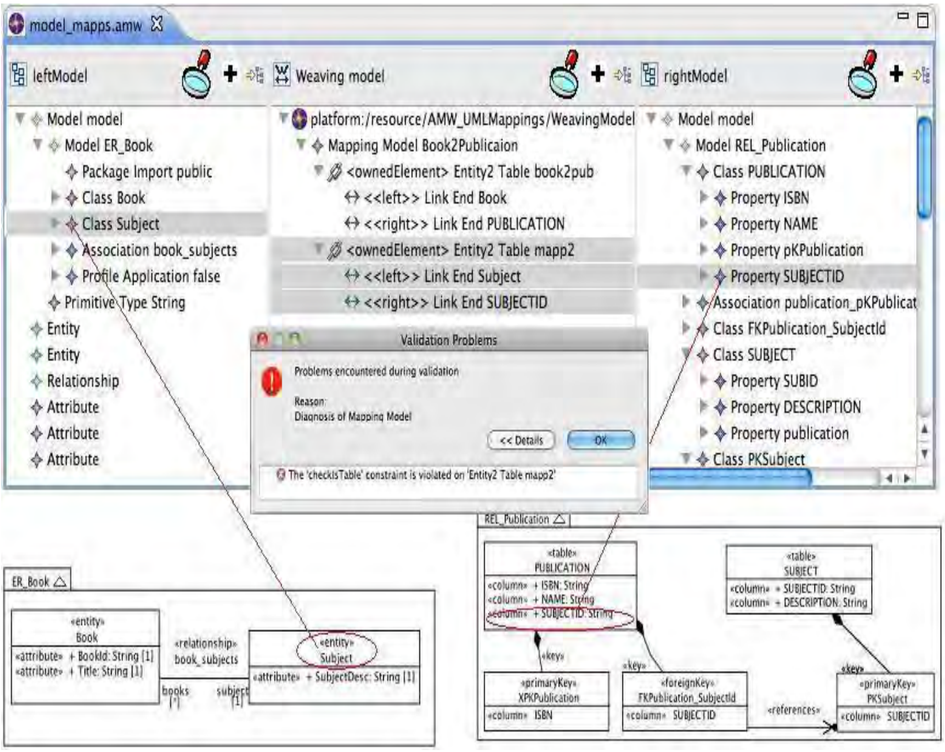

wmm_base
MappingRule
wmm_ER2REL
Entity2Table
Attribute2Colu
Relationship2Foreign
http://www.eclipse.org/emf/2002/Ecore/OCL
(from Entity2Table)
checkIsTable : {self.right.element.checkStereotype('Table')}
checkIsEntity : {self.left.element.checkStereotype('Entity')}

Fig. 11. Mappings rules between ER and relational model

The main effect of the transformation is in changing the way how mapping rules are represented. In the source model (LM model shown in fig. 9) they are represented through an UML object model, which is more suitable for their creation (i.e. specification of mapping rules). In the target LM weaving metamodel (shown in fig. 12), mapping rules are represented through an UML class diagram augmented with OCL constrains, which is more suitable for their instantiation (i.e. creation of concrete data schema mappings complying with the defined mapping rules).

5.4. An example of schema mappings

An example of concrete data schema mappings is given in figure 12. *ER_Book* and *REL_Publication* are two concrete data schemas expressed using ER and relational data model respectively.


Validation Problems

Problems encountered during validation

Reason:
Diagnosis of Mapping Model

The 'checkIsTable' constraint is violated on 'Entity2 Table mapp2'

Fig. 12. WM mapping model between two concrete models

The schemas are represented in the lower part of the figure as UML class diagrams using corresponding UML stereotypes defined for these two data models[3]. In the upper part of the figure the same schemas are depicted in the way how they are represented within AMW Weaving Editor, i.e. as trees of schema elements shown in the left and right tree-view subwindows within the main window of AMW Weaving Editor. Lines in the figure designate correspondences between the same elements in the UML class diagrams and tree nodes. The weaving model, specifying concrete mappings between

[3] Definition of these UML stereotypes is omitted in the paper for brevity.

elements of the two schemas, is also represented as a tree shown in the middle tree-view subwindow.

The shown example includes specification of several mappings between elements of the two concrete data schemas. One of the shown mappings is book2pub, which is an instance of Entity2Table mapping rule. This mapping maps entity Book from ER_Book schema into table PUBLICATION from the REL_Publication schema. Since both OCL constraints checkIsEntity and checkIsTable of Entity2Table mapping rule are satisfied, book2pub mapping is valid. Another shown mapping is *mapp2*, which maps entity *Class Subject* to column *SUBJECTID.* However, this mapping is invalid because it violates the OCL constraint *checkIsTable*, i.e. the target is not a table. Figure 12 shows the message displayed after checking the validity of the weaving model.

5.5. Discussion

In order to specify schema mappings, the fundamental requirement is to first identify correspondences between elements of schemas. Correspondences can be either generated through an automatic matching process, or can be provided manually by an expert. In our solution the process matching is not automated but supported through mapping rules which restrict semantically valid correspondences. However, our approach can be also used as a basis for the so called constraint-based automatic matching approach [12].

Our solution is primarily aimed to support manual specification of operational schema mappings. This manual specification of schema mappings is supported by the set of software tools consisting of open-source plug-ins for model weaving (AMW [8]), model transformation (M2M QVT Operational) and model validation (Eclipse OCL). All these tools are built on top of the Eclipse Modeling Framework (EMF) [19] and are available as open source from the GMT (Generative Modeling Technologies), M2M (Model-to-Model Transformation) and MDT (Model Development Tools) Eclipse modeling projects [20]. To provide support for our specific approach, existing software tool are extended by LM weaving metamodel (defined in section 5.3.), which restrict the number of match candidates for mappings between two concrete schemas.

To validate our approach, we conducted a set of experiments. Despite presented example being simple, it is easy to envisage the creation of mappings between very large source and target models. However, this solution is successfully evaluated in practice in the telecomunication domain. Our experiments tested mappings between internal and standard-based models (addressing both structural and behavioral aspects) of business processes, information entities and applications.

It is worth of noting that this approach to specify mappings could be applied on the data level as well. Models on the data level contain data which adhere to the corresponding schemas, so that we can also make correspondences between them using the weaving metamodel. We previously discussed how to use mapping rules which would restrict correspondences to meaningful ones. That discussion is also valid here.

Generally, the correspondences between data instances would not be defined explicitly, as is shown in figure 12. Instead, mapping rules would be used that are given in the schema mapping model (SM), and which define the transformation algorithms from source data to target data. SM model is used as a specification to produce transformation models in different target languages, such as ATL, XSLT, or SQL-like languages. The transformation models are extracted to the corresponding concrete syntax and additionally can generate a data weaving model (DM). Of course, in that case DM must be in the correspondence to the schema mapping SM. Also, the weaving metamodel that is generated, contains constrains which are used to prevent semantically forbidden links between the elements (data instances) of source and target data models. For example, it is not possible to link an instance *a1* of an entity *Author*, with an instance *p1* of the table *Publication* (Fig, 13). This is because there is no defined mapping between *Author* and *Publication* on the schema level. This weaving model, which is the result of a transformation, is usually called traceability model. Traceability model can be also used to check if transformation models are valid.

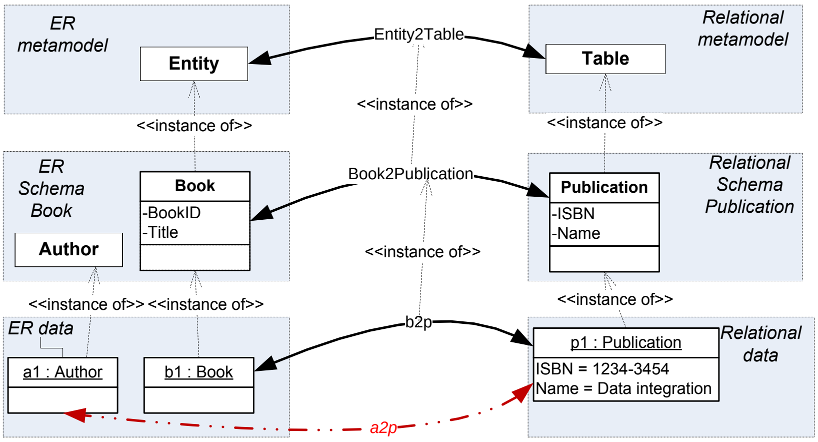


Fig. 13. Example of data mappings

6. Conclusion

This paper proposed an approach for data schema mappings which is based on weaving models. Main contributions of the paper are the following:

– A conceptual data integration framework introduced to identify kinds of models that occur in the context of data schema mappings and to precisely define their roles and mutual relationships;

- An analysis of suitability of weaving models for specification of schema mappings, which reveals that the existing approach supported by AMW toolkit does not properly support schema mappings. This is mainly due to inability of AMW approach to properly define mapping rules between schema languages;
- A proposed solution enabling the proper definition of mapping rules between schema languages, which is based on the introduction of special weaving models and metamodels with OCL constrains.

It can be concluded that the existing AMW approach based on weaving models has been carefully conceived from both theoretical and technological points of view to be general and flexible enough. This generality and flexibility enables weaving models to be applicable in a wide range of MDE related tasks. However, such generality and flexibility have shortcomings when applied in the specific domain of data schema mappings. Therefore, to overcome these shortcomings, a domain specific solution is proposed, which extends and improves the general one of AMW approach.

Acknowledgement. Research presented in this paper was supported by Ministry of Science and Technological Development of Republic of Serbia, Grant III-44010, Title: Intelligent Systems for Software Product Development and Business Support based on Models.

References

1. Lenzerini, M.: Data Integration: A Theoretical Perspective. PODS, 233-246. (2002)
2. Cui, Y., Widom, J.: Lineage Tracing for General Data Warehouse Transformations. VLDB J. 12, 41-58. (2003)
3. Ehrig, M., Sure, Y.: Ontology Mapping - An Integrated Approach. ESWS, 76-91. (2004)
4. Miller, J., Mukerji, J.: Model Driven Architecture (MDA). OMG Document. [Online]. Available: http://www.omg.org (current December 2011)
5. Del Fabro. M.D., Bézivin, J., Jouault, F., Valduriez, P.: Applying Generic Model Management to Data Mapping. In: Benzaken, V. (ed): BDA, Saint Malo, Actes. (2005)
6. Del Fabro. M.D., Valduriez, P.: Semi-automatic Model Integration using Matching Transformations and Weaving Models. SAC, 963-970. (2007)
7. Del Fabro, M.D., Valduriez, P.: Towards the Efficient Development of Model Transformations Using Model Weaving and Matching Transformations. Software and System Modeling 8(3), 305-324. (2009)
8. Del Fabro, M.D., Bézivin, J., Valduriez, P.: Weaving Models with the Eclipse AMW plugin. In: Eclipse Modeling Symposium, Eclipse Summit Europe. (2006)
9. Miller, R. J., Hernandez, M. A., Haas, L. M., Yan, L.-L., Ho, C. T. H., Fagin, R., and Popa, L.: The Clio Project: Managing Heterogeneity. SIGMOD Record 30(1), 78–83. (2001)
10. Melnik, S., Rahm, E., Bernstein P. A.: Rondo: A Programming Platform for Generic Model Management, Proc. SIGMOD 2003, 193-204. (2003)
11. Melnik, S.: Generic Model Management: Concepts and Algorithms Springer. Ph.D Dissertation, University of Leipzig. Springer LNCS 2967. (2004)

12. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB Journal 10(4), 334–350. (2001)
13. Atzeni, P., Cappellari, P., Torlone, R., Bernstein, P.A., Gianforme, G: Model-independent schema translation. The VLDB Journal 17, 1347–1370. (2008)
14. Atzeni, P., Gianforme, G., Cappellari, P.: A Universal Metamodel and Its Dictionary. A. Hameurlain et al. (Eds.): Trans. on Large-Scale Data- & Knowl.-Cent. Syst. I, LNCS 5740, 38-62. (2009)
15. Kensche, D., Quix, C., Li, X., Li, Y., Jarke, M.: Generic Schema Mappings for Composition and Query Answering. Data & Knowledge Engineering 68, 599-621. (2007)
16. Object Management Group: Meta Object Facility (MOF) 2.0 Query/View/ Transformation (QVT). [Online]. Available: http://www.omg.org/spec/QVT (current December 2011)
17. Jouault, F., Allilaire, F., Bézivin, J., Kurtev, I.: ATL: A model transformation tool. Sci. Comput. Program. (SCP) 72(1-2), 31-39. (2008)
18. Object Management Group (OMG): Object Constraint Language OMG Version 2.2. [Online]. Available: http://www.omg.org/spec/OCL/2.2/ (current December 2011)
19. Eclipse Modeling Framework Project (EMF). [Online]. Available: http://www.eclipse.org/modeling/emf/ (current December 2011)
20. Eclipse Modeling Projects. [Online]. Available: http://www.eclipse.org/modeling/ (current December 2011)

Nenad Aničić received the M.Sc. and Ph.D. degrees in Information Systems from Belgrade University, Serbia in 2001 and 2006, respectively. He is currently an associate professor in the Department of Information System and Technologies at the Faculty of Organizational Sciences, Belgrade University. He teaches courses in Business Process Modeling, Databases, XML Technologies and Applications, and Information Systems Design. His research interests include information systems development methodologies, model driven development, semantic technologies, and interoperable application systems.

Siniša Nešković received the Ph.D. and M.Sc. degrees in information systems from the University of Belgrade. He is a lecturer at the Faculty of Organizational Sciences (FOS), University of Belgrade, where he teaches courses in Data Structures and Algorithms, Business Process Modeling and Software Architectures. He currently heads the Information Systems Department and Laboratory for Information Systems of FOS. His research interests include information systems development, business process modeling and automation, model driven development, software product line engineering and advanced software architectures.

Milica Vučković works as an assistant professor in the Department of Information Systems, at the Faculty of Organizational Sciences, University of Belgrade. She received B.Sc., M.Sc. and Ph.D. degrees in information systems from the Faculty of Organizational Sciences. The areas of her research interest include: search of ontologically heterogeneous distributed resources on the Web, Model Driven Engineering, Model mappings and transformations. So far, she has authored/co-authored approximately 50 research papers. She participated in a number of research and industrial projects in the area of information systems development and software engineering.

Radovan Cvetković is an expert for Telco Information System development and implementation in company “Telekom Srbija”, Belgrade. In the same company he was the Director of IT development, and after that CIO. He received M.Sc. from Belgrade University in 1989 in domain of Information Systems and Technology. It is expected at the same university the defense of doctoral thesis "An approach to developing IS of telecommunication company which is based on the models", during 2012.

Received: August 23, 2011; Accepted: March 30, 2012.

DOI: 10.2298/CSIS111215015P

A Scale for Crawler Effectiveness on the Client-Side Hidden Web

Víctor M. Prieto[1], Manuel Álvarez[1], Rafael López-García[1], and Fidel Cacheda[1]

Comunication and Information Technologies Department - University of A Coruña
Campus de Elviña s/n - 15071 (A Coruña - Spain)
{victor.prieto, manuel.alvarez, rafael.lopez, fidel.cacheda}@udc.es

Abstract. The main goal of this study is to present a scale that classifies crawling systems according to their effectiveness in traversing the “client-side” Hidden Web.
First, we perform a thorough analysis of the different client-side technologies and the main features of the web pages in order to determine the basic steps of the aforementioned scale. Then, we define the scale by grouping basic scenarios in terms of several common features, and we propose some methods to evaluate the effectiveness of the crawlers according to the levels of the scale. Finally, we present a testing web site and we show the results of applying the aforementioned methods to the results obtained by some open-source and commercial crawlers that tried to traverse the pages.
Only a few crawlers achieve good results in treating client-side technologies. Regarding standalone crawlers, we highlight the open-source crawlers Heritrix and Nutch and the commercial crawler WebCopierPro, which is able to process very complex scenarios. With regard to the crawlers of the main search engines, only Google processes most of the scenarios we have proposed, while Yahoo! and Bing just deal with the basic ones.
There are not many studies that assess the capacity of the crawlers to deal with client-side technologies. Also, these studies consider fewer technologies, fewer crawlers and fewer combinations. Furthermore, to the best of our knowledge, our article provides the first scale for classifying crawlers from the point of view of the most important client-side technologies.

Keywords: Web Search, Crawlers, Hidden Web, Web Spam, JavaScript

1. Introduction

The World Wide Web (WWW) is the biggest information repository ever built. There are huge quantities of information that are publicly accessible, but as important as the information itself is being able to find, retrieve and gather the most relevant data according to users’ needs at any given time.

Crawling systems are the programs that traverse the Web, following URLs to obtain the documents to be indexed by web search engines. From their origins, these systems have had to face a lot of difficulties when accessing human-oriented Web sites because some technologies are very hard to analyse: web forms, pop-up menus, redirection techniques, mechanisms for maintaining user sessions, different layers which are shown or hidden depending on users' actions, etc. The pages that can only be accessed through these technologies constitute what we call Hidden Web [2] and, particularly, the set of pages that are "hidden" behind client-side technologies are called the "client-side Hidden Web". They represent a big challenge for programmers, since it is not easy to implement techniques that deal with them in a fully automatic way.

The objective of this paper is to present a scale for classifying crawlers according to their treatment of the client-side Hidden Web. This crawling scale will allows us a) to determine the capacities of each crawler dealing with client-side web technologies b) to know the percentage of the Web which has not been covered by crawlers and c) to know whether it is necessary to improve current crawling systems in order to deal with client-side web technologies in a better way.

The paper is organized as follows. In section 2 we analyse the related works. Section 3 shows the most important client-side technologies. In section 4 we list the difficulties that crawlers can find during their traversal and we classify them to create a scale. Section 5 shows this scale, which takes into account the level of use of each technology on the Web, the computational difficulty of each scenario and whether the crawlers have dealt with them. We also discuss some methods to assess the effectiveness of the crawlers at processing the aforementioned difficulties. In section 6 we show the website we have created to evaluate the crawlers and the results that each of them has obtained when traversing its pages. Finally, in sections 7 and 8 we explain the conclusions of the study and possible future works respectively.

2. Related works

There are many studies on the size of the Web and the characterisation of its content. However, there are not so many studies on classifying Web pages according to the difficulty for crawlers to process them. Weideman and Schwenke [20] published a study analysing the importance of JavaScript in the visibility of a Web site, concluding that most of the crawlers do not deal with it appropriately. However, according to the data submitted in [19] [4], currently the 90% of the Web pages use JavaScript.

From the point of view of crawling systems, there are many works aimed at creating programs that are capable of traversing the Hidden Web. There are some researches that tackle the challenges established by the server-side Hidden Web, by searching, modelling and querying web forms. We highlight HiWE [18] because it is one of the pioneer systems. Google [14] has also provided the techniques that it has used to access information through forms. Nevertheless, there are fewer studies about the client-side Hidden Web and they basically follow these approaches: they either access the content and links by means of interpreters that can execute scripts [16] [9], or they use mini-browsers such as the system proposed by Alvarez *et al.* [1].

There are also studies about the client-side technologies from the point of view of Web Spam. Their objective is to analyse the use of technologies in order to detect Web Spam techniques [10] such as Cloacking [21] [23], Redirection Spam [6] or Link Farm Spam [22].

Nevertheless, we do not know any scales that allow researchers to classify the effectiveness of the crawling systems according to their level of treatment of client-side Hidden Web technologies.

3. Client-side Web technologies

In order to define the scale, the first step is to understand the impact of the different client-side technologies on the Web ("Web Coverage"). For this purpose, we have used "The Stanford WebBase Project"[1] dataset, which is part of the "Stanford Digital Libraries Project"[2].

This dataset contains different kinds of sources and topics and more than 260 TB of data. For this analysis, we have created a 120 million page subset from the latest 2011 global datasets. In this dataset we have analysed technologies typically used to improve the user experience, such as:

- JavaScript [8], dialect of the ECMAScript standard, is an imperative and object-oriented language. It allows programmers to generate the interface dynamically.
- Applet, Java component of an application that is executed in a Web client.
- AJAX [11], technology that uses JavaScript for sending and receiving asynchronous requests and responses.
- VBScript [13], interpreted language created by Microsoft as a variant of Visual Basic.
- Flash [3], application that allows programmers to create vectorial interfaces.

Table 1 shows the results obtained by each technology (row) per month (column). According to these results, 60.30% of the documents contain JavaScript while only 2.58% contain ActionScript (Flash). The degree of occurrence of technologies like VBScript, Python and Tcl, is merely symbolic.

In short, these results confirm our assumption about the importance of knowing whether the crawlers treat the client-side technologies, since the use of JavaScript is generalized.

4. Occurrence of the main technologies

To define the basic levels of the scale we have identified the most commonly used mechanisms for generating links. To do this, we have relied on the technology analysis done in the "Client-side Web technologies" section and on a

[1] http://dbpubs.stanford.edu:8091/ testbed/doc2/WebBase/

[2] http://diglib.stanford.edu:8091/

Table 1. Analysis about Web technologies.

TECHNOLOGY	MONTH							TOTAL
	January '11	Febrary '11	March '11	April '11	May '11	June '11	July '11	
Pages	8519300	20057638	18700000	1230000	12677481	24505000	33453991	**119143410**
JavaScript	5791148	10436415	12672553	902780	6960075	16959179	18013475	**71735625**
	67.98%	**52.03%**	**67.77%**	**73.40%**	**54.90%**	**69.21%**	**53.85%**	**60.30%**
VBScript	7871	38243	19361	213	6253	20700	32529	**125170**
	0.09%	0.19%	0.10%	0.02%	0.05%	0.08%	0.10%	**0.11%**
Flash	257134	430772	584661	29376	273882	829380	669211	**3074416**
	3.02%	**2.15%**	**3.13%**	**2.39%**	**2.16%**	**3.38%**	**2.00%**	**2.58%**
Python	9	27	31	0	15	33	44	**159**
	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**0.00%**
Tcl	228	42	228	3	26	1821	5083	**7219**
	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.02%	**0.01%**
Applets	139	2832	1049	5	2142	1176	2671	**10014**
	0.00%	0.01%	0.01%	0.00%	0.02%	0.00%	0.01%	**0.01%**

manual analysis of a subset of 1000 pages selected randomly from the dataset discussed above. We have also considered the Web design common practices [7] [12].

The following features have been identified:

- Text links, which constitute the lowest level of the scale.

```
<a href="a_110001001000000000_test...html">Luis
Ramirez Lucena</a>
```

- Simple navigations, generated with JavaScript, VBScript or ActionScript. This feature includes links that are generated by means of "document.write()" or similar functions in other languages, which allow designers to add new links to the HTML code dynamically.

```
<a href="JavaScript: ...">Paolo Boi</a>
```

- Navigations generated by means of an Applet.
- Navigations generated by means of Flash.
- In the case of Applets and Flash, we divide links into two types: those which are passed as an argument to the application and those whose URL is created as a string inside the code.
- Navigations generated by means of AJAX.
- Pop-up menus, generated by a script that is associated to any event.
- Links that are defined as strings in .java, .class files or any other kinds of text and binary files.
- Navigations generated in script functions. They can be written in any scripting language and the script can be either embedded inside the HTML or located inside an external file.
- Navigations generated by means of several kinds of redirections:

- Redirections specified in the `<meta>` tag.
- Redirections generated by the “onLoad” event of the `<body>` tag.
- Redirections generated by a script when an event in another page is fired (e.g.: the “onClick” event).
- Other redirections in script blocks, which are executed the moment the page is loaded.
- Redirections executed in an applet when the page is loaded.
- Flash redirections, executed when the page and its corresponding Flash file are processed.

In addition, the navigations that are generated with any of the methods we identified can create absolute or relative URL addresses. For the addresses that are generated by means of any scripting language, it is possible to recognize the following construction methods:

- A static string inside the Script.

```
menu_static_embedded_relative(){
document.location="a_10010101100000000...html";
}
```

- A string concatenation.

```
function menu_concatenated_embedded_relative(){
var out="";
out="a_10010010100000000000_test_menu" +
"_concatenated_embedded_relative.html";
document.location=out;
}
```

- Execution of a function that builds the URL in several steps.

```
function menu_special_function_embedded_relative(){
var a1="win",a2="dow",
a3=".location.",="replace",a5;
a5="('a_1001000110000000000_test_menu_special_function";
var a6="_embedded_relative.html')"; var i,url="";
for(i=1;i<=6;i++){
url+=eval("a"+i);
}
eval(url);
}
```

Furthermore, the distinct methods we have listed above can be combined. For example, some Web sites build pop-up menus dynamically, by means of “document.write()” functions. The number of possible combinations would be too high. Hence, this study only takes into account a reduced but significant subset, which is shown in Table 2.

The 70 scenarios we have proposed are combinations of the types of URLs, technologies, locations and methods of construction of URLs we identified, all of them shown in the columns of Table 2. At the bottom of the table we show the different types of redirections. The number of scenarios could be higher, but it would not provide more information about the use of client-side technologies on the Web or about the methods that crawlers use to discover links. For example, it is not necessary to take into account the combination of menus and "document.write()" because we can deduce the capacity of the crawler from the two base cases that have been included separately.

Table 2. Combination of the types of links.

Url Type	Technology	Location	String type	Test
Relative / Absolute	Text	Embedded	Static	1 - 2
	JavaScript	Embedded / External	Static / Concatenated / Special Function	3 - 38
	Document.write()	Embedded / External	Static / Concatenated / Special Function	3 - 38
	Menu JavaScript	Embedded / External	Static / Concatenated / Special Function	3 - 38
	Link in .java	External	Static	39 - 40
	Link in .class	External	Static	41 - 42
	Applet-Link HTML	Embedded	Static	43 - 44
	Applet-Link Class	External	Static	45 - 46
	Flash-Link HTML	Embedded	Static	47 - 48
	Flash-Link SWF	External	Static	49 - 50
	AJAX	Embedded	Static	61 - 62
	JavaScript with #	Embedded	Static / Special Function	63 - 66
	VBScript	Embedded	Static / Special Function	67 - 70

Relative / Absolute	Redirect	External	Static	Tag Meta	51 - 52
				Tag body	53 - 54
				JavaScript	55 - 56
				Applet	57 - 58
				Flash	59 - 60

5. The scale for web crawlers

To create the scale from the 70 scenarios we have proposed, we have implemented a grouping process based on their difficulty:

- Theoretical difficulty of processing each scenario.
- Practical difficulty (for existing crawling systems) to treat each scenario.

Regarding the first issue, we have performed a theoretical and an empirical analysis of the scenarios in order to determine the necessary modules for a crawler to process them. The theoretical analysis consists in calculating the processing modules that a crawling system would need to process each kind of difficulty ("Extraction Process" column in Table 3). For example, to treat the plain text scenario (level 1), we only need the crawler and a basic URL extractor. However, to extract URLs embedded in JavaScript code (level 2), we would also need a JavaScript interpreter, thus increasing the complexity.

After this, we have assessed our analysis by empirically processing the scenarios by means of well-known existing tools (JavaScript interpreters, Flash decompilers, and so on). Depending on the number and complexity of the modules that are needed to process a scenario, we have assigned a discrete score to it, which is shown in "Complexity" column of Table 3. The values are based on the difficulty of developing such modules and the computational expenses of their execution. Our grouping is based on the assumption that the scenarios which require a bigger number of modules are more difficult to treat.

To learn how crawling systems process the different scenarios, we have implemented a web site (see section 6.1) that contains all the scenarios shown in Table 2. Table 6 and Table 7 (see section 6) show the results obtained by the different crawlers that traversed the example website, grouped by technology and link building method. In this case, we have grouped the scenarios that have been processed by the same number of "crawling systems", because they are considered to have the same complexity.

Based on these analyses, we have clustered the pre-defined scenarios to form the 8 levels of the scale we propose, which is shown in Table 3. For each level there is a brief description, the scenarios that are part of it and their complexity. Table 3 also shows the frequency of appearance on the Web of each web technology that has been used in each scenario (column "F"). This frequency has been obtained from the results of the Web analysis we show in section 3 (Table 1). For example, the percentage of use of JavaScript is 60.30%, so, normalising to 1, the scenarios based on JavaScript will have a frequency of 0.6. In the case of text links, we have assumed that all the web pages have at least a text link, therefore we will use a frequency of 1. This information complements our scale and allows determining more accurately what portion of the Web is treated by each crawling system, as well as the maximum complexity level reached.

In the next section, we propose various evaluation methods for measuring the effectiveness of a crawler according to the scale.

5.1. Scale evaluation method

In order to evaluate a group of crawlers, one must set them up and run them on the testing site, creating a table similar to Table 6 to compare the results. Then we propose a list of evaluation methods in order to classify the crawlers according to the level of complexity of the "links" they have processed. As a

Table 3. Scale & Link classification by difficulty.

Level	Description	Scenarios	F	Complexity	Extraction Process
1	Text link	1,2	1	Very Low	Module of Crawling - Extractor of URLs
2	JavaScript/Document.Write/Menu - Static String - Embedded	3,4,15,16,27,28	0.6	Low	Module of Crawling - JS Interpreter
	JavaScript - Concatenated String - Embedded	6	0.6		Analyser of Redirects - Extractor of URLs
3	HTML/onBody/JavaScript Redirect	51,52,53,54,55,56	1	Low	Module of Crawling - JS Interpreter
	JavaScript # - Static String - Embedded	63,64	0.6		Extractor of URLs
4	VBScript - Static String - Embedded	67,68	0.01	Medium	Module of Crawling - Interpreter VBS
	VBScript - Special Function - Embedded	70	0.01		Extractor of URLs
5	JavaScript/Document.Write - Static String - External/Embedded	9,10,18	0.6	Medium /	Module of Crawling - Interpreter VBS
	Document.Write/Menu - Static String - External	21,22,33,34	0.6	High	Extractor of URLs
	Menu - Concatenated String - Embedded	30	0.6		
6	JavaScript/Document.Write/Menu-Concatenated String-External	12,24,36	0.6	Medium /	Module of Crawling - JS Interpreter
	Applet - Static String in HTML	43,44	0.03	High	Advanced extractor of URLs
7	JavaScript - Concatenated String - External/Embedded - Relative	5,11	0.6	High	Module of Crawling
	JavaScript - Special Function - External/Embedded - Relative	7,8,13,14	0.6		Advanced JS Interpreter
	Document.Write - Concatenated String - External/Embedded	17,23	0.6		Java decompiler
	Document.Write - Special Function - External/Embedded	19,20,25,26	0.6		Analyser of external files
	Menu - Concatenated String - External/Embedded - Relative	29,35	0.6		Advanced extractor of URLs
	Menu - Special Function - External/Embedded	31,32,37,38	0.6		
	Link in .java	39,40	0.03		
	AJAX Link - Absolute	62	0.6		
8	Link in .class	41,42	0.03	Very High	Module of Crawling
	Applet - Static String in .class	45,46	0.03		Advanced JS Interpreter
	Flash - Static String in HTML/SWF	47,48,49,50	0.3		Java decompiler - Flash decompiler
	Applet/Flash Redirect	57,58,59,60	1		Advanced VBS Interpreter
	AJAX Link - Relative	61	0.6		Analyser of external files
	JavaScript with # - Special Function - Embedded	65,66	0.6		Analyser of refirects
	VBScript - Special Function - Embedded	69	0.01		Advanced extractor of URLs

preliminary step, we specify the following concepts to help us formally define each method:

- Let be $S = \{Full\ set\ of\ scenarios\}$, being $|S|$ the total number of scenarios and S_i a binary value which establishes "1" if the crawler achieved the $i-th$ scenario of S or "0" in another case.
- Let be $N = \{Full\ set\ of\ levels\ that\ the\ crawler\ processes\ successfully\}$ being $|N|$ the total number of levels and $|N_i|$ each element of N.
- Let be $C = \{Full\ set\ of\ crawlers\}$, being $|C|$ the total number of crawlers, $C_i = \{Set\ of\ crawlers\ that\ achieved\ the\ scenario\ i\}$. The values of C_i are shown in Table 4. The set of 70 scenarios are shown in a matrix where the scenario ids are built taking the value of the upper row as the units, and the value of the left column as the tens.

Table 4. Values of C_i.

Scenarios		Units									
		0	1	2	3	4	5	6	7	8	9
	0	7	7	6	6	1	6	1	1	3	3
T	1	1	2	1	1	6	6	1	3	1	1
e	2	3	3	1	2	1	1	6	6	1	3
n	3	1	1	3	3	1	2	1	1	0	0
s	4	1	1	2	2	0	0	0	0	0	0
	5	5	5	5	5	5	5	0	0	0	0
	6	0	1	5	5	0	0	5	5	0	4

- Let be f_i the frequency (see Table 3) of occurrence of i.

We now describe each of the methods we propose:

- Simple Average: it treats the scenarios defined for each level in the same way, without taking into account their difficulty. It highlights the crawlers which treat the highest number of scenarios, and consequently pay more attention to the Hidden Web in general.

$$SA = \left(\frac{1}{|S|}\right) \sum_{i=1}^{|S|} S_i$$

- Maximum Level: this model sorts crawlers according to the highest level of difficulty they can process. A crawler obtains a score i if it has the capacity of processing the scenarios of that level and the levels below. Some crawlers process a certain level, but they cannot obtain pages from scenarios of a lower level. This could be due to some problems, like the low PageRank of a Web page and others. However, this evaluation method assumes that, if a crawler is capable of dealing with a level i, it should be able to deal with lower ones.

$$ML = i \mid \forall i, j \in 1 \ldots |N|\, N_i, N_j = 1 \Lambda\, i \geq j$$

- Weighted Average: this measure depends on the number of crawlers that have previously been able to process each scenario (C_i) and whether the new crawler has been able to process it (the binary value we have called S_i). Assuming that all crawlers can deal with most of the easiest scenarios, this method gives importance to the crawlers that can obtain the highest number of difficult resources in the client-side Hidden Web, since they are the only ones with can get points in the difficult scenarios.

$$WA = \left(\frac{1}{|C| + 1}\right) * \sum_{i=1}^{|S|} ((C_i + S_i) * S_i)$$

- Eight Levels: in this model each level has a value of one point. If a crawler processes all the scenarios of one level it obtains that point. For every scenario that the crawler processes successfully, it gets $1/n$ points, where n is the total number of scenarios that have been defined for that level.
  Let $L = \{L_1 \ldots L_8\}$ be the sets of scenarios which represent the 8 levels previously defined. Then L_{ij} is the $j\!-\!th$ scenario of $i\!-\!th$ level. The number of elements of each set can be different from the others.

$$\begin{array}{ccccc} L_{11} & \ldots & L_{1p} & \rightarrow & L_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ L_{81} & \ldots & L_{8w} & \rightarrow & L_8 \end{array}$$

$$EL = \sum_{i=1}^{8} \sum_{j=1}^{|L_i|} \frac{L_{ij}}{|L_i|}$$

These models allow researchers to measure crawling systems capacity of obtaining links with a certain difficulty. However, in order to get a more conclusive opinion about which crawler processes the client-side Hidden Web the best, it is important to contextualise the technologies and their degree of occurrence in the Web. A crawler is not better than another if it processes a higher level in the scale when that level has few Web resources.

To solve this problem, we have defined the “Crawling Web Coverage” metric as the number of scenarios that a crawler solves, weighing up the technologies involved according to their frequency on the Web. We have also extended Weighted Average and Eight Levels, creating two new methods, called “Ranked Weighted Average” and “Ranked Eight Levels”, which work like the basic methods but consider the degree of occurrence of each scenario on the Web.

To do this, we compare the results we have obtained in our study of Web technologies (section 3) and those obtained in the W3techs [19] and builtwith [4] web sites. The results are similar except for the frequent use of JavaScript.

According to their data 90% of the web pages use it, but in our study this number has decreased to 60%. We have chosen our result because their data are based on an analysis of pages included in the top 1 million or top 10,000 visits. We do not consider that those pages represent the reality of the Web, since they have high traffic and use innovative technologies to offer a better user experience.

The F (Frecuency) column of Table 3 shows the occurrence frequency of the base technology of each level. In this way, a crawler that processes a level with frequency 1 will be more valuable than one that processes a level with a lower frequency, since the probability of finding a link of the first type on the Web is much higher. On the other hand, a crawler that obtains links with a value of 0.01 will not get a high result, since this kind of links are scarce on the Web.

- Ranked Weighted Average

$$RWA = \left(\frac{1}{|C|+1}\right) * \sum_{i=1}^{|S|} ((C_i + S_i) * S_i \, * f_i)$$

- Ranked Eight Levels

$$REL = \sum_{i=1}^{8} \sum_{j=1}^{|L_i|} f_{i,j} * \frac{L_{i,j}}{|L_i|}$$

The Simple Average and Weighted Average methods and their ranked counterparts should be normalized in order to obtain values between 0 and 8, which make the methods easier to use and to compare.

As we can see, each method provides a different point of view on the capabilities of a crawler. Therefore, it is precisely this set of all methods that gives us a global vision of the processing capacity of the client-side technologies and the percentage of the web that is being treated.

6. Experimental results

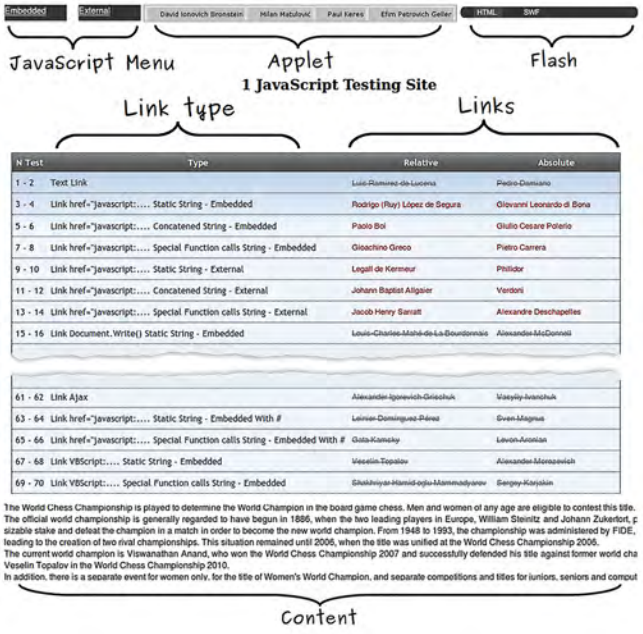
In this section we describe the results of the experiments for the creation of the scale and its application in the classification of open-source and commercial crawlers. As a preliminary step, we describe the website we have designed to perform the tests.

6.1. Testing web site

In order to check how crawling systems deal with the different scenarios, we have created a Web site[3] for performing experiments. This Web site contains 70 links representing the 70 scenarios shown in Table 2. Those links use and combine the different technologies explained previously. With the purpose of

[3] http://www.tic.udc.es/∼mad/resources/projects/jstestingsite/

stimulating the indexing of the Web site, we have linked it from the main Web page of our institution Web site, we have written its content in English and we have added the biography of a famous chess player to each scenario. The latter measure was taken in order to avoid confusion with Web Spam sites.



1 JavaScript Testing Site

N Test	Type	Relative	Absolute
1 - 2	Text Link	~~Luis Ramirez de Lucena~~	~~Pedro Damiano~~
3 - 4	Link href="javascript:.... Static String - Embedded	Rodrigo (Ruy) López de Segura	Giovanni Leonardo di Bona
5 - 6	Link href="javascript:.... Concatened String - Embedded	Paolo Boi	Giulio Cesare Polerio
7 - 8	Link href="javascript:.... Special Function calls String - Embedded	Gioachino Greco	Pietro Carrera
9 - 10	Link href="javascript:.... Static String - External	Legall de Kermeur	Philidor
11 - 12	Link href="javascript:.... Concatened String - External	Johann Baptist Allgaier	Verdoni
13 - 14	Link href="javascript:.... Special Function calls String - External	Jacob Henry Sarratt	Alexandre Deschapelles
15 - 16	Link Document.Write() Static String - Embedded	~~Louis-Charles Mahé de La Bourdonnais~~	~~Alexander McDonnell~~
61 - 62	Link Ajax	~~Alexander Igorevich Grischuk~~	~~Vasyliy Ivanchuk~~
63 - 64	Link href="javascript:.... Static String - Embedded With #	~~Leinier Domínguez Pérez~~	~~Sven Magnus~~
65 - 66	Link href="javascript:.... Special Function calls String - Embedded With #	~~Gata Kamsky~~	~~Levon Aronian~~
67 - 68	Link VBScript:.... Static String - Embedded	~~Veselin Topalov~~	~~Alexander Morozevich~~
69 - 70	Link VBScript:.... Special Function calls String - Embedded	~~Shakhriyar Hamid oglu Mammadyarov~~	~~Sergey Karjakin~~

The World Chess Championship is played to determine the World Champion in the board game chess. Men and women of any age are eligible to contest this title.
The official world championship is generally regarded to have begun in 1886, when the two leading players in Europe, William Steinitz and Johann Zukertort, p
sizable stake and defeat the champion in a match in order to become the new world champion. From 1948 to 1993, the championship was administered by FIDE,
leading to the creation of two rival championships. This situation remained until 2006, when the title was unified at the World Chess Championship 2006.
The current world champion is Viswanathan Anand, who won the World Chess Championship 2007 and successfully defended his title against former world cha
Veselin Topalov in the World Chess Championship 2010.
In addition, there is a separate event for women only, for the title of Women's World Champion, and separate competitions and titles for juniors, seniors and comput

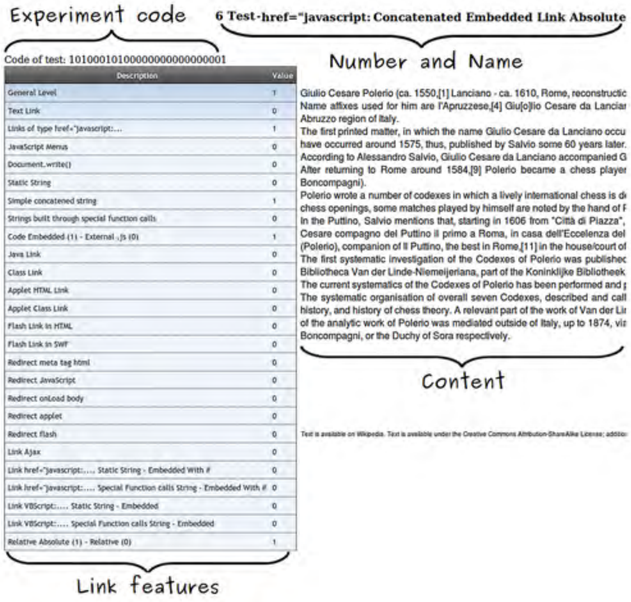
Fig. 1. Main page of the prototype.

Fig. 1 shows the main page of the prototype. It contains the following parts:

- At the top, from left to right, there are links in a JavaScript menu, links generated in an Applet and links in Flash.
- In the middle of the page, there is a table with four columns containing the test number, its type and the relative and absolute links respectively.
- At the bottom, below the table, it is the content.

Apart from that, we have created a page for the result of each test, so if a crawler can access certain contents, it will mean that it has been able to process the corresponding links. Fig. 2 shows a result page, consisting of the following elements:

- At the top-centre, it shows the test number and name.
- At the top-left it shows the reference code of the test, a binary mask that represents the features of the test.
- On the left side, it shows a table that enumerates the features of the test.
- The centre includes the biography of a chess grandmaster.



6 Test-href="javascript: Concatenated Embedded Link Absolute

Code of test: 1010001010000000000000001

Description	Value
General Level	1
Text Link	0
Links of type href="javascript:...	1
JavaScript Menus	0
Document.write()	0
Static String	0
Simple concatened string	1
Strings built through special function calls	0
Code Embedded (1) - External .js (0)	1
Java Link	0
Class Link	0
Applet HTML Link	0
Applet Class Link	0
Flash Link in HTML	0
Flash Link in SWF	0
Redirect meta tag html	0
Redirect JavaScript	0
Redirect onLoad body	0
Redirect applet	0
Redirect flash	0
Link Ajax	0
Link href="javascript:.... Static String - Embedded With #	0
Link href="javascript:.... Special Function calls String - Embedded With #	0
Link VBScript:.... Static String - Embedded	0
Link VBScript:.... Special Function calls String - Embedded	0
Relative Absolute (1) - Relative (0)	1

Giulio Cesare Polerio (ca. 1550,[1] Lanciano - ca. 1610, Rome, reconstructic
Name affixes used for him are l'Apruzzese,[4] Giu[o]lio Cesare da Lanciar
Abruzzo region of Italy.
The first printed matter, in which the name Giulio Cesare da Lanciano occu
have occurred around 1575, thus, published by Salvio some 60 years later.
According to Alessandro Salvio, Giulio Cesare da Lanciano accompanied G
After returning to Rome around 1584,[9] Polerio became a chess player
Boncompagni).
Polerio wrote a number of codexes in which a lively international chess is d
chess openings, some matches played by himself are noted by the hand of F
In the Puttino, Salvio mentions that, starting in 1606 from "Città di Piazza",
Cesare compagno del Puttino il primo a Roma, in casa dell'Eccelenza del
(Polerio), companion of Il Puttino, the best in Rome,[11] in the house/court of
The first systematic investigation of the Codexes of Polerio was publishec
Bibliotheca Van der Linde-Niemeijeriana, part of the Koninklijke Bibliotheek
The current systematics of the Codexes of Polerio has been performed and
The systematic organisation of overall seven Codexes, described and call
history, and history of chess theory. A relevant part of the work of Van der Li
of the analytic work of Polerio was mediated outside of Italy, up to 1874, vi
Boncompagni, or the Duchy of Sora respectively.

Text is available on Wikipedia. Text is available under the Creative Commons Attribution-ShareAlike License; additio

Fig. 2. Target Web page associated to the link of one scenario.

6.2. Experimental setup

For both the realization of the scale and the classification of crawling systems according to the evaluation methods, we have obtained two types of results.

The first one is based on the fact that a crawling system indexes the resources it obtains, and puts the data in a certain repository. Hence, we can extract information by querying this repository. On the other hand, we can extract access information from the log files of the Web server that maintains the testing Web site, in order to identify the crawlers that were accessing any page of that server. With this study we want to get further detail on how crawlers try to process each level. We have used the User-Agent and IP pairs published in the official robots page [4] to identify and remove non-independent crawlers, that is to say, those which depend on the crawler of a third party. In order to accelerate the process of visiting and indexing the site, we have registered the site in these crawlers: GoogleBot[5], Bing[6], Yahoo![7], PicSearch[8] and Gigablast[9].

For both the data in the logs and those obtained in the repositories, we have analysed the results of the crawlers of the major search engines as well as the

[4] http://www.robotstxt.org/
[5] http://www.google.es/addurl/
[6] http://www.bing.com/webmaster/SubmitSitePage.aspx
[7] http://siteexplorer.search.yahoo.com/submit
[8] http://www.picsearch.com/menu.cgi?item=FAQ
[9] http://www.gigablast.com/addurl

results of open-source and commercial ones. We have compared the global results in terms of the type of link and in terms of the method of construction of the URL string. First, we have studied each type of crawler independently, and then, comparing the results of both types.

These results have been used to evaluate each crawling system by means of the proposed methods and to obtain a final classification of each one in the scale we have proposed.

6.3. Open-source and commercial crawlers

We have analysed 23 open-source and commercial crawlers. Table 5 shows their most important characteristics from the point of view of the client-side technologies, along with the treatment of forms.

Among the features that they share, but we have not shown, we highlight: options about the use of a proxy, limitations in the number of documents, different protocols, inclusion/exclusion of file extensions, cookies, User-Agent, options about domains/directories, logging and some more.

Table 5. Open-source and Commercial Crawlers.

Crawler	Licence	JavaScript	Flash	Forms	Authentication	Thread
Advanced Site Crawler	Free	Yes	No	No	Yes	Yes
Essential Scanner	Free	No	No	No	Yes	No
Gsite Crawler	Free	No	No	No	No	Yes
Heritrix	Free	Yes	No	No	Yes	Yes
Htdig	Free	No	No	No	Yes	No
ItSucks	Free	No	No	No	Yes	Yes
Jcrawler	Free	No	No	No	No	No
Jspider	Free	No	No	No	Yes	Yes
Larbin	Free	No	No	No	Yes	Yes
MnogoSearch	Free	No	No	No	Yes	No
Nutch	Free	No	Yes	No	No	Yes
Open Web Spider	Free	No	No	No	No	Yes
Oss	Free	No	No	No	No	Yes
Pavuk	Free	Yes	No	Yes	Yes	Yes
Php Crawler	Free	No	No	No	No	No
WebHTTrack	Free	Yes	Yes	No	Yes	Yes
JOC Web Spider	Shareware	No	No	No	Yes	Yes
MnogoSearch	Shareware	No	No	No	Yes	Yes
Teleport Pro	Shareware	Yes	No	Yes	Yes	Yes
Visual Web Spider	Shareware	No	No	No	Yes	Yes
Web Data Extractor	Shareware	No	No	No	Yes	Yes
Web2Disk	Shareware	Yes	No	Yes	Yes	Yes
Web Copier Pro	Shareware	Yes	Yes	Yes	Yes	Yes

After examining these and other features, and taking into account the evaluation of their results in the treatment of client side technologies, we have selected the seven best crawlers among those shown in Table 5.

Among the open-source crawlers, we have chosen Nutch[10] [5], Heritrix[11] [15], Pavuk[12] and WebHTTrack[13], and among the commercial crawlers we have chosen Teleport[14], Web2disk[15] and WebCopierPro[16].

6.4. Results for open-source and commercial crawlers

First, we have studied the results according to the content of the Web site that has been indexed by the different open-source and commercial crawlers. To that end, we have analysed the repositories that the crawler generated during its execution. The crawler that achieves the best results is WebCopierPro (left side in Table 6), which processed 57.14% of the levels, followed by Heritrix with 47.14% and Web2Disk with 34.29%. Only a few of them get values beyond 25% in most types of links, and even those are unable to get links in many cases.

It is also important to notice the poor results that they have obtained for redirections, especially in the case of WebCopierPro, which has not been able to deal with any of them, although it has got results of 100% in harder levels. None of the crawlers has reached the 100% in redirections, because they have not been able to process pages with redirections embedded in Applets or Flash. Although they have downloaded the pages, they have not executed the Flash or the Applet that generates the redirection.

If we analyse the results by type of link, which are shown in the left columns of Fig. 3 (section 6), we will obtain a vision about which types of links are processed by a bigger number of crawlers. Apart from the 100% obtained for text links, we can verify that they can complete 35% to 40% of the scenarios of href="javascript..."; "document.write()"; menu links; links with "#" and VBScript.

They only achieve 7% of the links in .class or .java files and those which have been generated with AJAX. None of them has got links from Flash, but this can be due to the scarce attention that crawlers pay to these kinds of links.

The last rows of Table 6 classify the links by their construction method and show how many of them have been processed by each crawler. If we calculate the mean of the percentages, we have 42.52% for static links, 27.38% for links generated by string concatenation and 15.18% for links that generated by means of functions. This indicates that most of the crawlers search URLs by means of regular expressions.

[10] http://nutch.apache.org/

[11] http://www.archive.org

[12] http://www.pavuk.org

[13] http://www.httrack.com

[14] http://www.tenmax.com/teleport/pro/home.htm

[15] http://www.inspyder.com/ products/Web2Disk/Default.aspx

[16] http://www. maximumsoft.com/products/wc_pro/overview.html

Table 6. Summary of results for open-source and Commercial crawlers. On the left side in the repositories and on the right side in the logs.

REPOSITORIES (Passed — %)								LOGS (Passed — %)	
	Heritrix	Nutch	Pavuk	Teleport	Web2Disk	WebCopierPro	WebHTTrack	Teleport	WebHTTrack
Text link	2 100.00%	2 100.00%	2 100.00%	2 100.00%	2 100.00%	2 100.00%	2 100.00%	2 100.00%	2 100.00%
Href="JavaScript link	6 50.00%	3 25.00%	0 0.00%	3 25.00%	5 41.67%	12 100.00%	3 25.00%	4 33.33%	3 25.00%
Document.write link	6 50.00%	3 25.00%	0 0.00%	2 16.67%	4 33.33%	12 100.00%	2 16.67%	3 25.00%	2 16.67%
Menu link	6 50.00%	3 25.00%	0 0.00%	2 16.67%	4 33.33%	12 100.00%	2 16.67%	3 25.00%	2 16.67%
Flash link	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	2 50.00%
Applet link	2 50.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	2 50.00%	0 0.00%	2 50.00%
Redirects	6 60.00%	6 60.00%	2 20.00%	6 60.00%	4 33.33%	0 0.00%	6 60.00%	6 60.00%	6 60.00%
Class or Java link	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	2 50.00%	0 0.00%	0 0.00%	0 0.00%
AJAX link	0 0.00%	1 50.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%
Links with #	2 50.00%	2 50.00%	0 0.00%	2 50.00%	2 50.00%	0 0.00%	2 50.00%	3 75.00%	2 50.00%
VBScript link	3 75.00%	3 75.00%	0 0.00%	3 75.00%	3 75.00%	0 0.00%	2 50.00%	3 75.00%	2 50.00%
Static string link	26 61.90%	19 45.24%	4 9.52%	18 42.86%	22 52.38%	16 38.10%	20 47.62%	18 42.86%	22 52.38%
Concatenated string link	6 50.00%	2 16.67%	0 0.00%	1 8.33%	1 8.33%	12 100.00%	1 8.33%	1 8.33%	1 8.33%
Special function string link	1 6.25%	2 12.50%	0 0.00%	1 6.25%	1 6.25%	12 75.00%	0 0.00%	5 31.24%	0 0.00%
Tests passed	33 47.14%	23 32.86%	4 5.71%	20 28.57%	40 57.14%	40 57.14%	21 30.00%	24 34.29%	23 32.86%

From these results, we conclude that the probability of finding links by means of regular expressions or treatment of text is inversely proportional to the difficulty of generation of the link.

Summarizing the data of Table 6, we conclude that only one third of the links that are generated with client-side technologies are treated by open-source or commercial crawlers.

As a complement to these results, we have analysed the entries that each crawler has generated in the logs of our Web server. The right side of Table 6 shows the results of the crawlers in the logs for the cases where these are different from the results of analysing the repository (left side). Those differences are due to:

- WebHTTrack has tried to access Flash scenarios, but it has not taken into account that the structure to pass parameters to Flash is “parameter1=data1& parameter2=data2&...” so it considers the whole string as an URL although it is not.

  ```
  "GET mad/resources/projects/jstestingsite/1_test/
  link_relative=a_10000100100001000000000000
  _test_flash_link_html_relative.html&link_absolute=
  http://www.tic.udc.es/~mad/resources/projects/
  jstestingsite/1_test/a_10000100100001000000000001
  _test_flash_link_html_absolute.html"
  ```

- Teleport has tried to access absolute URLs that have been generated by a function. It detected “http...” patterns and searched the corresponding “.html” patterns that indicate the end of the address. Nevertheless, it has not detected that some variables have been defined in the middle, so the URL it has generated is incorrect.

  ```
  "GET /~mad/resources/projects/jstestingsite/1_test/";
  var%20a6="a_10001001100000000000000001
  _test_document_write_function_special_embedded_absolute.html"
  ```

6.5. Results for crawlers of the main web search engines

In order to perform the analysis of the crawlers that the main Web search engines use, we registered the Web site on the 23rd of May of 2011 in the different search engines. Once it was available, we waited for 70 more days. Then, we checked the logs of the Web site in order to know which crawlers had accessed at least the main page of the Web site. Table 7 contains the results we obtained. It does not show the results for Alexa, Ask, Bing, Gigablast and PicSearch, since they have not indexed the testing Web site. Only Google and Yahoo! indexed it. Among the factors that could induce some of the crawlers to skip the Web site, we could remark the following:

- Low PageRank [17].

Table 7. Summary of the results for crawlers of the main Web search engines.

	Google (Passed — %)		Yahoo! (Passed — %)	
Text link	2	10000%	1	50.00%
Href="javaScript link	**6**	**50.00%**	0	0.00%
Document.write link	**6**	**50.00%**	0	0.00%
Menu link	**6**	**50.00%**	0	0.00%
Flash link	0	0.00%	0	0.00%
Applet link	0	0.00%	0	0.00%
Redirects	**6**	**50.00%**	2	20.00%
Class or java link	0	0.00%	0	0.00%
AJAX link	0	0.00%	0	0.00%
Links with #	**4**	**100.00%**	0	0.00%
VBScript link	**1**	**25.00%**	0	0.00%
Static string link	**17**	**40.48%**	3	7.14%
Concatenated string link	**6**	**50.00%**	0	0.00%
Special function string link	**8**	**50.00%**	0	0.00%
Total passed	**31**	**44.29%**	3	4.29

- It is stored in a deep level inside the Web site of the department.
- High content of code. This can make crawlers' heuristics decide that the Web site might contain Malware or links too difficult to analyse.

Google processed 44.29% of the links. This implies that neither the design nor the content nor the other variables inhibited crawlers to try to traverse the Web site. Yahoo! also processed some links, but only the easiest ones.

GoogleBot processed 50% of many of the levels we proposed. The other half was not processed because the code was stored in an external file that the crawler did not analyse. If these files were analysed, we suspect that GoogleBot would achieve much better results. The links that GoogleBot did not process included technologies like Flash, Applets and AJAX or file types like .class and .java. These crawlers have shown the same results in the indices (querying their search engine) and in the logs of our Web server. Hence, we do not show the results separately.

6.6. Result comparison

Looking at the overall results (last row) of Tables 6 and Table 7, we see that generally, open-source and commercial crawlers achieve better results. Only Google, getting processed 34 of the 70 scenarios, has similar performance.

This can be due to the fact that an open-source or a commercial crawler can be configured by the user, who establishes what kind of links should be followed without taking into account the performance or the security. These features cannot be ignored by the crawlers of the main Web search engines.

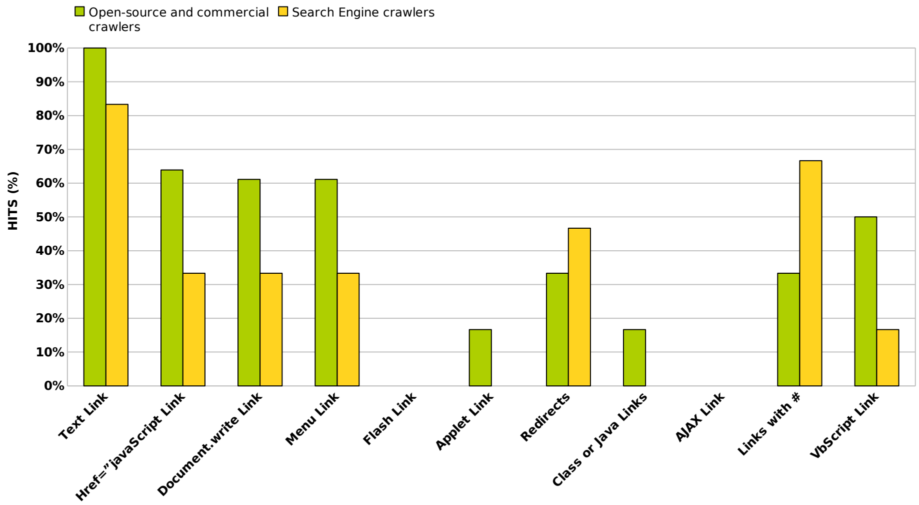


Fig. 3. Comparison among crawlers by link type.

Fig. 3 compares the results by technology used in the links. Once again, open-source crawlers have been more effective. One thing we would like to highlight is that the curve that every group of crawlers draw is similar. Hence, we can conclude that despite achieving a lower number of links in each technology, crawlers show the same interest in processing the same kind of technologies.

6.7. Classification of the crawlers according to the scale

After obtaining the results of each crawler in the website, we evaluate the crawling systems in the scale, using the evaluation methods we have proposed. Fig. 4 shows these results.

- For Simple Average, WebCopier gets the best results, followed by Heritrix and Google.
- For Maximum Level, Google places first since it processes level 8 links. It is followed by WebCopier, which obtains 7 points, and Heritrix, which obtains 6 points. As Google achieves the maximum level in this model but not in others, we can conclude that it has the capacity to deal with every scenario, but it does not try some of them because of internal policies.
- For the Weighted Average measure, WebCopier is followed by Google, Heritrix and Nutch.
- For Eight Levels, top places are for Heritrix, Web2Disk and WebCopier. GoogleBot places fourth. This means that the three top crawlers have dealt with a big quantity of levels in each group or they have gone through links that were part of a group with few links, which makes each link more valuable.

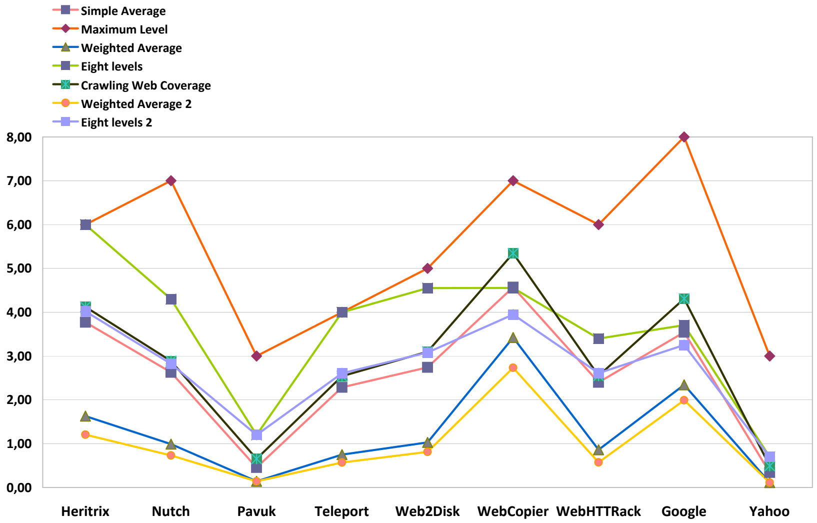


Fig. 4. Results according to the proposed scales.

However, the frequency of the technologies in the Web should be taken into account too. That is why we also show the Crawling Web Coverage metric and the results of the Ranked Weighted Average and Ranked Eight Levels methods, which are based on such metric. WebCopier, Google and Heritrix get the best results, so they are the ones that deal with the biggest portion of the Web.

- For Ranked Weighted Average, crawlers obtain lower results, but quite similar to the previous ones. Once again, WebCopier and Google are at the top at a certain distance from the rest.
- For Ranked Eight Levels, almost all crawlers decrease their performance between 1 and 2 points.

When a crawler gets a high score in the previous assessments and a low one in these, it means that it has been able to traverse difficult links, but that the technologies these links are made of do not have a significant number of occurrences on the Web.

Weighted results evaluate if a crawler can analyse client-side technologies on Web sites, but they also analyse if those capabilities are appropriate for treating the current state of the Web. We conclude that the best crawlers in both quantity and quality are Google and WebCopier, followed by Heritrix, Nutch and Web2Disk. It is important to highlight the results of GoogleBot. It takes into account a wide range of technologies although it is oriented to traverse the entire Web and it makes a lot of performance and security considerations.

7. Conclusions

This article proposes a scale that allows us to classify crawling systems according to their effectiveness when they access the client-side Hidden Web. It takes into account a global study on 120 million pages to determine the most commonly used client-side technologies and another one to explain the specific methods to generate links. It also includes the crawlers of the main search engines, as well as several open-source and commercial ones. Hence, we believe it is more exhaustive than the work by M. Weideman and F. Schwenke [20].

In order to classify the distinct crawling systems, we have created a Web site implementing all the difficulties that we had included in the scale.

The results we have obtained show that, for both the levels that the crawlers have tried and the levels they have achieved, most of the times they try to discover new URLs processing the code as text and using regular expressions. It is true that this allows them to discover a big amount of scenarios and that the computational expenses are not high, but we conclude that most of the URLs which are located inside client-side technologies are not discovered. The only crawlers that achieve good results at processing the client-side Hidden Web are WebCopier and GoogleBot. They surely are using an interpreter that allows them to execute code.

8. Future works

One of the future works is the design, implementation and testing of an algorithm that is capable of dealing with all the levels proposed in the scale. The algorithm should be composed of different modules for each level, starting with an analysis based on text and regular expressions, and finishing with the use of interpreters and mini-browsers. As a consquence, each module would have more computational cost, but they should be executed only if there are evidences pointing to the possible finding of new URLs in web pages that are not Web Spam. The tests should measure both the efficacy and the efficiency of the algorithm, since it would be designed to be included in a global crawling system.

A long term work would be modelling and implementing a high-performance crawler that includes the aforementioned algorithm and the Web Spam detection module. With both features we would have a high-performing safe crawler which would grant access to the hidden web content that, as we proved in this article, is not completely processed by crawlers.

References

1. Álvarez, M., Pan, A., Raposo, J., Hidalgo, J.: Crawling Web Pages with Support for Client-Side Dynamism. In: Yu, J., Kitsuregawa, M., Leong, H. (eds.) Advances in Web-Age Information Management, Lecture Notes in Computer Science, vol. 4016, chap. 22, pp. 252–262. Springer Berlin / Heidelberg, Berlin, Heidelberg (2006)

2. Bergman, M.K.: The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing 7(1) (Aug 2001)
3. Braunstein, R., Wright, M.H., Noble, J.J.: ActionScript 3.0 Bible. Wiley (Oct 2007)
4. BuiltWith: Web Technology Usage Statistics. http://trends.builtwith.com/ (2011)
5. Cafarella, M., Cutting, D.: Building Nutch: Open Source Search: A case study in writing an open source search engine. ACM Queue 2(2) (2004)
6. Chellapilla, K., Maykov, A.: A taxonomy of javascript redirection spam. In: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web. pp. 81–88. AIRWeb '07, ACM, New York, NY, USA (2007)
7. Danielson, D.R.: Web navigation and the behavioral effects of constantly visible site maps. Interacting with Computers 14(5), 601–618 (2002)
8. Flanagan, D.: JavaScript: The Definitive Guide. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 3rd edn. (1998)
9. Google: V8 JavaScript Engine. http://code.google.com/p/v8/ (2011)
10. Gyongyi, Z., Molina, H.G.: Web Spam Taxonomy. In: First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005) (2005)
11. Holdener, III, A.T.: Ajax: the definitive guide. O'Reilly, first edn. (2008)
12. Kalbach, J., Bosenick, T.: Web page layout: A comparison between left- and right-justified site navigation menus. J. Digit. Inf. 4(1) (2003)
13. Kingsley-Hughes, A., Kingsley-Hughes, K., Read, D.: VBScript Programmer's Reference. Wrox Press Ltd., Birmingham, UK, UK, 3rd edn. (2007)
14. Madhavan, J., Ko, D., Kot, L., Ganapathy, V., Rasmussen, A., Halevy, A.: Google's deep web crawl. Proc. VLDB Endow. 1, 1241–1252 (August 2008)
15. Mohr, G., Kimpton, M., Stack, M., Ranitovic, I.: Introduction to heritrix, an archival quality web crawler. In: 4th International Web Archiving Workshop (IWAW04) (2004)
16. Mozilla: Mozilla rhino javascript engine (2011), http://www.mozilla.org/rhino/
17. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford Digital Library Technologies Project (1998)
18. Raghavan, S., Garcia-Molina, H.: Crawling the hidden web. In: Proceedings of the 27th International Conference on Very Large Data Bases. pp. 129–138. VLDB '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
19. W3Techs: World Wide Web Technology Surveys. http://w3techs.com/ (2011)
20. Weideman, M., Schwenke, F.: The influence that JavaScript has on the visibility of a Website to search engines. Information Research 11(4) (Jul 2006)
21. Wu, B., Davison, B.D.: Cloaking and redirection: A preliminary study. In: AIRWeb '05: Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (May 2005)
22. Wu, B., Davison, B.D.: Identifying link farm spam pages. In: Special interest tracks and posters of the 14th international conference on World Wide Web. pp. 820–829. WWW '05, ACM, New York, NY, USA (2005)
23. Wu, B., Davison, B.D.: Detecting semantic cloaking on the web. In: Proceedings of the 15th International World Wide Web Conference. pp. 819–828. ACM Press (2006)

Víctor M. Prieto graduated in Computing Science at the University of A Coruña in September of 2007 and he is a Ph.D. candidate in Computer Science at the same university. At this moment, he is part of the Telematics Group of the Department of Information and Comunications Technologies at the same university. His main research fields are web crawling, hidden web and web spam.

Manuel Álvarez is an Associate Professor in the Department of Information and Communication Technologies, at the University of A Coruña (Spain). He received his Bachelor's Degree in Computer Engineering from the University of A Coruña in 1999 and a Ph.D. Degree in Computer Science from the same University in 2007. His research interests are related to data extraction and integration, semantic and Hidden Web. Manuel has managed several projects at national and regional level in the field of data integration and Hidden Web accessing. He has also published several papers in international journals and has participated in multiple international conferences. He also teaches a Master's degree at the University of A Coruña and is consultant for Denodo Technologies (a telematic engineering company).

Rafael López-García got his Degree in Computing Science at the University of A Coruña in September of 2006. In the years that followed, he has focused on teaching and research, studying the postdegree at the same university. His primordial fields of interest in Computing Science are web technologies and distributed systems in Information Retrieval (IR), field in which he wants to write his Ph.D. Thesis.

Fidel Cacheda is an Associate Professor in the Department of Information and Communications Technologies at the University of A Coruña (Spain). He received his Ph.D. and B.S. degrees in Computer Science from the University of A Coruña, in 2002 and 1996, respectively. He has been part of the Department of Information and Communication Technologies, the University of A Coruña, Spain, since 1998. From 1997 to 1998, he was an assistant in the Department of Computer Science of Alfonso X El Sabio University, Madrid, Spain. He has been involved in several research projects related to Web information retrieval and multimedia real time systems. His research interests include Web information retrieval and distributed architectures in information retrieval. He has published several papers in international journals and has participated in multiple international conferences.

Received: December 15, 2011; Accepted: April 9, 2012.

DOI: 10.2298/CSIS110606004M

Wavelet trees: a survey

Christos Makris[1]

[1] Department of Computer Engineering and Informatics,
University of Patras, 26500 Patras, Greece
makri@ceid.upatras.gr

Abstract. The topic of this paper is the exploration of the various characteristics of the wavelet tree data structure, a data structure that was initially proposed for text compression applications but has found a plethora of other uses in text indexing and retrieval. Issues concerning the efficient maintenance of the structure, plus its handling in various applications are explored. Our main aim is to provide to computer science researchers that would like to explore the specific area, an up-to-date comprehensive material covering a wide range of applications. This kind of up-to-date survey is missing from the current bibliography and we hope that it will help young researchers to get familiar with the notions of this research area.

Keywords: information retrieval, text algorithms, data structures.

1. Introduction

The constant increase in the volume of transmitted and stored data makes imperative the design of efficient algorithms and data structures for handling them. The field of data and text compression has been traditionally involved in providing tools that could face effectively the problems emerging in large scale information retrieval applications. Moreover in the time course of the previous decade, a challenging and interesting issue has flourished: self-indexing data structures, that is data structures that do not need the storage of the source text to operate, but embed in them both the text, and the indexer that operates on it. Self indexed data structures move the compression target from the source text to the indexing module and permit further functionalities.

There are a lot of algorithms and compression results that fall under this realm of research activity [39], [40], [43], [44], [45], [94] and from this emergence of results, a data structure called wavelet tree and having its roots in the arena of range searching data structures has emerged. This data structure that begun as a main component of compressed suffix arrays [64] and as a complement to other indices (such as the FM-index [94]), has been fruitfully extended in order to be used in other applications such as image compression, posting lists' handling, spatial searching, XML queries and many more.

The basic theme in the present paper is to examine the various issues involved and the basic characteristics of the specific data structure, dealing with a set of selected applications. There exists various works [38], [51], [68] exploring different aspects of the wavelet tree data structure however there does not exist in the literature an up to date survey covering its plethora of applications and referring to the till now space and time complexities bounds for them, in order to be provided as a roadmap to young researchers entering the area. We aim to cover this gap in the scientific literature by condensing all relevant information in one single reference, appropriately putting all the pieces together, thus helping young researchers and enter them smoothly in the notions of the specific scientific area.

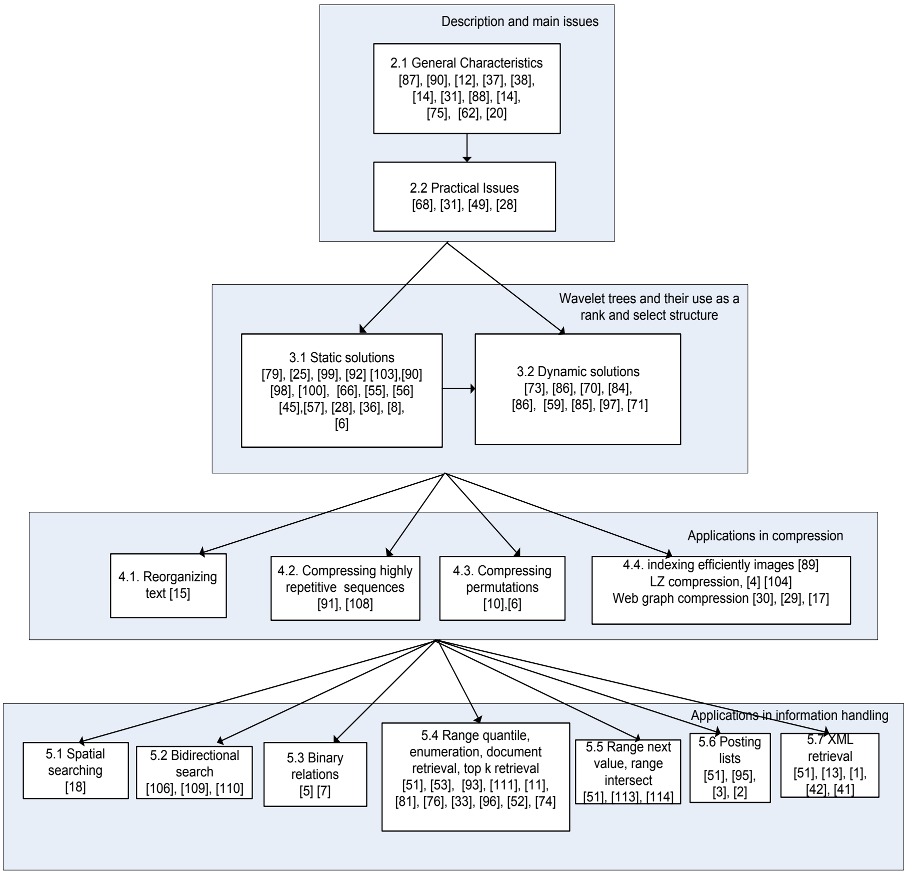

Description and main issues
2.1 General Characteristics [87], [90], [12], [37], [38], [14], [31], [88], [14], [75], [62], [20]
2.2 Practical Issues [68], [31], [49], [28]
Wavelet trees and their use as a rank and select structure
3.1 Static solutions [79], [25], [99], [92] [103],[90] [98], [100], [66], [55], [56] [45],[57], [28], [36], [8], [6]
3.2 Dynamic solutions [73], [86], [70], [84], [86], [59], [85], [97], [71]
Applications in compression
4.1. Reorganizing text [15]
4.2. Compressing highly repetitive sequences [91], [108]
4.3. Compressing permutations [10],[6]
4.4. indexing efficiently images [89] LZ compression, [4] [104] Web graph compression [30], [29], [17]
Applications in information handling
5.1 Spatial searching [18]
5.2 Bidirectional search [106], [109], [110]
5.3 Binary relations [5] [7]
5.4 Range quantile, enumeration, document retrieval, top k retrieval [51], [53], [93], [111], [11], [81], [76], [33], [96], [52], [74]
5.5 Range next value, range intersect [51], [113], [114]
5.6 Posting lists [51], [95], [3], [2]
5.7 XML retrieval [51], [13], [1], [42], [41]

Fig. 1. A tree diagram of the paper's structure

Before proceeding, it would be interesting for the reader to have as motivation, during reading the paper, a practical application. For example, consider that the access log to a set of Web sites is stored as a sequence of Web pages, ordered by their access time, with each page being mapped to a distinct symbol, and that the length of the sequence is *n*. That is, the produced sequence *S*, is of the form, $S=p_1,p_2,\ldots, p_n$ where $p_i \epsilon \Sigma$, and Σ denotes the alphabet of symbols. Then the wavelet tree can be used to support two elementary operations: (1) locate the number of occurrences of a given page, in a given position of the sequence (starting counting from the beginning of the sequence), and (2) locate the position in the sequence of the i-th occurrence of a given page. Using these two elementary operations, it is possible to answer interesting queries in the available sequence, such as counting how many times a given page was visited in a period of time, find the most visited pages, and queries of similar nature that could be a value to Web designers and/or Web miners. It would be helpful for the reader to have this simple example as a working exercise while reading the paper, and find out how wavelet trees efficiently handle the described operations.

We organize the whole material in the following sections: in section 2 we present the main characteristics of the structure, in section 3 we describe its connection with rank and select structures, in section 4 we describe various of its applications in data compression, and in section 5 we describe its main applications in information handling (information retrieval, data retrieval and spatial objects retrieval). Finally in section 6 we conclude with discussion and open problems. In figure 1 we provide the reader with a tree diagram depicting the structure of the paper, in relation with the various application areas that are covered and the references used in them.

Before proceeding and in order to make the paper self contained we provide some definitions that will be used continuously in the sequel. Consider a sequence *S* of *n* symbols from an alphabet $\Sigma=\{c_1,\ldots,c_\sigma\}$ of cardinality σ, we call as *entropy H* the sum[1]:

$$H=\sum_{i=1}^{\sigma} p_i \log \frac{1}{p_i}$$

where p_i, is the probability of appearance in *S*, of the *i*-th symbol in the alphabet. According to the coding theorem of Shannon [107], this notion of entropy represents a lower bound to the average numbers of bits needed to represent each symbol in *S*, provided that the symbols appear independently. This probabilistic notion of the entropy that takes into account the statistical nature of the source, is usually replaced in the scientific literature by the most practical notion of *empirical entropy*.

In particular the zero-order empirical entropy of *S* is defined as:

$$H_0 = H_0(S) = \sum_{c_i \in \Sigma} \frac{n_i}{n} \log \frac{n}{n_i}$$

where n_i is the number of appearances of character c_i in *S*. If we take also into account the context of appearances this definition can be extended to the so called *k*-th order empirical entropy. For a string $w \in \Sigma^k$ let us denote with w_S the subsequence of characters that follow[2] *w* in *S*, then the *k*-th order empirical entropy of *S*, is defined as follows:

[1] Throughout the paper, the symbol *log*, will denote base 2 logarithms.

[2] In some works they refer to the set of characters that *precede* w, though this seems counterintuitive, it helps in the analysis; either way the difference is small [43].

$$H_k = H_k(S) = \frac{1}{n}\sum_{w \in \Sigma^k} |w_S| H_0(w_S)$$

Both zero-order and the *k*-th order empirical entropy are more practical measures when one needs to lower bound the compressibility of a text (the first when considering symbols as independent from each other, and the second, when taken into account the size *k* context, where the symbols appear), since they do not need to make any assumption concerning the probabilistic nature of the source of the text.

A related notion is the *modified empirical entropy,* a notion that is helpful for highly compressible text [39]. The zero-order modified empirical entropy is defined as:

$$H^*{}_0(S) = \begin{cases} 0 & \text{if } |S| = 0 \\ (1 + \lfloor \log |S| \rfloor) / |S| & \text{if } |S| \neq 0 \text{ and } H_o(S) = 0 \\ H_0(S) & \textit{otherwise} \end{cases}$$

A set S_k of substrings of *S* of length at most *k* is termed a *suffix cover* if any string in Σ^k has a unique suffix in S_k; in this case define:

$$H^*{}_{S_k}(S) = \frac{1}{|S|}\sum_{w \in S_k} |w_S| H^*{}_0(w_S)$$

Then, the *k*-th order empirical entropy of *S* is defined as:

$$H_k{}^*(S) = \min_{S_k} H^*{}_{S_k}(S).$$

2. Description and Main Issues

2.1. General characteristics

The wavelet trees have their roots in a range searching data structure described in a paper by Bernard Chazelle [20] and were introduced as a distinct structure in [64], as a component of a self-indexed compressor based on suffix arrays, and afterwards a set of other papers explored and extended their use (the majority of the papers referred to in the bibliography listing). A wavelet tree acts basically as a mechanism that permits the efficient implementation of specific operations on sets of objects from non-binary alphabets, via the use of the respective bit vector operations on sets of objects from the binary alphabet, with a logarithmic slowdown in the attained performance.

The main advantage of this transformation, that from a worst case point of view and from a query time perspective is inferior to other solutions proposed for these operations on non-binary alphabets, is that this transformation can embed entropy bounds in the attained space complexity. It thus can be used as a general tool for reducing computationally the compression of a string from an arbitrary alphabet to the compression of binary strings.

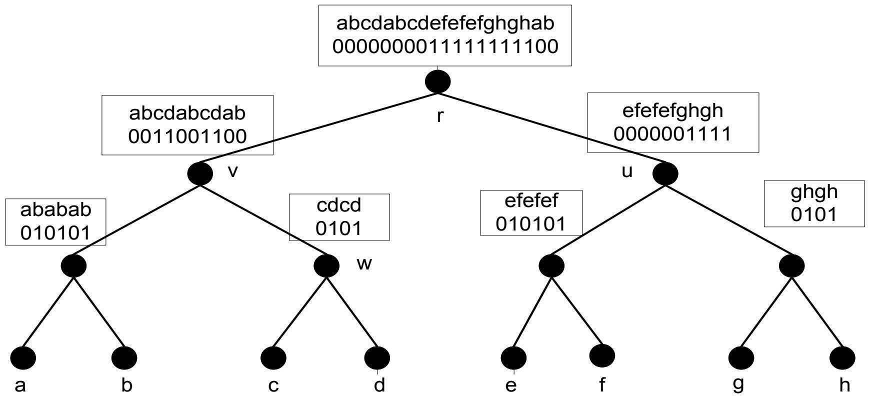


Fig. 2. The wavelet tree for S=abcdabcdefefefghghab, Σ={a,b,c,d,e,f,g,h}

A wavelet tree can act both as a compression mechanism on its own, and as a component of a compression algorithm; that is a wavelet tree can simply store a text by itself [38] and provide both an indexing and a compression mechanism or it can be used as a data structural component storing auxiliary information for a compression algorithm [4], [94].

More analytically consider a sequence S, of length n where the elements belong to an alphabet Σ, of size σ. A wavelet tree T is a static complete binary tree whose leaves store the symbols of the alphabet Σ in order from left to right (that is the number of leaves of the tree is equal to the size of the alphabet) and where each internal node v corresponds to a subsequence S_v of S that is stored through a binary vector B_v. The subsequence S_v is formed by the symbols of the sequence S that are stored in the subtree with root v, while the binary vector B_v has the same length as S_v, and each of its positions corresponds to a distinct symbol in the specific subsequence, such that $B_v[i]=0$ if the i-th symbol of S_v is stored in the left subtree of v, and $B_v[i]=1$ if the i-th symbol of S_v is stored in the right subtree of v.

If we state this recursively each subtree T_v of T is itself a wavelet tree for S_v. As an example see figure 2 for an exposition of a wavelet tree for the sequence $S=abcdabcdefefefghghab$, with alphabet $\Sigma=\{a,b,c,d,e,f,g,h\}$ where at each node we depict the subsequence with the accompanying bit vectors. For example, in the root r since a, d are stored in the left subtree the respective bits are set to 0, while the bits for e and g are set to 1, since they are stored in the right subtree. In node v the symbol a is stored in the left subtree, hence its corresponding bit have value 0, while symbol d being stored in the right subtree has its corresponding bits being set to 1. Finally in node u the symbol e is stored in the left subtree hence its corresponding bits have value 0, while the symbol g being stored in the right subtree has its corresponding bits being set to 1.

In each node of the wavelet tree, the bitvector B_v is stored in a structure suitable for accessing any of its stored elements, and answering efficiently rank and select queries [25], [79], [99], [103]. More analytically, rank and select queries on bit vector are defined as follows:

- $B_v(i)$, for returning the i-th bit of the bit vector B_v,
- $rank_b(B_v,i)$, that returns the number of times bit b appears in the prefix of B_v consisting of its i first symbols and
- $select_b(B_v,i)$, that returns the position of the i-th appearance of bit b in B_v.

Attaching these structures to each node we can use the wavelet tree to answer rank and select queries for any symbol in the sequence S, with a time complexity that is equal to the depth of the tree (that is $log\sigma$) multiplied by the time to perform a rank and select query in a bit vector of size at most n. Note, that the rank and select queries for S are defined similarly to their binary counterparts that is:

- $rank_c(S,i)$, return the number of times symbol c appears in the prefix of S consisting of its i first symbols, and
- $select_c(S,i)$, that returns the position of the i-th appearance of symbol c in S)

In order to answer the rank query $rank_c(S,i)$, we start from the root r of the wavelet tree and if c is stored in the right child we compute as new i the value $rank_1(B_r,i)$ otherwise the new i is the value $rank_0(B_r,i)$, and the procedure continues till reaching a leaf. On the other hand and for the query $select_c(S,i)$ we move bottom up, starting from the leaf corresponding to c, and if it is a right child the corresponding position to its parent v is $select_1(B_v, i)$ otherwise it is $select_0(B_v,i)$; the procedure moves similarly upwards until reaching the root, where the answer is reported.

We provide a toy example of how these algorithms work by treating queries $rank_d\ (S,9)$ and $select_d\ (S,2)$. The procedure for $rank_d\ (S,9)$ starts at r. Since d is stored in the left subtree, we compute $rank_0(B_r, 9)=8$, and we go to v. Since d is stored in the right child of v, we compute $rank_1(B_v, 8)=4$, to locate the proper position in the right child and move to w, there d is stored as a leaf in the right child of w, and hence the answer is $rank_1(B_w,4)=2.$ In order to answer $select_d\ (S,2),$ we start from the leaf of the tree corresponding to d and we move upwards. Since the leaf storing d is at the right child of w, the new position is $select_1(B_w,2)=4$. Moving to v, since we come from the right child the new position is $select_1(B_v,4)=8$. Finally, at the root the seeked position is $select_0(B_r,8)=8.$

If the bitvectors are implemented using the structure in [25], then the overall space complexity of the construction is $nlog\sigma(1+o(1))$ bits, while the time to answer a query is $O(log\sigma)$. If we use for the implementation of the rank and select structure the construction described in [103] then the space complexity becomes entropy bounded by $nH_0(S)+O(nloglogn/log_\sigma n)$ bits, while the time complexity remains $O(log\sigma).$

There is a deep connection between the aforementioned structure and a construction proposed by Chazelle for range searching in [20], this connection will appear more clearly in the following sections, where spatial searching applications are presented. We explore this connection, that was noticed also by various researchers, see e.g. [87], more deeply.

Therefore, consider a set of points in the *xy*-plane with the *x*- and *y*-coordinates taking values in *{1,…,n}*; assume without loss of generality that no two points share the same *x*- and *y*-coordinates, and that each value in *{1,…,n}* appears as a *x*- and *y*- coordinate. We need a data structure in order to count the points that are in a range $[l_x, r_x] \times [b_y, u_y]$ in time *O(logn),* and permits the retrieval of each of these points in *O(logn)* time. The structure proposed in [20] for this problem needs *nlogn(1+o(1))* bits, and we will describe it as in the version provided in [87], which is simpler, and more clearly described; this structure is essentially equivalent to the wavelet tree. The structure is a perfect binary tree, according to the *x*–coordinates of the search points, with each node of the tree storing its corresponding set of points ordered according to the *y*-coordinate. In this way the tree mimics the distinct phases of a mergesort procedure that sorts the points according to the *y*-coordinate, assuming that the initial order was given by the *x*-coordinate. This construction takes *O(nlogn)* space, that can be reduced to *nlogn(1+o(1))* bits by the *functional approach* introduced in [20] that principally states that each node of a data structure does not need to store the sets of points that correspond to it, but needs only to provide the means (functions) for computing them. Hence each node instead of the real points stores a bitmap B_v with $B_v[i]=0$ iff the *i*-th point in the subsequence mapped to *v* belongs to the left child, otherwise $B_v[i]=1$. This bitmap is preprocessed, for the computation of rank and select queries in constant time.

Using this repertoire of structures we can retrieve the identity of any point, giving a specific node, and a specified bit of its bit vector, by descending a path in the tree to a respective leaf by performing suitable rank queries; the time cost is *O(logn)*. In order to count the points that are in the range $[l_x, r_x] \times [b_y, u_y]$ we just need to find in the tree by two searches the O(log*n*) maximal tree nodes that cover $[l_x, r_x]$, and for each of these nodes perform a subtraction in the result of two rank queries (for more details see [20] and [87]). If we do not want to count but also to retrieve the points we have to identify each of these points, paying an extra logarithmic term for each of them. The aforementioned construction, that works for distinct sequences can also be extended for handling arbitrary sets of points in a continuous space, as shown in [19] .

The connection between wavelet trees and range searching was more clearly depicted in [12] and in [37]. In [12] a solution was presented for storing *n* points with *x*- and *y*- coordinates from the rank space, using *nlogn+o(nlogn)* bits, supporting range counting in *O(logn/loglogn*) time and range reporting in *O(klogn/loglogn)* time, where *k* is the size of the output; a basic ingredient of this solution is a multi-ary wavelet tree. In [37] the entropy of the stored two dimensional set came into play. In particular consider a set of *m* points with *x*- and *y*- coordinates from the space *{1,…n};* the "entropy" *H* of this set is defined as the total number of the different possible grids that is: $\log\binom{n^2}{m}$.

It was proved in [37] that there exists a representation of this set, that takes $H+O(H)$ bits of space; the proposed representation employs a block partitioning of the points combined with wavelet trees for handling these partitions. The solution can answer range counting queries in $O(log(n^2/m))$ time, range reporting queries in $O(log^2(n^2/m))$ time per reported element, while point selection queries require $O(log^2n)$ time. By taking into account the density of the involved rank space it is possible to have the various times reduced down to $O(1)$ per reported element.

Wavelet trees have been mainly exploited as auxiliary structures for self-indexes based on the Burrows-Wheeler transformation [38], [40], [43], [44], [94]. In these algorithms it has been shown that indexing compressed text is reduced to performing rank and select operations in the sequence that is produced by applying the Burrows-Wheeler transformation on the text. Hence, these works have shown that the Burrows-Wheeler transform is not only useful for compressing a given sequence but it also allows to support search operations, if rank queries are supported over BWT. In these cases the binary vectors of the wavelet tree are compressed using run length encoded or gap encoding [38]. The application of run length encoding[3] compresses the sequence to its *k*-th order entropy, while the gap encoding does not achieve analogous bounds. However it is possible to apply gap encoding in combination with the compression boosting technique of [39] and have a compression algorithm of size $2.2618|S|H^{*}_{k}(S)+log|S|+\Theta(|\Sigma|^{k+1})$ bits, for any positive *k*, thus improving the bound achieved in [39] into an almost optimal result. It should be noted that if the wavelet trees store the initial sequence (not its Burrows-Wheeler transformation) then they can be considered as standalone general purpose compressors.

Concerning the performance of wavelet trees as standalone compressors, it was noted in [38] that their use reduces the problem of compressing a string to that of compressing a set of binary strings, with the specification of the set being determined by the topology of the underlying tree structure, and the coding of the alphabet symbols at the tree leaves. The authors pursue their remarks even further by introducing the paradigm of the *generalized wavelet tree* and noting that it is possible by pruning whole subtrees of the initial tree, to have a mixed compression strategy where only some strings are binary and the others use symbols that are compressed with general purpose order zero compressors (for example Huffman or arithmetic coding). The authors apply this pruning strategy, by developing a combinatorial optimization framework based on the notion of leaf covers and by providing a polynomial time algorithm for selecting the optimal tree shape for some special cases.

In [44] it has been shown that run-length compression on the wavelet tree is not mandatory, if the wavelet tree is applied not to the whole Burrows-Wheeler transformed text, but separately to each one of the partitions of the

[3] The casual reader should be careful here to note that run-length compressing or gap-compressing a wavelet tree, means compressing *each* bitmap and *not* the whole sequence, before building the wavelet tree.

transformed text (provided an optimal partitioning has been found). The technique is called *compression boosting* [39] and permits the formation of *k*-order compressors, by using 0-order compressors. If compression boosting is used then the bit vectors in the nodes of the used wavelet trees, can be compressed by simply using the structure in [102]. This technique was improved in [88] (see also [86]) where it was shown that there is no need to apply the wavelet tree approach to every piece produced by the partitioning provided by the Burrows-Wheeler transform, but one can simply use the wavelet tree (using however [103] for representing the bit vectors) for the complete Burrows-Wheeler transformed text. In essence, the technique provides a way to do compression boosting implicitly, with a trivial linear time algorithm, but using a specific zero order compressor [103].

There have been proposed also variants of the above compressed representations of the wavelet tree, that exploit the observation, that the binary skeleton of a wavelet tree, if labeling the left edge of each node with 0 and the right edge of each node with 1, is basically a trie storing the binary string representations corresponding to each distinct symbol of its leaves. Henceforth it could be possible to achieve greater compression, by using instead of the binary representation for each symbol its code according to a prefix free code such as Huffman code (this appeared in [14], [28], [49], [68], [88]). In this case the tree is not balanced, and hence the time complexity of the query operations is no longer $O(log\sigma)$, but can be as bad as $O(\sigma)$; however the average query time is bounded by $O(log\sigma)$ since it is bounded by the entropy plus one, and the space needed to store it is reduced, since the average length of the paths to reach each leaf nodes is smaller.

Extending this approach we can use the wavelet tree by codifying words instead of characters and employing multi-ary wavelet trees. An example is provided in [14] where the wavelet tree is built over the codes associated to each word of the text using End-Tagged Dense Code (ETDC). This code produces sequences of bytes, and the constructed wavelet tree is no longer binary but multi-ary since the label of each edge equals a byte (the edge from the root signifies the first symbol) and each node can have up to 256 children, moreover each node will contain now a byte vector instead of a vector of bits. In [14] it is mentioned that, when working with natural languages, the codes generated by ETDC will never have length greater than 4, and hence the produced wavelet tree will have at most four levels.

It should be noted here that wavelet trees can be efficiently externalized and used in a set of applications in external memory, see for example [75] and [62]. In [75] it has been shown how the wavelet tree can be externalized optimally (linear number of blocks in space consumption, and an *logB* speedup in answering queries) by replacing the binary tree skeleton and the accompanying bit vectors by a *B*-ary tree, augmented with vectors of characters each with *logB* bits, while in [62] and in the context of presenting a practical self-index for secondary memory based on the FM-index alternative secondary memory implementations, were discussed.

2.2. Practical Issues

Concerning practical implementation and experimental outcomes of the wavelet tree a set of results have been published in [49], [28], [68]. In particular in [49] wavelet trees are implemented using mainly run length encoding and γ encoding for codifying the involved bit vectors. Moreover Huffman shaped wavelet trees, are involved with the provision of applying fractional cascading, in order to access efficiently the various dictionaries. An interesting aspect of the [49] construction is that besides using Huffman shaped wavelet trees (note that in this case of already compressed bit dictionaries variations in the tree shape does not seem to affect the total space) a frequency based construction is applied where the wavelet tree is built as an optimal Huffman prefix tree not on the frequency of appearance of the symbols, but on the a-priory distribution of the queries on them.

Experiments are performed that depict the usefulness of the frequency based heuristics, in search scenarios, though the fractional cascading does not seem to lead to considerable improvements. Moreover the proposed scheme was applied to the Burrows-Wheeler streams of various text files from the Canterbury and Calgary corpora and besides γ coding the following codes (for the run lengths of the bit vectors of the wavelet tree) were used: δ code, Golomb code, Maniscalo code, Bernouli code, or MixBernoulli code. It was depicted that run length encoding in combination with γ coding constitutes a simple solution capable of providing efficient compression.

In [28] a practical implementation of the [103] bit vector dictionary is provided, and its use in wavelet trees is experimentally tested. In particular the [103] dictionary is experimentally tested against the dictionaries described in [58] and it is shown that for uniformly and independently distributed bitmaps the dictionary in [103] is superior for high density data. Moreover by using [103] to implement the bit vectors at the nodes of a balanced wavelet tree of the Burrows Wheeler transform of natural language english text, it is proved that this structure is clearly superior to the [58] implementations; this is natural since the [103] structure exploits local regularities in the bit sequence. Moreover efficient implementations of the Huffman-shaped wavelet trees are provided that use no pointer information (or just $log\sigma$ pointers) while the bitmaps at each level are stored concatenated. Experiments performed in various datasets show that the [103] structure in combination with a Huffman shaped wavelet tree (with or without pointers) provide excellent compression, and helps reaching entropy bounds, even for very large alphabets.

Finally, and extending the findings of [49], [28] a set of alternatives concerning the tree topology and the implementation of the accompanied bit vectors is explored in [68]. The authors initially prove theoretically that for data of low entropy (formally when the 0-th order empirical order is asymptotically less than $log\sigma$), the run length δ coding is superior to the run length γ encoding, achieving $nH_0(S)$ (plus lower order terms) compression with leading constant 1, while run length γ encoding has leading constant 2. Moreover a software package is presented with a parametric implementation of wavelet trees able to embed a large number of available options both in the tree topologies and in the involved bit dictionaries. The various topologies explored are: a balanced shape, an alphabetic weight balanced wavelet tree [10], and a Huffman shaped wavelet tree; moreover the attached bit vectors are implemented as: the structure in [103], run length γ-encoding [112], run length δ-encoding [112], pure arithmetic coding [112], selective compression at the lower levels of the tree, and compression with *t*-subset encoding.

A set of experiments are performed that depict that run-length encoding gives the more space efficient implementation especially when compressing Burrows Wheeler transformed text and low entropy data; they show that generally run length γ encoding is superior but for very low entropy data run length δ encoding becomes better. However the query performance of the run length implementation is quite poor, due to the need for handling the run length encoding. In these cases the weight balanced implementation using [103] or with compression at the lower levels are superior. Moreover concerning building times, run length compressed wavelet trees are the slowest, the other implementations are competitive to each other, and all implementations take advantage of Burrows Wheeler transformed text that is amenable to faster compression. Finally, and for applications where a good tradeoff is what is needed, Huffman shaped wavelet trees and weight balanced wavelet trees in conjunction with the structure in [103], are proved to be the best choice, a fact that revalidates the findings of [28].

In [49] and [31] the issue of efficiently building a wavelet tree was handled. In particular in [49] a time efficient algorithm is presented for constructing a wavelet tree for the Burrows-Wheeler output of a sequence S of size n in $O(n+min(n,H_k(S))log\sigma)$ time; the bit vector dictionaries are run length γ encoded and are concatenated together to heap order. The method is a bottom up procedure, that traverses the tree upwards, processes consecutive runs of equal symbols and extends appropriately the bitvectors of each node. The specific construction depicts that data that are highly compressible can be indexed faster.

In [31] the focus is on space efficient implementations, and novel algorithms are presented for constructing wavelet trees with virtually no space. The construction works for both binary and multi-ary wavelet trees, however it does work only for uncompressed wavelet trees (the bit vectors are not compressed in contrast to [31]), and extending their techniques to compressed wavelet trees, is left as an open problem. The techniques used are based on in place sorting and exploit properties of permutations, that permit the execution of the necessary *partitioning* (moving top down) and *merging* (moving bottom up) steps of the construction algorithms, without using extra storage. Two classes of algorithms are presented *non-destructive*, that do not alter the initial array of symbols and *destructive* that alter it, and use it for storing the created dictionaries of the tree nodes.

In particular for a bit vector of size m, let $C(m)$, $E(m)$, $S(m)$ be the construction time, extra bits required for construction and total space occupation in bits respectively. It should be noted here that these space and time complexities when constructing the bitmap refer to the complexities of the *extra* data structures for supporting rank and select on the bit dictionaries and *not* on the space of the bitmap itself. The authors initially describe a non-destructive algorithm for a binary wavelet tree that needs $O(n\log\sigma + C(n\log\sigma))$ time and $O(\log n\log\sigma)+E(n\log\sigma)$ bits beyond the space for the array of symbols and the tree nodes. Then they describe destructive algorithms and show that it is possible to store a sequence of *n* symbols in a binary wavelet tree in (i) $O(n\log n\log^2\sigma)+C(n\log\sigma)$ time using $O(\log n\log\sigma)+E(n\log\sigma)$ extra bits beyond the space required for the tree, (ii) $O(n\log\sigma+C(n\log\sigma))$ time and $n+O(\log n\log\sigma)+E(n\log\sigma)$ extra bits beyond the space required for the tree. Moreover if incrementally constructed the $O(\log n\log\sigma)$ bits term in the space bound can be replaced with $O(\log n)$ bits.

3. Wavelet Trees and their Use as a Rank and Select Structure

3.1. Static Solutions

Wavelet trees, though having many uses, can be considered mainly as rank and select data structures for large alphabets and during their lifetime have been used as component substructures in various solutions to the rank and select problem. We remind the reader that in that problem we are given a sequence *S* of *n* symbols from an alphabet *Σ* of size *σ,* and we want to access the *i*-th element of the sequence, and answer, for every symbol *c* in *Σ,* the following queries: $rank_c(S,i)$ (that returns the number of times symbol *c* appears in the prefix of *S* consisting of its *i* first symbols), and $select_c(S,i)$ (that returns the position of the *i*-th appearance of symbol *c* in *S*).

In [79] the problem was handled for the case of sequences from a binary alphabet and a data structure was presented that needed $n+o(n)$ bits and supported both rank and select in $O(\log n)$ time. This solution was extended and improved in [25] and [92] where using the same space, both rank and select operations were handled in O(1) time. We will discuss these solutions following the presentation given in [25]. The general idea behind the construction is to store in tables, precomputed answers for the query arguments, and then at query time, just retrieve in O(1) time, with a simple lookup, the appropriate answer. If this idea is applied naively then the space complexity will be $O(n\log n)$ bits, but this can be overcome by precomputing answers for properly chosen samples in the query ranges and by employing multilevel schemes.

This multilevel scheme could lead to a non-constant query time, but when the size of the treated bit vector falls bellow $\log n/c$, (*c* designates a constant greater than 1) then, all possible subsets, can be treated with a global lookup table that needs $o(n)$ space and answers queries in constant time. These constructions as simpler for rank than select, and one should build the needed tables separately for rank and select queries. As in [94] we will focus on the *rank*$_1$ and *select*$_1$ queries, *rank*$_0$ follows from *rank*$_1$, while *select*$_0$ can be computed as *select*$_1$.

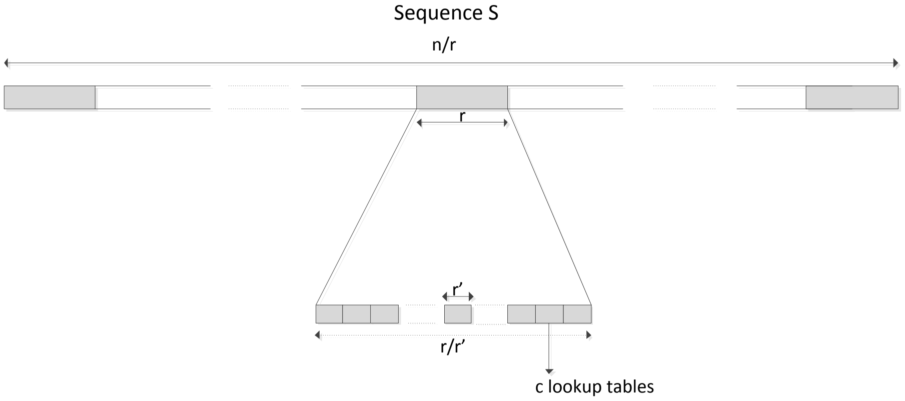


Fig. 3. *The rank multilevel structure*

In the rank case a multi-level directory scheme (see figure 3) is employed where the first level divides the sequence *S*, into consecutive chunks of size $r=\lceil logn\rceil^2$, and stores precomputed answers for arguments that are multiples of *r.* The second level deals separately with each chunk, dividing it in consecutive subparts of size $r'=\lceil logn\rceil$, and storing precomputed answers for arguments that are multiples of *r'.* Finally each subpart having size *r'* is divided into a constant number *c* of subsegments of size *r' /c*. These segments are handled by a global lookup table that for each possible segment (there are $2^{r'/c}=n^{1/c}$ of these) and for each possible argument stores the answer (this technique can be considered [94] as a different application of the classical *four-Russians* technique).

In order to answer rank queries we first locate, with two table lookups, the correct entries in the first and the second level, and the sum of the appropriate values gives an *r'*-sized range. There we locate in constant time the appropriate subsegment, and using the global lookup table produce the rank in the segment (adding at most *c* values). Overall the query time is constant while the space consumption is: *O(n/logn)* bits for the first level, *O(nloglogn/logn)* bits for the second level, and $O(n^{1/c}lognloglogn)$ bits for the global lookup table, thus in total *o(n)* bits.

Considering now the *select*$_1$ operation, it is a folklore solution to solve it by *O(logn)* calls to *rank*$_1$ with each call taking *O(1)* time. We can solve the problem by employing again a multilevel dictionary structure that is a bit more complicated than the construction for the *rank*$_1$ query. The first level records the position of every $l_1=\lceil logn\rceil\lceil loglogn\rceil$ 1 bits. If the size of a range *r* in the initial vector between two of these 1-bits is greater than $(l_1)^2$ then the answers are explicitly stored, otherwise the range is partitioned by storing the relative position of each $l_2=\lceil logr\rceil\lceil loglogn\rceil$ 1-bits. If the size *r'* of a subrange is greater than $\lceil logr' \rceil \lceil logr \rceil \; \lceil loglogn \rceil^2$, then we store them explicitly, otherwise we note that the size of the range is less than $(loglogn)^4$. For these small ranges we can again employ a trick similar to the four-Russians technique and a lookup table suffices for performing select in constant time. Hence by a constant number of table lookups we can compute the answer to the select operation.

This construction was improved in [99] and [103] with a data structure of entropy bounded size $nH_0(S)+o(n)$ bits that could answer queries in constant time. In particular the idea is to replace the original bit sequence representation of $O(n)$ bits, with an entropy bounded representation, upon which the extra information and techniques needed by the [25] and [92] of O(1) query time tank and select, is employed. In particular the original sequence is divided into chunks of $b=(logn)/2$ size. Each of these chunks can belong to one of $(logn)/2$ classes according to its number of 1-bits, and a class with identifier j (that is having j bits set), can contain at most $\binom{b}{j}$ elements (that is, different chunks). A chunk therefore can be coded with two numbers: its class and its position inside its class. It is proved in [25] and [92] that the total storage of this representation is $nH_0(S)+o(n)$ bits, while the constant time query complexities for rank and select remain unaffected, by employing the same techniques but on the new representation.

In [90] it is shown how to implement efficiently the rank and select operations on top of sparse bit sequences using the gap encoding technique (see also [69]); in particular the gap between consecutive 1's was encoded using arbitrary random access self-delimiting integer codes. The provided construction needs constant time for both queries and takes $l\alpha log(n/l)+O(l)+o(n)$ bits of space, where α is a constant depending on the coding used and l is the number of 1 bits. In the same paper, a new problem is also introduced, the so called *position restricted substring searching*, and based on this, the rank and select operations are extended by taking as input not a single symbol but a whole string, while the presented solution entails the heavy use of wavelet trees. In particular if the text is of length n, from an alphabet of size σ, and the searched substring is p, then the provided solution needs $O(ntlog\sigma)$ bits of space and supports the various rank and select operations in $O(|p|log\sigma/loglogn)$ time; here t is a given construction parameter such that $|p|<t$.

Finally, and still remaining in binary alphabets, in [98] several practical alternatives were presented, while in [100], [55], [56], [66] a suite of solutions with various tradeoffs concerning the space and time complexities were presented. In particular, the sparse problem where the number of ones is small attracted attention; note that if the sequence contains k 1's then the optimal space consumption is: $B = \log\binom{n}{k}$.

The best solution was provided in [100] depicting that $O(t)$ query time can be attained using[4] $B+n/(\log n/t)^t+\tilde{O}(n^{3/4})$ bits of space, while in [66] a parametric construction was provided that for $0\leq\delta\leq 1/2$, $0\leq\varepsilon\leq 1$, (δ, ε are any real constants), and positive integer s, has $O(s\delta^{-1}+\varepsilon^{-1})$ query time and uses $B+O(k^{1+\delta}+k(n/k^s)^{\varepsilon})$ bits.

Moving now to the case of large alphabets, wavelet trees provide a solution that needs $nH_0(S)+o(n\log\sigma)$ (or $nH_0(S)+o(n)\log\sigma$) bits of space and $O(\log\sigma)$ time for rank and select. This construction was improved in [45] by designing efficient structures for alphabets of small size and then moving to alphabets of larger size by employing properly defined multi-ary wavelet trees. In that paper initially a solution using $nH_0(S)+O((\sigma n\log\log n)/\log_\sigma n)$ bits was provided answering queries in constant time, when the alphabet size is of the order $o(\log n/\log\log n)$.

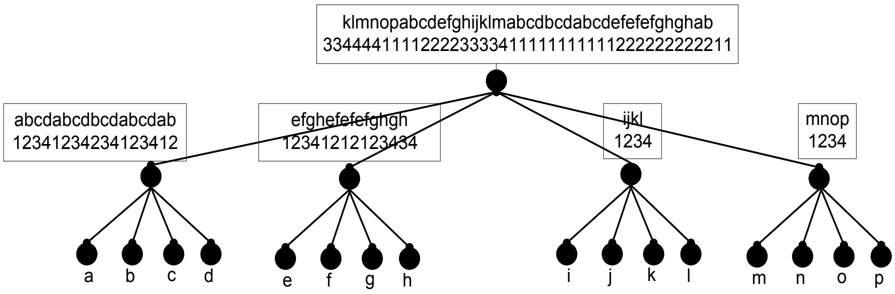


Fig. 4. *The 4-ary wavelet tree*

The construction is inspired by the ideas of [99] and [103], concerning the representation used for getting entropy bounds. The original ideas are delicately modified in order to handle non-binary alphabets, and are carefully tuned in order to attain efficient compression, this is the reason for the upper bound on the alphabet size. Let $\Sigma=\{c_1, .., c_\sigma\}$ be the alphabet. The basic idea, as in [99] and [103], is to separate the sequence S into chunks of size $\lceil(\log_\sigma n)/2\rceil$, and codify the classes of the chunks according to the participation of the symbols in the chunk; a *symbol decomposition* is defined as an σ-tuple $(n_1, n_2, \ldots, n_\sigma)$, where n_i is the frequency of appearance of symbol c_i. Hence a chunck is codified by its symbol composition, and its unique relative position among all the chunks having the same symbol composition. Moreover precomputed tables store the answers for all possible combinations of rank and select queries, while the rank and select procedures follow the same logic as previously described.

The encoding of the sequence using the block decomposition and the relative position of the block in its symbol decomposition gives rise to the $nH_0(S)$ space complexity, with the various other tables employed needing a total of extra $O((\sigma n\log\log n)/\log_\sigma n)$ bits; these extra bits are the reason for the

[4] With the Õ notation poly-logarithmic factors are ignored, that is $\tilde{O}(f(n))=O(f(n)\log^{O(1)}n)$

constraint on the size of the alphabet, since larger values could make the space complexity larger than $nlog\sigma$, which is the uncompressed space complexity.

The result is then generalized to alphabets of arbitrary size by employing a multi-ary generalization of the binary wavelet tree. In particular an *h*-ary tree is used with height $1+log_h\sigma$, with each node of the tree containing instead of binary vectors, vectors of integers in the range *[1,h].* Let *v* be a node with children $v_1 \ldots v_h$, then the subsequence S_v is stored in an array A_v such that $A_v[j]=i$, iff the *j*-th symbol in the sequence is stored in the *i*-th subtree of *v*. All the arrays at a specific level are physically stored concatenated, and they are handled by using the rank and select structures for the small alphabet case. In particular the rank query is implemented as in the binary tree case by starting from the root moving to the proper child in O(1) time, and calculating as new argument of the rank operation the result of the rank at the root, the same procedure is followed until reaching the appropriate leaf. Similarly and for the select operation we start from the leaf of the given symbol, and move upwards by setting as new argument the result of the select operation at the current node; the answer is the value returned at the root of the tree. Finally the value of *h* is set to be equal to $O(log^{\delta}n)$ with $\delta < 1$, and since descending each level takes constant time, it follows that the whole construction uses $o(nlog\sigma)$ extra space over the $nH_0(S)$ term and $O(log\ \sigma/\ log\ log\ n)$ query time. In figure 4, a 4-ary wavelet tree is depicted where for reasons of simplicity we depict the arrays at each level separately at each node.

Besides wavelet trees, in [57] two data structures were proposed, one that can access an element and handle the rank operation in $O(log\ log\ \sigma)$ time, and the select operation in O(1) time, using $nlog\sigma + no(log\sigma)$ bits of space, and another that supports the operations in the same time, but using space $nH_0(S)+O(n)$ bits. The main idea is to reduce the problem over one sequence to chunks of length σ, handle efficiently each chunk, and represent in a unified way all the chunks together.

As mentioned, in [28] a practical study was presented on the compact representation of sequences supporting rank, select, and access queries. This paper implements the structures described in [103] and in [57] (the first such implementation at the time [28] was written) and offers a well tuned implementation of Huffman-shaped wavelet trees showing that this provides an excellent solution for representing a sequence within zero-order entropy bounded space, even when the alphabet size is very large.

In [36] different implementation strategies are presented (without however providing new complexity bounds), while in [8] (see also [46]) a structure was provided using $nH_k(S)+o(nlog\sigma)$ bits supporting access/rank/select operations in $o((loglog\sigma)^2)$ time, but without using wavelet trees. In particular in [105], [46], [61], general techniques are presented for compressing general sequences to the *k*-th entropy bound, without affecting the query complexities; these constructions permit the compression of the various rank/select structures to $nH_k(S)+o(n)$ bits.

Finally interesting constructions are provided in [6] and a novel technique of combining previous results on rank and select data structures was presented in order to handle large alphabets. The specific technique uses the frequencies of appearances of the characters, in order to divide the alphabet into subalphabets, and produces a new string by replacing each character on the original string by identifiers of their subalphabets; this string is stored in a multi-ary wavelet tree. Moreover the projections of the original string to the characters belonging to each sub-alphabet, are stored using the structure in [57] for large alphabets.

The outcome of this construction is a data structure that stores a string of n characters over an alphabet of size σ, in $nH_0(S) + o(n)(H_0(s)+1)$ bits. This structure supports the operators access and rank in time $O(loglog\sigma)$, and the select operator in constant time; while it can be extended and be easily implemented. The technique can be improved achieving $nH_k(S)+o(n)log\sigma$ bits space complexity and needing poly double logarithmic time, for the query time. Moreover the construction has applications to the design of full-text self-indices, to the representation of compressed permutations, and to efficient handling the compressed representation of dynamic collections of disjoint sets.

3.2. Dynamic Solutions

In the dynamic version of the problem, the rank and select operations are extended by the update operations of inserting an arbitrary object (bit/symbol) between two specified objects in the sequence, and by deleting an object (bit/symbol) from the sequence. The problem was dealed with in [73] for binary sequences with a solution requiring $n+o(n)$ bits of space, $O(b)$ amortized update time, and $O(log_b n)$ time for rank, select, where $b=\Omega(polylogn)$.

In [86] the aforementioned solution was improved producing a construction able to handle binary strings in $nH_0(S)+o(n)$ bits of space, and $O(logn)$ worst case time complexity for all operations. In that paper the dynamic solution was generalized to symbols in an arbitrary alphabet by using the wavelet tree machinery and preprocessing the bit vectors at each level of the tree structure with the dynamic structure for handling bit sequences. The result is a structure that uses the same $nH_0(S)+o(nlog\sigma)$ bits of space, and supports all the operations in $O(lognlog\sigma)$ time (that is the use of the wavelet tree, as anticipated, incurred a $log\sigma$ slowdown in the operations time complexity).

In particular the authors consider firstly the case where the sequence is binary, and they sketch a simple solution with $O(n)$ space and $O(logn)$ worst case update and query time. They simply divide the sequence into blocks of $O(logn)$ bits, attach every block to a leaf of a balanced binary tree, store at each internal node the size and number of bits set in the respective subtree (see figure 5 for a snapshot of the structure), and they merge/split leaves when the capacity of blocks shrinks or expands at $(1/2)logn$ and $2logn$ respectively. However this simple solution is not (asymptotically) entropy bounded and is not even succinct. In order to make the structure succinct the authors employ a two level blocking scheme at the lower levels of the tree.

Now a leaf is attached to a *superblock* that is equivalent to a group of simple blocks (though it can be considered simply as a block of a larger capacity).

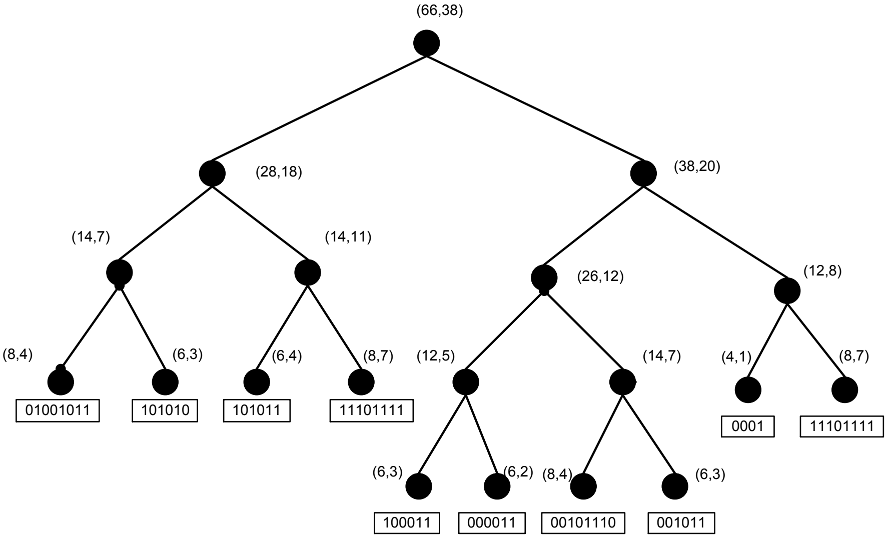


Fig. 5. The tree structure

A superblock is composed of $f(n)$ consecutive blocks ($f(n)$ is a parameter judiciously chosen for guaranteeing acceptable time and space bounds) with the size of them being equal to $(1/2)logn$. The superblocks are all full, with the exception of some partial leafs; the authors impose the invariant that every $f(n)$ full leaves there exists a partial leaf, and implement each leaf as circular arrays except the partial leafs that are implemented as simple arrays. It is shown that the whole scheme needs $O(n/f(n))=o(n)$ extra bits and permits the handling of the various operation in $O(logn)$ time. In particular the space is $n+O(n/f(n))$ bits and the update/query time is at most $O(logn + f(n))$, hence by choosing $f(n)=(logn)$ we get a total construction of $n+O(n/logn)$ bits and $O(logn)$ worst case time for all query and update operations.

Then the authors propose two approaches for making the scheme entropy bounded either by applying gap encoding or by applying block identifier encoding, in order to compress the original sequence. Note that in both cases the dynamic structure follows the same design principles and the time bounds remain $O(logn)$ worst case with the difference that in the second case the $f(n)$ function should be $(logn)^{1/2}$. Assuming that w designates the machine's word size, then it is proved that: (i) using gap encoding, the attained space is $n(1+o(1))H_0(S)+O(l+n^{1/2}polylog(n)+w)$, where l denotes the number of 1's, and (ii) using block identifier encoding, the attained space is $n(1+o(1))H_0(S)+O(nloglogn/logn + w)$.

The construction is then extended to small alphabets of size σ (with the limitation that $\sigma=o((logn)^{1/2})$ by using the same techniques (superblocks contain $2f(n)$ blocks, while blocks are $(1/2)logn$ bits long that is $(1/2)log_\sigma n$ symbols long), and by employing a similar representation scheme as that used in [45] (that is exploiting the symbol decomposition and encoding a block by its position in its specific class). By carefully handling the various technical details arising from this new representation and by using again (as in the block identifier scheme mentioned before) $f(n)=(\log n)^{1/2}$ the attained complexity is $nH_0(S)+o(n\log\sigma)+O(w)$ bits of space, $O(\log n)$ worst case time for rank and select, and $O(\sigma\log n)$ worst case time for insert/delete. The careful reader would notice here the difference between query and update time, a difference that has focused the attention of several authors for possible improvement. Finally, the construction (as [45]), employs a multilevel wavelet tree in order to extend the solution to larger alphabets. In particular a *q*-ary wavelet tree is employed with $q=o(\log n)$ and the resulted construction needs $nH_0(S)+o(n\log\sigma)+O(w)$ bits of space, handles query operations in $O(\log n\log_q\sigma)$ worst-case time, and update operations in $O(q\log n\log_q\sigma)$ worst case time; for $q=2$ the worst case query and update time are asymptotically the same: $O(\log n\log\sigma)$.

The above structures achieve entropy bounds in the space complexity, but are not so fast; this was tried to be handled in subsequent efforts that removed the entropy from the space complexity, in order to become faster in the time complexity. In particular in [70], a construction was provided without using wavelet trees that used $n\log\sigma+o(n\log\sigma)$ bits in the space complexity, providing $O((1/\delta)\log\log n)$ time for rank and select queries, and $O((1/\delta)n^{\delta})$ time for updates.

In another attempt, in [84] an improvement to the [86] data structure was presented in the context of a general dynamic rank and select framework that was later explored in other works too. In particular the authors presented initially a solution for a small alphabet with size σ less than *logn*, and then they extended this solution for an alphabet of arbitrary size. Their small alphabet solution follows the [86] solution with the difference that it separated their construction into two schemes a counting and a storing scheme, thus having the opportunity to optimize both of them simultaneously; the proposed construction supports rank and select queries in $O(\log n)$ worst case time, and update operations in $O(\log n)$ amortized time, moreover the space required is $n\log\sigma+o(n\log\sigma)$ bits. In order to handle alphabets of arbitrary size the authors employ once more multi-ary wavelet trees, for a carefully chosen branching factor. In particular for an alphabet of size σ they regard each symbol of $\log\sigma$ bits as $\log\sigma/\log\log n$ digits from a *logn*-size alphabet. Hence the built *h*-ary wavelet tree is an *logn*-ary tree. As usual the sequence vectors at each node are represented by the dynamic rank and select structures employed previously for the small dictionaries, and it is proved that the provided structure can handle sequences of alphabets of arbitrary size using $n\log\sigma+o(n\log\sigma)$ bits, supporting updates in $O(\log n(1+\log\sigma/\log\log n))$ amortized time, and needing $O(\log n(1+\log\sigma/\log\log n))$ worst case rank and select query time.

As it can be noted the above two solutions [84] and [86] are complementary to each other, that is while [84] is fast and uncompressed, the solution in [86] is slower but compressed. In [59] an effort was made to provide a solution that could combine the best features of these two constructions. Besides combining ideas, the authors in [59] imported also new ideas such as embedding a substructure for handling dynamic partial sums and novel deamortization approaches. The proposed construction initially provides a solution for alphabets of small size *($\sigma=o(logn/loglogn)$)* that needs $nH_0(S)+O(nlog\sigma/ (logn)^{1/2})$ bits of space, and supports all operations in $O(logn)$ worst case time. This solution is then extended for alphabets of larger size by using a generalized *h*-ary wavelet tree where $h=\Theta(logn)$, and thus obtaining a solution with $nH_0(S)+O(nlog\sigma/(logn)^{1/2})$ bits of space and supporting all operations in $O((1+log\sigma/loglogn) logn)$ worst case time.

In [85] an alternative to the [59] structure is presented that is a compressed version of their construction in [84]. In particular while [59] uses a block identifier encoding to compress and a theoretical counting argument to guarantee worst case time, the authors in [85] employ gap encoding in order to achieve compression plus a simple counting structure to have amortized updated time. The proposed solution needs $nH_0(S)+o(nlog\sigma)+O(n)$ bits, the rank/select queries take worst-case time $O(logn)$ and the update operations are handled in amortized $O(logn)$ time for a *logn*-sized alphabet. This can be extended to arbitrary alphabets by using *logn*-ary wavelet trees that need the same space, rank/select queries are handled in $O((1+log\sigma/loglogn)logn)$ time, while update operations are handled in $O((1+log\sigma/loglogn)logn)$ amortized time.

Finally in [97], [71] an improvement to the above time complexities was attained.

In particular in [97] a data structure for manipulating efficiently bit vector operations the *range min max* tree was presented, in the context of dealing effectively with various operations on static and dynamic succinct trees. By using this structure and exploiting techniques presented in [21], [45], [86], the authors were able to show that: (i) a sequence *S* of *n* bits can be maintained in $nH_0(S)+O(nloglogn/logn)$ bits such that query and update operations can be handled in $O(logn/loglogn)$ time, (ii) a sequence *S* of *n* symbols of an alphabet $\sigma=O(log^{1-\varepsilon}n/loglogn)$ can be handled in $nH_0(S)+O(n\sigma loglogn/logn)$ bits of space, so that all the operations can be handled in $O(logn/loglogn)$ time. The result for alphabets of small size, can be extended by using a multi-ary wavelet tree with branching factor $O(1+log\sigma/((1-\varepsilon)loglogn)$. The resultant structure needs space $nH_0(S)+O(nlog\sigma/log^{\varepsilon}n + \sigma log^{\varepsilon}n)$ bits and supports the query and update operations in $O(logn/loglogn(1+log\sigma/loglogn))$ time.

On the other hand in [71], a construction was proposed that follows (and improves) the construction of [59], and employs generalized wavelet trees in order to support rank, select, insert and delete in $O(logn/loglogn(log\sigma/loglogn+1))$ time using $nH_0(S)+o(n)log\sigma+O(w)$ bits. It should be noted that the construction of [71] for small alphabets (of size σ=O(polylog(n)) support all the operations in O(logn/loglogn) time.

4. Applications in Compression

4.1 In [15] a novel application of wavelet trees, in order to effectively compress natural language text is proposed. In particular it is shown that wavelet trees can be used as a means of reorganizing natural text that has been word-compressed in order to guarantee self-synchronization, even for compression algorithms that are not self-synchronized.

Table 1. codewords

Symbol	Codeword
you	b_4
can	b_3
compute	b_5b_0
the	b_5b_1
complexity	b_5b_2
in	$b_5b_3b_0$
a	$b_5b_3b_1$
data	$b_5b_3b_2$
structure	$b_5b_3b_3$

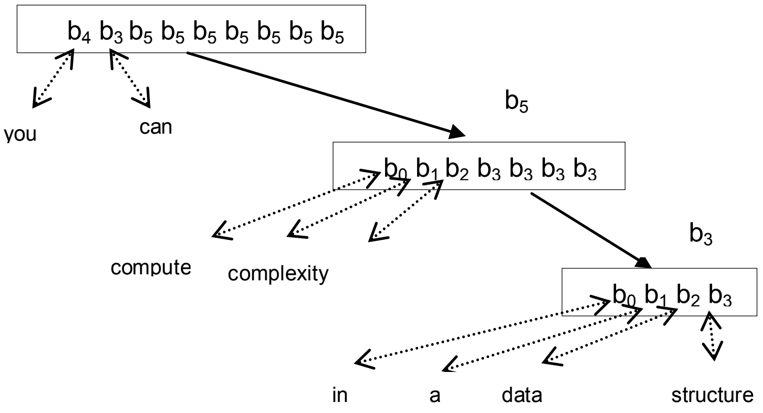


Fig. 6. The wavelet tree for reorganizing text P

Note here that self-synchronization for compressed text is a property that permits fast search and random access, while its lack means that locating a word in the compressed text needs to sequentially traverse it from the beginning. For example Huffman and Restricted Prefix Byte Codes are not self-synchronized while Tagged Huffman, End-Tagged Dense Code and (s,c)-Dense Code are; however their self-synchronization comes with a loss in efficiency.

The technique proposed by the authors of the paper can be applied to any word-based, byte-oriented semistatic statistical prefix-free compression algorithm, and it is shown that with this reorganization the compressed text is synchronized even for codes that by nature are not self synchronized and hence all the required advantages (good compression, fast search and random access ability) can be gained simultaneously.

In particular the main idea is to firstly compress the text and then reorganize the different bytes of each codeword laying them out as they appear in the different nodes of a tree structure resembling closely that of a wavelet tree. In this reorganization the first byte of each codeword appears in the root of the wavelet tree, in the same order as it appears in the text, while the root has so many children as are the different first bytes that a codeword can have; inductively at the *i*-th level, at node *v*, will be stored the *i*-th bytes of all codewords that start with the byte sequence from the root to *v*. It is clear that the produced tree will have depth equal to the maximum length that a codeword can have.

For example let us assume that the text is P="you can compute the complexity in a data structure" and assume that the codewords are as that provided in table 1. In figure 6 it is depicted how the respective wavelet tree is constructed and the bytes of the words of the text reorganized.

In [15] it is proved that by using little extra space, the compressed text can be indexed so that the resultant structure can compete block-addressing inverted indices (the natural choice for indexing natural language texts) and appear as a viable alternative to them; experiments are performed by the authors of the paper to validate such claims.

4.2. In [91], [108] the problem of compressing and indexing highly repetitive sequence collections is considered. Such collections appear in various applications of practice, such as version control systems, and in computational molecular biology for the storage of biological data. In such highly repetitive collections the classical self indexes do not work effectively since the notion of *k*-th order entropy around which these indices are built, fails to capture the notion of repetitiveness. In order to face this fact the authors present a variant of the wavelet tree data structure that embeds in it run length encoding; the compression method of choice when someone compresses repetitive collections. The result is the so called run length encoded wavelet tree and is described as follows: let *R* be the number of runs (that is sequences of identical symbols) in the sequence of *n* characters to be stored (either the initial text or its Burrows-Wheeler transformation). Let B^{all} be the concatenation of all the bit vectors, in a level of the wavelet tree. The bit vector B^{all} is encoded into two separate vectors B^{1} and B^{rl}, with B^{1} storing the starting positions of 1-bit runs, and B^{rl} encoding the run lengths of these runs in unary coding. The value $rank_1(B^{all}, i)$ can be computed, in this modified wavelet tree, by noting that the required answer is the number of 1-bits in $[1, j-1]$ and $[j, i]$, where *j* is the first place in the 1-bit run that precedes *i*; these quantities can be computed by appropriate rank and select queries to B^{1} and B^{rl} (see [91], [108]).

By representing succinctly, using the structures of [69] these bit vectors, then the wavelet tree takes a total of $R\log\sigma\log(2n/R)(1+o(1))+ O(R\log\sigma\log\log(2n/R))+O(\sigma\log n)$ bits of space, while the various queries take time *$O(t_{BSD}\log\sigma)$,* where t_{BSD} is the query time in [69] (u is the length of the handled bit vector and b is its number of 1's) :

$$t_{BSD}(b,u) = O(\min(\sqrt{\frac{\log b}{\log\log b}}\frac{\log\log u \log\log b}{\log\log\log u}, \log\log b + \frac{\log b}{\log\log u}))$$

Moreover the structure, similarly to the various structures mentioned previously for dynamic rank and select, can be made dynamic, by using various available dynamic bit vector representations for B^1 and B^{all}.

4.3 In [10] various techniques are examined to compress a permutation π over n integers, taking advantage of ordered subsequences in π, while supporting the application of $\pi(i)$ and the application of its inverse $\pi^{-1}(i)$ efficiently. The problem is interesting since the techniques presented can be used for various applications and its byproducts improve previous results; it should be remembered that the main idea behind the wavelet tree construction, the range search structure in [20] emerged as an application of mergesort, that can be defined as a series of element permutations. Before proceeding we should note that a *run* in a permutation is defined as the maximal range of consecutive positions that do not contain two consecutive elements in the wrong order. It is shown in [10] that there exists an encoding scheme with space complexity $n(2+H(Runs))(1+o(1))+O(\rho\log n))$ bits that can encode a permutation over the n first integer numbers, that is covered by ρ runs of length $Runs=<n_1, n_2,..,n_\rho>$; here $H(Runs) = \sum \frac{n_i}{n}\log\frac{n}{n_i}$ is the entropy of the runs. The construction supports $\pi(i)$ and $\pi^{-1}(i)$ in $O(1+\log\rho)$ time for any value i in $\{1,\ldots,n\}$; if i is chosen uniformly at random then the average time is $O(1+H(Runs))$.

The specific construction uses a wavelet tree that is built as follows: proceed from the root, and handle recursively each child; when at leaves do nothing and when coming from two siblings of a node, merge their permutations in order to produce the permutation at the father and append a 1 bit, when an element is taken from the right child and a 0 bit when an element is taken from the left child. When the construction is finished the permutation has been sorted in $O(n + \rho\log\rho)$ time, plus the total number of bits appended to all bitmaps, and if the wavelet tree is Hu-Tucker shaped then the total number of bits is at most $n(2+H(Runs))$. Using the wavelet tree and the attached bit maps, $\pi()$ and $\pi^{-1}()$ can be computed in time $O(1+\log\rho)$. The construction was improved in [6] where they used their rank and select data structure, and solved the problem in $2nH(Runs)+o(n)(H(Runs)+1)$ bits supporting the queries $\pi()$ and $\pi^{-1}()$ in $O(1+\log\log\rho)$ time.

4.4. Concerning other applications of the wavelet tree in data compression, in [89] it is shown how to use the wavelet tree in order to index efficiently images. Two such techniques are presented one that is based on transforming the problem to that of efficiently handling one dimensional suffix arrays and the other that is based on 2-dimensional suffix arrays and an application of well-known techniques to index images based on so called Lsuffixes. Experimental results in that paper validate the usefulness of the proposed approaches. In particular it is verified experimentally that wavelet trees can be seen as an improvement over the classical bit-plane encoding, and be used to obtain both lossless and lossy compression.

Use of the wavelet tree structure also appears in [4], [104] where it is applied as a range structure, for their compressed self-index on texts compressed with the LZ78 data compression, and in [82] as a component of various substructures, in their implementation of a self-index handling texts compressed with the LZ77 data compression algorithm. The last application is very interesting since LZ77 can benefit from the existence of frequent repetitions in a text, and as discussed previously these repetitions, appear in various applications. It is mentioned that the proposed self-index is smaller (up to one half) than the best current self-index for repetitive collections, and in many cases it is also faster.

Finally, in [30] it is depicted how the use of wavelet trees can support efficient representation and navigational operations in a compressed representation of the Web graph (for other approaches see [29], [17]). In particular the paper represents the adjacency list of each node of the graph as a sequence that is compressed using the re-pair [83] compression technique, and then it is handled by two alternative representations for rank and select, the structure in [57] and a wavelet tree, in order to support the operations of retrieving direct and reverse neighbors. Experiments with these two approaches and with various alternatives, depict that the wavelet tree representation though slower (sometimes by an order of magnitude) from its alternatives achieves the smaller space reported in the scientific literature (in comparison with similar techniques to represent the web graph), and can be considered to be the practically better choice.

5. Applications in Information Handling

5.1 In [18] the relationship between wavelet trees and range searching data structures, that in plain words makes the wavelet tree equivalent to an orthogonal range searching data structure suitable for handling points with discrete coordinates, was pushed a bit further by placing them as a viable alternative to spatial data structures that use and store minimum bounding rectangles in order to handle various spatial queries as they appear in geographical information system. In that paper experiments were performed comparing the proposed structure with variants of the R-tree family (the R*-tree and the STR R-tree) both in space and in time performance in synthetic and real datasets. Concerning the space complexity the structure needs less space than the R*-tree in both datasets, while it needs less space than the STR R-tree in real datasets, and a comparable space in synthetic. Concerning the time complexity if the dataset is uniform then the wavelet tree outperforms both variants of the R-tree while in the case that it is not it outperforms them for very selective queries.

In particular consider a set of *N* objects that are represented by their Minimum Bounding Rectangles (MBR) as defined by their lower left and their upper right corners (each corner is a pair of coordinates in x- and y- axis).

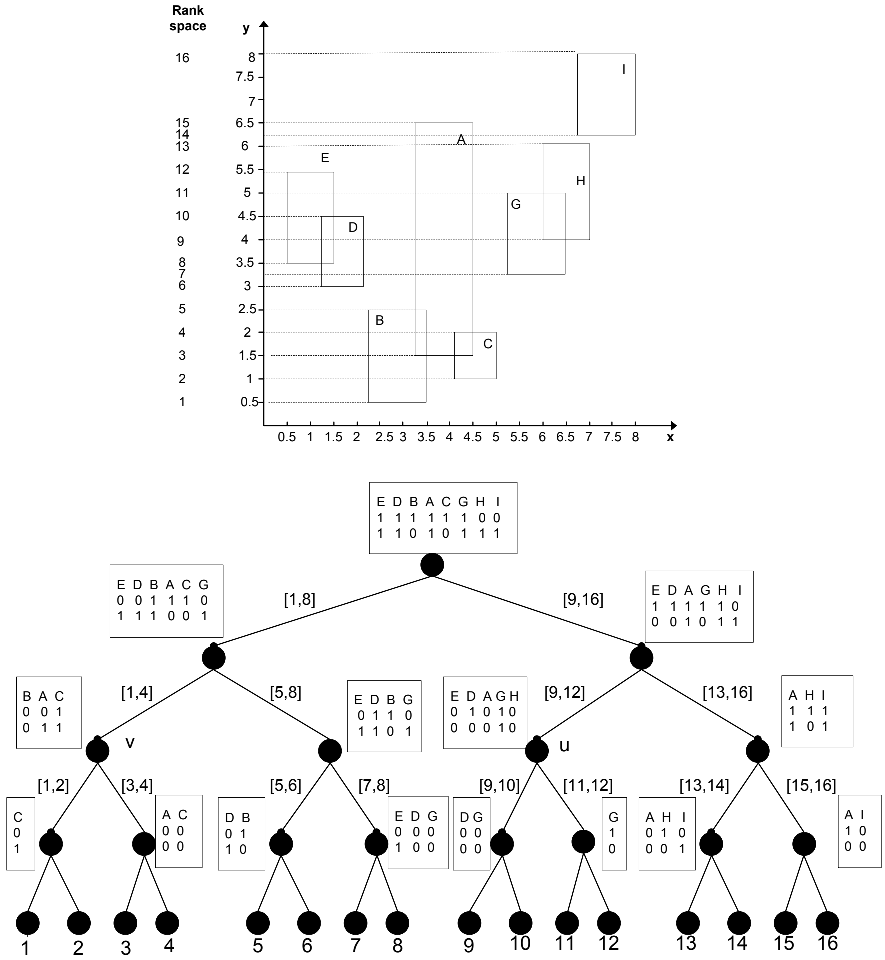


Fig. 7. The wavelet tree for representing a set of maximal Minimum Bounding Rectangles {A, B, C, D, E, F, G}

The proposed structure handles intersection queries by initially splitting the set of MBRs into *k maximal* sets and then building a wavelet tree separately for each maximal set. A set of MBRs is termed *maximal* if it is not possible for a projection over the x-axis of a rectangle in this set to contain the projection of another rectangle in the same set. Given a set of *N* rectangles the optimal partitioning to maximal sets can be performed in *O(NlogN)* time, by a procedure that is related to the longest increasing subsequence problem, and to the problem of decomposing permutations.

So let us suppose that the set at hand is maximal, this set can be represented in a square grid of *2N* rows and columns, with each corner point in each row and column. This grid can be represented simply in rank space, by storing the real coordinates of the MBRs in suitable structures, in order to translate a given query rectangle to the discrete coordinates in the rank space. Because the handled set is maximal, the order in the x-coordinate of the left and right sides is the same, hence two sorted arrays for the *x*-coordinates and one sorted array for the *y*-coordinates, plus arrays storing the identifiers of the MBRs in the same order suffice for the translation from the geographic space to rank space.

The constructed wavelet tree is built according to the *y*-coordinates, and each node of the tree stores the respective MBRs according to the *x*-order, with the novelty that instead of one, two bit vectors are used. The first bit vector projects onto the left child and has the value 1, if the respective MBR is processed in the left child, and the second bit vector projects onto the right child and has the value 1, if the respective MBR is processed in the right child.

An MBR is processed in a node if it intersects the *y*-range corresponding to the node, and hence an MBR can be processed in both the left and the right child of a node. In order to avoid the quadratic space explosion due to the fact, that an MBR could be stored in a linear number of nodes at a level, a modification in the structure is presented that resembles that of a segment tree, by demanding that when an MBR covers the whole range of a node, then it is not stored in any of its descendants.

Figure 7 depicts the construction for a set of maximal MBRs. In the upper part of the figure the set of rectangles is depicted plus the corresponding rank space for the y-coordinates, while in the constructed wavelet tree the bit vector for the left child of a node is depicted above the bit vector for the right child, In node *u* for example rectangle *D* intersects only the left and not the right child's range, hence it stores 1 in the first bit vector and 0 in the second, while concerning the space saving heuristic, in node *u* rectangles A, E and in node *v* rectangle *B*, contain the whole range of the respective nodes, hence the bits for both bit vectors are equal to 0.

The above structure and assuming that it stores a maximal set of MBRs can answer range queries by appropriately descending paths of the tree according to the provided query range, thus giving a time complexity of *O(logN)* for each path traversed. In particular the authors provide a recursive algorithm that projects the initial query range at a node to suitable query ranges in the first and right child by applying rank operations in the two bitmaps.

In the general case and since we have at our disposal *k* maximal sets, we need to query *k* wavelet trees, and hence the total time complexity of the tree paths traversed is bounded by *O(klogN).*

5.2. In [106] the wavelet tree is proposed as an index in order to implement bidirectional search in a string. The motivation for this application comes from the area of Bioinformatics and in particular genome encoding where after finding suitable matches, both ends of the string need to be extended by exploiting the complementarily in the bases. This kind of search can be implemented by using bidirectional search. Indices performing this kind of search have been proposed in the bibliography in the form of affix trees and affix arrays [109], [110]. The affix tree of a string is comprised by its suffix tree and the suffix tree of its reverse, while the affix array combines the suffix arrays of the string and the suffix array of its reverse. Generally, when answering queries, it is difficult to implement the interplay of these structures while the space complexity is usually large. In order to face these shortcomings the authors in [106] present the bidirectional wavelet index of a string which consists of two wavelet trees, the wavelet tree of the Burrows-Wheeler transformation of the string, and the wavelet tree of the Burrows-Wheeler transformation of its inverse. The difficult part in using these structures is how to synchronize the search on both indices, but the authors in [106] depict, that by a carefully designed search procedure it is possible to perform bidirectional search. Performed experiments depict that the proposed index decreases the space requirement by a large constant factor of 21 (in comparison to affix arrays) making it possible to apply their algorithm in very large strings.

5.3. In [5] the wavelet tree is extended in order to support binary relations [7]. Binary relations are an important abstraction that is inherent in many combinatorial objects such as trees, graphs, inverted lists, and web graphs. In particular a binary relation *B* between a set of objects with identifiers in *[1,n]* and a set of labels with identifiers in *[1,σ]* can be considered as a set of *t* pairs from a total universe of *nσ* possible pairs; it can also be represented as a matrix with *σ* rows (the labels) and *n* columns (the objects). The notion of entropy can be extended to handle these combinatorial objects by defining it as:

$$H(B) = \log\binom{n\sigma}{t} = t\log\frac{n\sigma}{t} + O(t).$$

In the specific paper besides listing operations of potential interest and giving reductions between operators, two data structures are presented. The first data structure uses the reduction of binary relation operators to string operators by representing a binary relation with a bitmap *B[1,n+t]* concatenating the cardinalities of the columns of the relation in unary, and with a string *S[1,t]* over the alphabet *[1,σ]* containing the labels of the pairs in column major order. Using this representation it is possible to get an efficient implementation of a binary relation that supports the various operations in *O(logσ)* time, and needs *tlogσ+o(t)logσ+O(min(n,t)log((n+t)/min(n,t))* bits of space.

The second structure is the so called *Binary Relation Wavelet Tree*. This tree is a wavelet tree like construction with the leaves corresponding to individual rows of the relation, and with each node containing two bitmaps per level. The first bitmap corresponds to objects in the left child, and has its *i*-th bit equal to 1, if there exists an object in the left child with label having identifier *i,* while the second bitmap corresponds to objects in the right child and has its *i*-th bit equal to 1, if there exists an object in the right child with label having identifier *i*. Note the similarity of the construction, with that in the extension of the wavelet tree in order to handle spatial objects; here also it is possible for an object *x* to propagate both left and right and this means that the sizes of the bitmaps at a level may add up to more than *n* bits. The authors prove that the specific construction use $log(1+\sqrt{2})tH(B)+O(tH(B))+O(t+n+\sigma)$ bits of space, and can support the various operations in $O(log\sigma)$ time.

5.4 In [53] (see also [51]) it is described how to use wavelet trees in order to support efficient range quantile queries in a list *S* of *n* numbers. A range quantile query takes a rank and the endpoints of a sublist and returns the number with that rank in that sublist. More analytically the wavelet tree is built for the list of numbers together with the appropriate rank and select data structures at the bit vectors of its nodes. Suppose that we are given the endpoints *l* and *r* of a sublist, and a rank *k*, and we want to report the element with rank *k* in the sublist between *l* and *r*.

Starting from the root, and accessing its binary vector *B*, we apply two rank queries with arguments *l-1* and *r*, in order to find the number of 0s and 1s in $B[1,l\text{-}1]$ and $B[l,r]$. If there are more than *k* zeroes in $B[l,r]$ then we have to move to the left subtree, and we set *l* to be one more than the number of 0s in $B[1,l\text{-}1]$, *r* to be the number of 0's in $B[1,r]$ and we recurse on the left subtree, otherwise we have to move to the right subtree of *T*, hence the new *k* is the old *k* minus the number of 0s in $B[l,r]$, the new *l* is one more than the number of 1's in $B[1,l\text{-}1]$, we set *r* to be equal to the number of 1s in $B[l,r]$ and we recurse to the right subtree; when a leaf is reached its label is returned. If σ is the total number of distinct elements (note that this number is smaller than both *n*, and the size of the universe from which the elements take value) then the tree has height $O(log\sigma)$, and since accessing each node's binary vector costs O(1), it follows that the time complexity is $O(log\sigma)$. Concerning space, the cost is equal to the space complexity of the wavelet tree, hence depending on the constant time dictionaries that are used for the bit vectors, it can have different values from $O(nlog\sigma)$ to $nH_0(S)+O(nloglogn/log_\sigma n)$ bits.

As a working example consider figure 8, depicting a wavelet tree for the sequence *S=15,14,1,5,6,4,11,12,13,8,9,7,16,2,3,10* and assume that we would like to locate the 4-th element in *S[3..11]* . We start from the root *r*, and compute $rank_0(B_r, 2)=0$, και $rank_0(B_r, 11)=5$, hence the number of 0s and 1s in *S[3,11]* is 5 and 4 respectively; since the number of 0s is larger than 4, we go to *v*, looking for the 4^{th} element in *[1,5].* There we have $rank_0(B_v,0)=0$ and $rank_0(B_v,5)=2$, hence the number of 0s and 1s in [1,5] is 2 and 3 respectively; since the number of 0s is less than 2, then we have to move to *w*, looking for the 4-2=2^{th} element in [1,3]. In *w* we have $rank_0(B_w,0)=0$ and $rank_0(B_w,3)=2$, hence the number of 0s and 1s in [1,3] is 2 and 1 respectively; since the number of 0's is equal to 2 we move to the left child *u* of *w* looking for the 2^{th} element in [1,2]. We have $rank_0(B_u,0)=0$ and $rank_0(B_u,2)=1$ hence the number of 0,s and 1s in [1,2] is 1 and 1 respectively; since the number of 0's is less than 2 we move to the right and the answer is affirmatively the number 6.

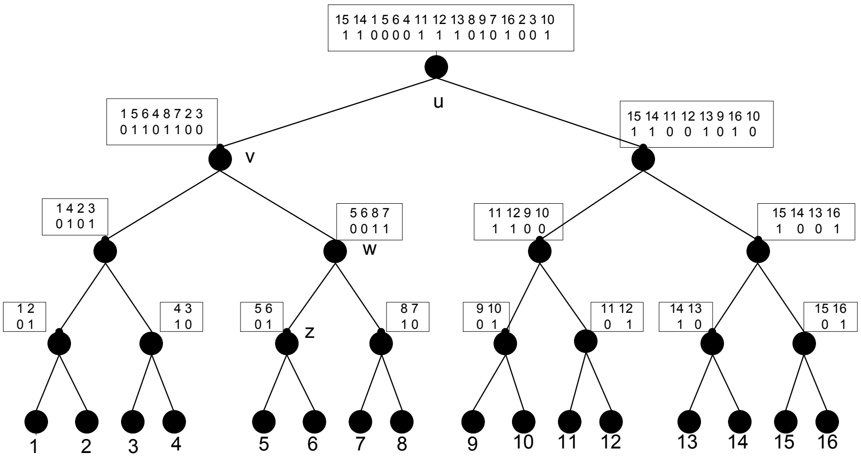


Fig. 8. A wavelet tree for the sequence S=15,14,1,5,6,4,11,12,13,8,9,7,16,2,3,10

The aforementioned structure can be generalized to any constant number of dimensions, by using the same base tree structure on the distinct elements, and by employing a multidimensional range counting structure in the internal nodes. If the structure is the one in [80] then it is proved in [53] (in the full version of the paper appearing as technical report) that for any positive constants *d* and *t* there exists a structure of size $O(n^{1+t}log\sigma)$ bits, that can answer *d*-dimensional range quantile queries in $O(log\sigma)$ time.

As another application of the solution to the 1-dimensional range quantile query, the above algorithm can be extended in order to enumerate the *f* distinct items in a given sublist in $O(flog\sigma)$ time. The reporting algorithm, can find the first element in a given range, in $O(log\sigma)$ time with a range quantile query, and then exploiting the fact that in $O(log\sigma)$ time a wavelet tree can compute the number of occurrences of a symbol in the prefix of a sequence, it can transform the search of successive elements by a suitable range quantile query costing $O(log\sigma)$ time. This operation is important since it can be used in the document listing problem [93].

In the document listing problem we are given a set of *n* documents of total size *N* characters from an alphabet Σ of size σ (without word separation) and we want to locate the documents that contain a pattern. In [93] a solution was provided using a suffix tree and a document array, being accompanied by a compressed array, using space $O(Nlogn)$ bits and $O(m+t)$ query time, where *m* is the size of the seeked pattern and *t* denotes the output size. The space was reduced in [111] by employing a compressed suffix array and a wavelet tree to represent the document array plus a range minimum query structure and the query time was worsened to $O(mlog\sigma+tlogn)$, where *t* denotes the size of the output. In particular the space complexity was improved to $|CSA|+2N+Nlogn(1+o(1))$ bits, where $|CSA|$ is the size of a suitable compressed array and is less than $Nlog\sigma(1+o(1))$ bits. The aforementioned extention of the range quantile query permits the use of only the wavelet tree structure without the accompanying machinery (the range minimum data structure) thus reducing even further the space of the solution in [111] (by $N+o(N)$ bits), while keeping the same time complexity. It should also be noted that in [111] a remark concerning autocompletion search [11], was also made. In particular it was noted that the structure presented in [11] can also be used for the document listing problem with the same space but worse time complexity. It is worth to be mentioned that the autocompletion search structure provided in [11] is similar to the wavelet tree construction, though different in its functionality and use of accompanying bit vectors.

A problem closely related to the document listing problem is the top-*k* query problem [33], [76], [81], in which we want to locate the top *k* documents where a pattern appears more often (top-*k* retrieval). Wavelet trees fit nicely in this extension of the problem since they can be used to compute pattern frequencies and report rankings. In particular in [33] the range quantile query approach in [53] was extended with two heuristics, that permitted the effective calculation of frequencies and top-*k* results reporting; the greedy traversal heuristic and the quantile probing heuristic. Both were tested experimentally and the greedy heuristic stood satisfactory as a competitor to inverted files, for natural language query processing. The above two heuristics do not come with bounded time estimations; this does not hold for the solution in [76], where it is shown that $O(m+k\log^{3+\varepsilon} N)$ query time can be attained using $|CSA|+o(N)+n\log(N/n)$ bits; as noted in [96] wavelet trees can be used in this solution as a substructure to count frequencies.

Finally, in [96] a further development was achieved by reducing the space complexity of the wavelet tree involved in the above constructions (see also for a similar attempt in [52]). This space compaction was achieved by applying grammar compression (re-pair compression [83]) to the bitmap dictionaries and then employing these compressed representation for storing the subsequences at each node of the wavelet tree. The proposed structure is shown experimentally that it achieves great space savings at the cost of a somehow inferior query time; moreover combinations of this structure with the approaches in [33], [76] are tested.

Concluding in [81] wavelet trees were engaged providing an asymptotically optimal solution for the top-*k* color problem. That is, consider an array with N elements each with a color and a priority, then there exists an $O(N\log c)$ bits data structure that reports k colors with the highest priority in a query interval in $O(k)$ time; here c denote the number of different colors. This solution can be exploited for the *top-k* document retrieval problem leading to an algorithm with $O(N\log N)$ bits and $O(m+k)$ query time.

Before closing, it should be noted that a further extension to the above problems was presented in [74] where it was shown how wavelet trees can be effectively applied in conjunction with the multiway search paradigm introduced in [32] in order to solve the document listing problem in the case when the query consists of $m>1$ patterns $\{P_1, P_2, .., P_m\}$. The proposed solution needs linear space and can answer these queries in time $O(p)+\tilde{O}(t^{1/m} n^{1-1/m})$, where p is the total length of the patterns and t is the size of the output. Moreover in the case of two patterns the achieved bound is

$O(|P_1|+|P_2|+\sqrt{nt}\log^2 n)$. It is also shown how other problems dealed with in [93] can be efficiently handled. A basic ingredient of the aforementioned construction is a variant of the wavelet tree, the so called *weight balanced wavelet tree*, that is interesting ın its own right, since it is different from other wavelet tree implementations. In this variation it is explicitly settled during construction that the number of 0's and 1's at the bit vector at each node is almost equal and hence the *total* number of symbols (and not the number of *distinct* symbols as usual) stored at each child of a node is almost equal. In particular the construction is performed top down, and at each node the symbols are processed in decreasing frequency order, and a symbol is going left or right depending on where the total number of symbols is bigger (right or left respectively). The weight balanced wavelet tree has the property that for every node at depth *d*, the corresponding bit vector has size at most $4n/2^d$, moreover its space consumption is bounded by the zero entropy of the stored sequence, without even needed to compress the bit vectors.

5.5 Using similar techniques on the wavelet tree [51] as in the range quantile query, it is possible to solve the *range next value* problem and the *range intersect* problem.

Let us first consider the range next value query on a wavelet tree *T*. In this problem we are given two endpoints *l, r* of a query interval plus a value *x*, and we want to compute the smallest value greater or equal to *x*. Initially we start from the root, if *x* is at a leaf at the right subtree of the root, we move to the right child by setting as new left value $rank_1(B_{root}, l-1)+1$, and as new right value $rank_1(B_{root},r)$; otherwise the search continues with the left child by setting as new left value $rank_0(B_{root},l-1)+1$, and as new right value $rank_0(B_{root},r)$. If at some point the interval becomes empty the recursion stops returning no value. If the recursion returns from the right child of the root, the answer is that no value exists, otherwise it could be possible that a number exists in the T_r, hence a final attempt is made to T_r, but in this time with a query that demands the minimum value in the sublist between *l* and *r*. This query is a special case of the range quantile query and hence can be handled as described previously. The whole approach overall needs to follow a path in the wavelet tree, performing constant time queries in each visited node, and thus the search time is $O(log\sigma)$, with *σ* being the number of distinct values in the list. The space complexity is $nlog\sigma+O(n)$ bits.

A similar problem, termed *range successor* problem was handled in [113], [114]. In this problem we are given a set of *n* points in a grid of distinct coordinates, and we would like to preprocess them so that for any given orthogonal range, we can report the point with the smallest *y*-coordinate. One of the solutions presented in [113], [114] takes $O(logn/loglogn)$ time to answer the query, needs $O(n)$ words space and makes heavy use of wavelet trees, improving a previous $O(logn)$ query time construction presented in [90]. The construction is basically a multiway wavelet tree with branching factor $(logn)^{1/2}$. In each node a substructure is attached that is also proposed by the authors that can handle $m<=n$ integers with range $r=O(logn/loglogn)$, in $O(mlogr)$ bits, so that together with an $o(n)$ bit table can handle range successor queries in $O(1)$ time. By effectively embedding the structure in the multiway wavelet tree, they show how to achieve the aforementioned time and space bounds.

In the *range intersect* problem, given k ranges in a list, we want to find the distinct common values in these ranges. In order to handle this problem we will first consider the case that we have two ranges; the case of multiple ranges can be handled easily by reducing it to multiple applications of the problem of intersecting two lists or it can be handled by an immediate extension of the basic procedure. The problem is extremely interesting since it has various applications and resembles the problem of intersecting the posting lists in an inverted file. In this kind of problems the inherent complexity is captured through a measure termed alternation α, which can be defined as the number of switches from one list to the other in the sorted union of the two ranges.

Hence assume that the two ranges are $[l_1,r_1]$ and $[l_2, r_2]$. The procedure starts from the root of the tree and tests if either of the ranges is empty, if they are the procedure stops, otherwise, by using the bit vectors at the root the algorithm descends to the left and to the right child, and continues recursively to the nodes where both ranges are non empty. When a leaf is reached then the corresponding element is in the intersection and we report its respective number of occurrences in the first and in the second list. The same procedure can be applied for multiple lists, and descending a path is stopped when one of the ranges becomes empty.

It is proved in [51] that the specific algorithm can compute the intersection of k ranges in time $O(\alpha k log(\sigma/\alpha))$ where k is the number of ranges, α is the alternation complexity of the problem and σ denotes the number of distinct values in the sequence.

Another algorithm given in [51] for the same problem, uses two variables x_1 and x_2, that scan the two lists simultaneously, and move through calls of the range next value procedure, thus simulating the well known merge procedure of two sorted lists, with the difference that the movement is achieved via calls of the range next value procedure; if at any time x_1=x_2, then an intersection is reported and the procedure continues by setting as x_1 the range next value in the first list of x_2+1. The specific procedure makes a switch each time an alternation is met and therefore the time complexity is $O(\alpha log\sigma)$. This can be improved to $O(\alpha log(\sigma/\alpha))$ by implementing the range next value procedure, so that it remembers the path that was traversed the last time, in a form of fingered search and thus getting a time complexity of $O(\alpha log(\sigma/\alpha))$.

5.6. In [51] and [95] it is shown how to use the wavelet tree in order to implement the search procedure that is needed when processing the posting lists of an inverted file. Inverted files are the method of choice when creating indices for both ranked and boolean retrieval.

In order to have both modes of inverted files operations operated in minimum space, it is needed to efficiently retrieve the documents both by decreasing weights (permits effective handling of queries) and by increasing document number (permits effective compression). In [95] and [51] it is depicted how to use the wavelet trees in order to implement efficiently, in the space required for a single compressed inverted index, the operations of retrieving efficiently the documents either in decreasing weight, or in increasing identifier order.

The construction basically concatenates and sorts all the posting lists of the terms (these should contain the document identifiers where a term appears in descending order of the term frequency of the specific term), thus creating a sequence list of document identifiers, and this sequence is stored in a wavelet tree. In parallel for each term in the collection, it is stored the starting position of its list in the sequence of identifiers, and the term frequencies are stored in differential and run-length compressed form in a separate sequence. Finally the starting positions of the list of each term in the concatenated sequence are stored in separate bitmap preprocessed for rank and select queries.

Let *D* be the total number of documents in the collection, let *L* be the length of the list with all document identifiers and let *N* be the total length in words of the text collection. Then, it is shown that the proposed construction takes space $NH_0(L)+o(NlogD)$ bits plus the cost for storing the various term frequency values in differential and run length compressed form. It is proved that with this construction the various operation of interest in both boolean retrieval and ranked retrieval can be efficiently supported.

Another construction was also employed in [3] for handling the posting lists retrieval problem, that needs $NH_0(L)+o(N)(H_0(L)+1)$ bits of space, such that a conjunctive query of *k* terms can be performed in $O(k\alpha loglog\sigma)$ time. This construction is mainly focused on the conjunctive query problem and represents the set of documents as a concatenated sequence, but this time with alphabet the distinct terms and employs a rank and select structure for arbitrary alphabets in order to handle it. The aforementioned bounds are valid if the fast rank and select structure proposed in [6] is used, however a valid alternative (especially in a practical implementation) is that of Huffman shaped wavelet tree. In general, using the wavelet tree structure one is able to retrieve the frequency information of a term, its positional information, the respective posting list, and answer conjunctive queries. Moreover the authors show how to retrieve effectively snippets enclosing the appearances of a given term. In comparison to the inverted files approach the wavelet tree though needing more time, need less space, and can providing more functionalities.

In [2] the approach of [3] is extended in order to handle the situation where the documents are stored in an *P* x *D* array of search node processors (*P* denotes the number of text partitions, and *D* denotes the replication level inside each partition). It is depicted both theoretically and experimentally that wavelet trees can efficiently work in this setting, while they can be used to dynamically maintain a cache with the most recently used posting lists. In particular it is shown that all processing phases (index decompression, ranking and snippet extraction) can be performed by a simple unifying structure thus permitting the use of less processors for the same performance or putting out it differently permits with the same number of processors better performance. The above is achieved by unifying into a single unit the pair of search nodes and document servers. Moreover besides employing the wavelet tree's ability to generate efficiently snippets the authors exploit its capability to produce posting lists and compute intersections in combination with caching strategies. In this case it is assumed that a term partitioning approach is applied in the replicas, and the employed wavelet trees are used to support various functionalities on a dynamically maintainable cash.

5.7 The wavelet tree has also found various applications in the retrieval of XML documents; see for example [1], [13], [41], [42], [51]. From these papers the most interesting are the constructions provided in [13] and [51].

In [13] the hypothesis is made that the XML tree can be modeled as a tree with document identifiers (that is passages of text) appearing only to the leaves of the tree, and with internal nodes being mapped to structural separators. The structure used for indexing is a parenthesis representation of the tree modeled as a sequence that is obtained through a preorder traversal. To each parenthesis a tag name is mapped and the sequence of tag names is represented using a wavelet tree; moreover for each distinct tag name a parenthesis representation of the nodes of the XML tree that are tagged by it, is stored. The total space for all these structures is $2n\log\tau+O(n)$ bits where n is the number of leaves of the tree (the basic text structures) and τ, is the number of distinct tags in the collection.

In [13] it is proved, that this representation can answer effectively various operations on the XML structure such as: locating the lowest ancestor of a text passage (leaf) where an occurrence appears, and return the range of text passages corresponding to this ancestor, listing text passages restricted to structural unit tagged with a specific name, counting number of occurrences of a query in a retrievable node, and computing intersections on retrievable units.

Finally, an interesting combination of the XML machinery and the wavelet tree literature appears in [51] where a novel structure the XML wavelet tree, is proposed as a self index for XML documents; this structure can handle XPath queries more efficiently than using the uncompressed counterparts, and also appears to be more competitive than inverted indices of the same space consumption. The idea is based in a combination of the (s,c)-DC compressor [16] with the wavelet tree. In particular the (s,c)-DC compressor is initially used in order to compress the XML document, and then the idea of [15] that reorganizes the compressed codewords in a structure resembling a wavelet tree is applied, in order to facilitate immediate access and retrieval of the appropriate codeword. Experiments performed in [51] validate the practical benefits of the proposed approach.

6. Conclusions and open problems

Wavelet trees represent a clear example where a data structure designed for a theoretical construction (Chazelle's work [20]) benefits the practitioners (designers of data compression algorithms) providing them with tools for efficiently handling several practical problems. Besides improving the bounds in the aforementioned uses of the wavelet tree (the majority of them constitutes in these days hot research topics and are the target of active research in the area of algorithms and information retrieval), it would be interesting to find out other applications of the specific technique and explore issues of extending the structure itself. Possible areas of future extensions could employ: self adjusting heuristics, embedding machinery of persistent data structures, compressing in various ways the structure itself (see examples for such approaches in [96], [81]) and cache oblivious extensions.

In particular weight balanced wavelet trees and frequency based wavelet tree have been proposed in [74], [49] for various applications, it would be interesting to see how these approaches combine together, and explore the suitability of splay heuristics if being applied to these structures (the frequency based heuristics of [49] provide such a clue, while it should be noted that splay trees are a natural data structural choice in various compression settings). Moreover compression algorithms of the self-indexed family generally fail to deal with repetitions in the text [91], [108] a phenomenon that could probably be dealt effectively by applying techniques from the persistent machinery [35]. Therefore it would be worthwhile to study the way that persistent techniques, in all of its various manifestations (partial, full, confluent) could be applied in wavelet trees. Moreover in [81] it has been shown (and it is trick that is used main times) how to jump levels of the structure in order to speed up it the descend of its levels; it would be nice to have this as a general technique that could be applied as a general framework in the various of the structure's applications. Finally since the structure itself is easy to be implemented in secondary memory [62], [75], it would be interesting to have a study of its various applications in the more advanced cache oblivious model.

Concluding, it would be also worthwhile to find out if achievements and research progress in the area of geometric search algorithms could possibly embed new tactics and methodologies, in the efficient maintenance of wavelet trees, since the roots of the structure are from this computer science research field. Especially in the last years, in the computer science literature, there are a lot of algorithms and variants of techniques dealing with the range searching problem, and it should be explored how these achievements reflect on the wavelet tree.

Acknowledgements The author would like to thank the two anonymous referees for their valuable suggestions that helped to improve the quality of the paper. Moreover, this research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thales. Investing in knowledge society through the European Social Fund.

References

1. Arroyuelo, D., Claude, F., Maneth, S., Makinnen, V., Navarro, G., Nguyen, K., Siren, J., Valimaki, N.: Fast in-memory XPath search over compressed text and tree indexes. ICDE, 417-428, (2010).
2. Arroyuelo, D., Gil-Costa, V., Gonzalez, S., Marin, M., Oyarzun, M.: Distributed search based on self-indexed compressed text. Information Processing and Management (2011).
3. Arroyuelo, D., Gonzalez, S., and Oyarzun.: Compressed self-indices supporting conjunctive queries on document collections. In SPIRE, LNCS 6393, pp. 43-54 (2010).
4. Arroyuelo, D., Navarro, G.: Space-efficient construction of Lempel-Ziv compressed text indexes. Information and Computation 209, 1070-1102 (2011).
5. Barbay, J., Claude, F., Navarro, G.: Compact Rich Functional Binary Relation Representations. LATIN, pp. 170-183 (2010).
6. Barbay, J., Gagie, T., Navarro, G., Nekrich, Y.: Alphabet partitioning for compressed rank/select and applications. ISAAC (2), 315-326 (2010), see also the version in *CoRR, abs/0911.4981, 2009.*
7. Barbay, J., Golynski, A., Munro, J., Rao, S.: Adaptive searching in succinctly encoded binary relations and tree structured documents. Theoretical Computer Science, 387 (3): 284-297, (2007).
8. Barbay, J., He, M., Munro, J., Rao, S.: Succinct indexes for strings binary relations and multi-labeled trees. ACM/SIAM SODA, 680-689 (2007).
9. Barbay, J., and Kenyon, C., Adaptive intersection and t-threshold problems. In SODA, pages 390-399, 2002
10. Barbay, J., Navarro, G.: Compressed representations of permutations and applications. Symposium on Theoretical Aspects of Computer Science. 111-122 (2009).
11. Bast, H., Mortensen, C.W., Weber, I.: Output sensitive autocompletion search. Information Retrieval. 11:269-286 (2008).
12. Bose, P., He, M., Maheswari, A., Morin, P.: Succinct orthogonal range search structures on a grid with application to text indexing. WADS, 98-109 (2009).
13. Brisaboa, N., Cerdeira-Pena, A., Navarro, G.: A compressed self-indexed representations of XML documents. ECDL, 273-284 (2009).
14. Brisaboa, N., Cillero, Y., Farina, A., Ladra, S., Pedreira, O.: A New Approach for Document Indexing using Wavelet Trees. DEXA Workshops, 69-73 (2007).
15. Brisadoa, N., Farina, A., Ladra, S., Navarro, G.: Reorganizing Compressed Text. ACM SIGIR, 139-146 (2008).
16. Brisaboa, N., Farina, A., Navarro, G., Esteller, M.: (s,c)-Dense Coding: an optimized compression code for natural language text databases. SPIRE, pp. 122-136, (2003).
17. Brisaboa, N., Ladra, S., Navarro, G.: k^2-trees for Compact Web Graph representations. SPIRE, pp. 18-30, (2009).
18. Brisaboa, N., Luaces, M., Navarro, G., Seco, D.: A fun application of compact data structures to indexing geographic data. FUN, 77-88 (2010).
19. Brisaboa, N., Luaces, M., Navarro, G., Siego, D.: A new point access method based on wavelet trees. ER Workshops, 297-306 (2009).
20. Chazelle, B.: A functional approach to data structures and its use in multidimensional searching. SIAM J. Computing, 17(3), 427-462 (1988).
21. Chan H.L., Hon W.K., Lam, T.W., and Sadakane K.: Compressed indexes for dynamic text collections. ACM Transactions on Algorithms, 3(2):article 21, 2007.

22. Chien, Y.F., Hon, W.K.: Geometric Burrows-Wheeler transform: linking range searching and text indexing. In Proceedings of the 2008 IEEE Data Compression Conference (DCC ’08), Snowbird, UT, March (2008).
23. Chien, Y.F., Hon, W.K., Shah, R, Vitter, J.S.: Compressed Text Indexing and Range searching. TR 2006-21, Purdue University, (2006).
24. Chiu, S., Hon, W., Shah, R., Vitter, J.S.: I/O Efficient Compressed Text Indexes: From Theory to Practice. Data Compression Conference, 426-434 (2010).
25. Clark, D.: Compact Pat Trees. Phd thesis, University of Waterloo, (1996).
26. Claude, F., Navarro, G.: Self-indexed text compression using straight-line programs. In: Proc. MFCS, pp. 235-246, LNCS (2009).
27. Claude, F., Farina, A., Martinez-Prieto, M., Navarro, G.: Compressed q-gram indexing for highly repetitive biological sequences. IEEE International Conference on Bioinformatics and Bioengineering, 426-434 (2010).
28. Claude, F., Navarro, G.: Practical Rank/Select Queries over Arbitrary Sequences. SPIRE, 176-187, (2008).
29. Claude, F., Navarro G.: Fast and compact Web graph representations. ACM Transactions on The Web, Vol. 4, No. 4, Article 16, Pub. date: September (2010).
30. Claude, F., Navarro, G.: Extended compact web graph representations. Algorithms and Applications, pp. 77-91 (2010).
31. Claude, F., Nicholson, P., Seco, D.: Space efficient wavelet tree construction. In SPIRE’11 (2011), pp. 185-196.
32. Cohen, H., Porat, E.: Fast set intersection and two-patterns matching. In LATIN, LNCS 6034, pp. 234-242, (2010).
33. Culpepper, S., Navarro, G., Puglisi, S., and Turpin, A.: Top-k ranked document search in general text databases. ESA, pp. 194-205, (2010).
34. Deorowicz, S.: Second step algorithms in the Burrows-Wheeler compression algorithm, Software- Practice and Experience. 32:99-111 (2002).
35. Driscoll, J., Sarnak, N. Sleator, D.D., Tarjan, R.E.: Making Data Structures Persistent. J. Comput. Syst. Sci. 38(1): 86-124 (1989)
36. Farina, A., Ladra, S., Pedreira, O., Places, A.: Rank and select for succinct data structures. Electr. Notes Theor. Comput. Sci. 236: 131-145 (2009).
37. Farzan, A., Gagie, T., Navarro, G.: Entropy bounded representation of point grids. ISAAC: 327-338 (2010).
38. Ferragina, P., Giancarlo, R., Manzinni, G.: The Myriad Virtues of Wavelet Trees. ICALP, pp. 561-572 (2006), journal version in Information and Computation 207 (2009), 849-866.
39. Ferragina, P., Giancarlo, R., Manzini, G.: Sciortino, M.: Boosting textual compression in optimal linear time. Journal of the ACM, 52(4),pp. 688-713, (2005).
40. Ferragina, P., Gonzalez, R., Navarro, G., Venturini, R.: Compressed Text Indexes: From theory to practice. ACM Journal of Experimental Algorithmics 13: (2008).
41. Ferragina, P. Lucio, F., Manzinni, G., Manzini, G., Muthukrishnan, S.: Structuring labeled trees for optimal succinctness and beyond. IEEE FOCS, pp. 1-9,(2005).
42. Ferragina, P., Lucio, F., Manzini, G., Muthukrishnan, S.: Compressing and Searching XML data via two zips. WWW, pp. 751-760, (2006).
43. Ferragina, P., Manzini, G.: Indexing Compressed Text. Journal of the ACM, 52(4), pp. 552-581, (2005).
44. Ferragina, P., Manzini, G., Mäkinen, V., Navarro, G.: An Alphabet-Friendly FM-Index. SPIRE: 150-160, (2004).

45. Ferragina, P., Manzini, G., Makinen, V., Navarro, G.: Compressed representation of sequences and full-text indexes. ACM Transactions on Algorithms 3(2): (2007)
46. Ferragina, P., Venturini, R.: A simple storage scheme for strings achieving entropy bounds. SODA: pp. 690-696, (2007).
47. Ferragina P., Venturini, R.: The compressed permuterm index. ACM SIGIR, 535-542, (2007).
48. Fischer, J.: Data structures for efficient string algorithms. Phd Thesis, (2007).
49. Foschini, L., Grossi, R., Gupta, A., Vitter, J.: When indexing equals compression: experiments with compressing suffix arrays and applications. ACM Transactions on Algorithms, 2(4), pp. 611-639, (2006).
50. Foschini, L., Grossi, R., Gupta, A., Vitter, J.: Fast compression with a static model in high-order entropy. Data Compression Conference: pp. 62-71 (2004).
51. Gagie, T., Navarro, G., Puglisi, S.: New Algorithms on Wavelet Trees and Application to Information Retrieval, SPIRE (2009), 1-6, also as full version in CoRR abs/1011.4532: (2010)
52. Gagie, T., Navarro, G., Puglisi, S.: Colored range queries and document retrieval. SPIRE, pp. 67-81 (2010).
53. Gagie, T., Puglisi, S., Turpin, A.: Range quantile queries: another virtue of wavelet trees, In Proc. 17th International Symposium on String Processing and Information Retrieval (SPIRE), LNCS 5721, pp. 1-6, (2009), also as arXiv:0903.4726v6.
54. Golynski, A.: Optimal lower bounds for rank and select indexes. In Proc. ICALP, 370-381, (2006).
55. Golynski, A., Grossi, R., Gupta, A., Raman, R., Rao, A.: On the size of succinct indices. ESA, pp. 371-382, (2007).
56. Golynski, A., Raman, R., Rao, S.: On the redundancy of succinct data structures. SWAT, pp. 148-159, (2008).
57. Golynski, A., Munro, J., Rao, S.: Rank/select operations on large alphabets: a tool for text indexing. SODA, pp. 368-373, (2006).
58. Gonzalez, R., Grabowski, S., Makinnen, V., and Navarro, G.: Practical implementation of rank and select queries. In the 4th Workshop on Efficient and Experimental Algorithms (WEA), 2005, pp. 27-38.
59. Gonzalez, R., Navarro, G.: Improved dynamic rank select entropy-bound structures. LATIN 2008, LNCS 4957, pp. 374–386, (2008) (also as Rank/Select on Dynamic Compressed Sequences and Applications, Theor. Comput. Sci. 410(43): 4414-4422 (2009))
60. Gonzalez, R., Navarro, G.: Compressed Text Indexes with Fast Locate. CPM, pp. 216-227, (2007).
61. Gonzalez, R., Navarro, G.: Statistical encoding of succinct data structures. CPM: pp. 294-305, (2006).
62. Gonzalez, R., Navarro, G.: A compressed text index in secondary memory. IWOCA, 80-91, (2007).
63. Greve, M., Jorgensen, A.G., Larsen, K.D., Truelsen, J.: Cell probe lower bounds ad approximations for range mode. ICALP, pp. 605-616 (2010).
64. Grossi, R., Gupta, A., Vitter, J.: High-Order Entropy-Compressed Text Indexes. ACM-SIAM SODA, 841-850, (2003).
65. Grossi, R., Orlandi, A., Raman, R.: Optimal trade-offs for succinct string indexes. ICALP (1): pp. 678-689, (2010).
66. Grossi, R., Orlandi, A., Raman, R., and Rao, S.:, More haste less waste: lowering the redundancy in fully indexable dictionaries. Symposium on Theoretical Aspects on Computer Science, 517-528 (2009).

67. Grossi, R., Vitter J.S..: Compressed suffix arrays and suffix trees with applications to text indexing and string matching. SIAM Journal on Computing, vol. 35, no. 32. pp. 378-407 (2005).
68. Grossi, R., Vitter J., Xu, B.. Wavelet trees from theory to practice. In Proc. of International Conference on Data Compression, Communication and Processing (2011)
69. Gupta, A., Hon, W., Shah, R., Vitter, J.: Compressed data structures: Dictionaries and data-aware measures. In Proc. 16th DCC, 213-222, (2006).
70. Gupta, A., Hon, W., Shah, R., Vitter, J.: A framework for dynamizing succinct data structures. In Proceedings of the 34th International Colloquium on Automata, Languages and Programming, pp. 521-532, (2007).
71. He, M., Munro, J.: Succinct Representations of Dynamic Strings. SPIRE 2010: 334-346, see also CoRR abs/1005.4652: (2010)
72. He, M., Munro, J., Rao, S.: A Categorization Theorem on suffix arrays with applications to space efficient text indexes. ACM SIAM SODA, pp. 23-32, (2005).
73. Hong, W.K., Sadakane, K., Sung, W.K.: Succinct data structures for searchable partial sums. In Proc. ISAAC, LNCS 2906, pp. 505-516, (2003).
74. Hon, W.K., Shah, R., Thankachan, S., and Vitter, J.S., String Retrieval for Multi-pattern queries. In SPIRE, LNCS 6393, pp. 55-66 (2010).
75. Hon, W.K., Shah, R., Vitter, J.: Ordered Pattern Matching: towards full-text retrieval. TR 2006-8, Purdue University, (2006).
76. Hon, W.K., Shah, R., Vitter, J.: Space-efficient framework for top-k string retrieval problems. In FOCS, pp. 713-722, (2009).
77. Hon, W.K., Shah, R., Vitter, J.: Efficient index for retrieving top-k most frequent documents. SPIRE, pp.182-193, (2009), journal version in J. Discrete Algorithms 8(4): 402-417 (2010).
78. Hon, W.K., Shah, R., Vitter, J.: Compression, Indexing and retrieval for Massive String Data. CPM 2010: pp. 260-274, (2010).
79. Jacobson, G.: Space efficient static trees and graphs. In Proc. of FOCS, pp. 549-554, (1989).
80. Karpinski, M., Nekrich, Y.: Space efficient multidimensional range reporting. CoRR abs/0806.4361: (2008)
81. Karpinski, M., Nekrich, Y.: Top-K Color queries for document retrieval. SODA: 401-411, (2011).
82. Kreft, S., Navarro, G.: Self-indexing based on LZ77. DCC: pp. 239-248 (2010).
83. Larsson, J., Moffat, A.: Off-line dictionary-based compression. Data Compression Conference: 296-305 (1999).
84. Lee, S., Park, K.: Dynamic rank/select structures with applications to run-length encoded texts. Theoretical Computer Science 410, pp. 4402-4413 (2009).
85. Lee, S., Park, K.: Dynamic Compressed Representation of Texts with Rank/Select. Journal of Computing Science and Engineering, Vol. 3, No. 1, March, pp. 15-26 (2009).
86. Makinen, V., Navarro, G.: Dynamic Entropy Compressed Sequences and Full-Text indexes. ACM Transactions on Algorithms, Vol. 4, No. 3, Article 32, Publication date: June (2008).
87. Makinen, V., Navarro, G.: Position Restricted Substring Searching. LATIN 2006, pp. 703-714, (2006).
88. Makinnen, V., Navarro, G.: Implicit Compression Boosting with Applications to Self Indexing. SPIRE, pp. 229-241 (2007).
89. Makinnen, V., Navarro, G.: On self-indexing images – image compression with added value. DCC, 422-431, (2008).

90. Makinnen, V., Navarro, G.: Rank and select revisited and extended. Theoretical Computer Science 387(3): pp. 332-347 (2007)
91. Makinnen, V., Navarro, G., Siren, J., Valimaki, N.: Storage and retrieval of highly repetitive sequence collections. JCB (2009).
92. Munro, J.: Tables. In Proc. Of FSTTCS, pp. 37-42, (1996).
93. Muthukrishnan, S.: Efficient algorithms for document retrieval problems. In SODA 2002, pp. 657-666 (2002).
94. Navarro, G., Makinen, V.: Compressed full text indexes. ACM Computing Surveys, 39(1): (2007).
95. Navarro, G., Puglisi, S.: Dual sorted inverted lists. In Proc. 17th International Symposium on String Processing and Information Retrieval (SPIRE), LNCS 6393, pp. 310-322, (2010).
96. Navarro, G., Puglisi, S.J., Valenzuela, D.: Practical Compressed Document Retrieval. SEA: 193-205, (2011).
97. Navarro, G., Sadakane, K.: Fully functional static and dynamic succinct trees. CoRR abs/0905.0768: (2009), see also the conference version entitled Fully functional succinct trees, in SODA, 2010, pp. 134-149.
98. Okanohara, D., Sadakane, K.: Practical entropy-compressed rank-select dictionaries. In ALENEX (2007).
99. Pagh, R.: Low redundancy in static dictionaries with constant query time. SIAM J. Computing, 31(2):353-363, (2001).
100. Patrascu, M.: Succincter. IEEE FOCS, pp. 434-443, (2008).
101. Puglisi, S., Smyth W.F., Turpin, A.: Inverted files versus suffix arrays for locating patterns in primary memory. In SPIRE, LNCS 4209, pp. 122-133, (2006).
102. Raman, R., Raman, V., Srinivasa, S.: Succinct dynamic data structures. WADS, 426-437, (2001).
103. Raman, R., Raman, V., Rao, S.S.: Succinct indexable dictionaries with applications to encoding k-ary trees and multisets. In SODA 232-242, 2002, see also the full version in ACM Transactions on Algorithms 3(4): (2007).
104. Russo, L., Oliveira, A.: A compressed self-index using a Ziv-Lempel Dictionary. Information Retrieval, 5(3):501-513, (2008).
105. Sadakane, K., Grossi, R.: Squeezing succinct data structures into entropy bounds. In Proc. 17th SODA, pp.1230-239, 2006.
106. Schnattinger, T., Ohlebusch, E., Gog, S.: Bidirectional search in a string with wavelet trees. CPM, LNCS 6129, pp. 40-50, (2010).
107. Shannon, C.E., and Weaver W.: A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423, 623-656, (1948).
108. Siren, J., Valimaki, N., Makinnen, V., Navarro, G.: Run-Length compressed indexes are superior for highly repetitive sequence collections. SPIRE, 164-175, (2008).
109. Stoye, J.: Affix trees. Technical report 2000-04, University of Bielefeld (2000).
110. Strothmann, D.: The affix array data structure and its applications to RNA secondary structure analysis. Theoretical Computer Science 389, 278-294, (2007).
111. Valimaki, N., Makinnen, V.: Space efficient algorithms for document retrieval. CPM, pp. 205-215, (2007).
112. Witten, I., Moffat, A., Bell, T.: Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kauffman, 1999.
113. Yu, C.-C., Hon, W.-K., and Wang B.-F.: Efficient data structures for the orthogonal range successor problem. In COCOON, LNCS 5609, pp. 96-105, (2009).

114. Yu, C.C., Hon, W.-K., Wang, B.-F.: Improved data structures for the orthogonal range successor problem. Computational Geometry 44, pp. 148-159, (2011)

Christos Makris is an Assistant Professor in the Department of Computer Engineering and Informatics, School of Engineering, University of Patras, Greece. His research interests include Data Structures, Web Algorithmics, Computational Geometry, Data Bases and Information Retrieval. He has published over 80 papers in various scientific journals and refereed conferences and has over 250 references. Lists of paper topics**:** information retrieval, text algorithms, data structures, compression.

Received: June 06, 2011; Accepted: February 06, 2012.

DOI: 10.2298/CSIS110525003F

Impact of Personnel Factors on the Recovery of Delayed Software Projects: A System Dynamics Approach

Mostafa Farshchi, Yusmadi Yah Jusoh, and Masrah Azrifah Azmi Murad

Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 Serdang, Selangor, Malaysia
farshchi; yusmadi; masrah@fsktm.upm.edu.my

Abstract. Delay in a software project may result in the loss of a market opportunity or the postponement of a dependent project. Therefore, software project managers take various steps to ensure that their project is completed on time, such as adding new members to the project team. However, adding new manpower to a delayed project may cause a negative impact on the team's productivity due to assimilation time, training overhead and communication overhead. Consequently, project managers have difficulty in making the decision on whether or not to add new members to the team. Thus, this research aims to examine whether a significant schedule improvement can be achieved with consideration of the new manpower's capabilities, skills and experience. A System Dynamics Model is proposed to simulate the behaviour of a project's progress when new members are added. The proposed model was evaluated through experiments using two types of case studies. The results of the experiments indicate that a significant schedule improvement of a late project can be achieved if people with certain levels of personnel factors are added to the project.

Keywords: software project management, personnel factors, system dynamics, schedule delay.

1. Introduction

Despite recent advances in software project management technology and methods, project failure rates are still considerably high. Many software project failures are reportedly due to problems related to delayed delivery and deferred deployment [1; 2]. Late completion of a software development project cannot be tolerated because it may cause delay to a dependent project or result in the potential loss in profits due to a missed business opportunity. Time delay in such situations could incur a much higher cost than the cost of the project itself. Consequently, the management would undertake any possible measure to ensure that their project is delivered on time. One of the common practices in dealing with schedule delays is adding new manpower [3; 4; 5; 6; 7].

Although adding new members to an ongoing software project team is anticipated to increase the available effort to the team, it may result in the loss of productivity due to assimilation delay, inter-personnel communication overhead, and training overhead. The above phenomenon was first explained by Frederick Brooks [8] and became known as Brooks' Law: "adding manpower to a late software project makes it later".

Accordingly, several studies have been reported to address the problem: the issue was investigated formally by Abdel-Hamid and Madnick [3]; in another study, an attempt to address the problem through mathematical expression was carried out by Stutzke [9]; later, an investigation of the impact of task constraints was performed by Hsia *et al.* [10]; and a study on the impact of pair programming was conducted by Williams *et al.* [11].In addition, an exploration of the issue through a newly-structured System Dynamics approach was presented by Madachy [12], and some assessments of the impact of the issue on open-source projects were recently conducted [13; 14; 15; 16; 17] .

As a result of the noted effects, a project manager faces difficulty in making the decision on whether or not to add new members to his team. Software development is a human-intensive process [18]. Several studies have reported the high contribution of personnel attributes on software development productivity, such as those carried out by Finnie *et al.* [19], Boehm *et al*. [20], Trendowicz and Munch [21], and Hannay *et al.* [22]. As highlighted above, scholars have studied the issue from various aspects. However, the effect of new manpower capability, skills and experience in addressing the problem has yet to be studied further. Therefore, this study tries to understand:

- How software project managers can minimise the negative effects of adding new manpower to delayed software projects using personnel factors trade-off analysis; and
- Whether a significant schedule improvement of a late project can be achieved by taking into consideration the new manpower capabilities, skills and experiences.

If a project manager wishes to add new members, then to succeed, the manager must look for ways to minimise the negative impact of communication overhead, training overhead, and assimilation delay. Hence, in this research, we attempt to assess how effectively the variety of personnel factors and capabilities can influence the progress of a late software project. Obviously, it's likely that a more skilful person would add more productivity to a project team than a less skilful one, but whether the difference would be significant enough to prevent schedule slippage or would just be a waste of resources is the concern of this study. This research aims to improve software project management by reducing the failure rate in meeting the schedule. Therefore, this study extends the work of previous research in both software process simulation and software project management in the following ways: Firstly, it proposes a System Dynamics simulation model that is capable of effectively simulating the impacts of adding new manpower with distinct levels of skills, experience and capabilities to the progress of ongoing software projects. Secondly, it assesses the effects of the personnel factors of new manpower on selected factors through an examination of industry experiments, and provides an analysis of the benefits and drawbacks of considering the capabilities of new manpower added to ongoing projects. Thirdly, it provides a tool that empowers software project managers to become proactive in their decision making, enabling them to perform trade-off analysis and forecast the likely consequences of their decisions in altering manpower in a project.

The rest of this paper is organised as follows: In Section 2, related work in the study is presented and reviewed. Section 3 explains the approach taken in conducting the research and constructing the System Dynamics simulation model. Section 4 gives the results and discusses the findings of executing the proposed model in two case studies by comparing them with a reference model, and then validating the model against a real case study. Next, in Section 5, the model evaluation is discussed. In last section, we summarise our insights into the study and propose directions for future research.

2. Related Work

This section discusses the related literature concerning the impact of adding new manpower to delayed software projects.

2.1. The Impact of Adding New Manpower to Delayed Software Projects

As mentioned in the previous section, adding more manpower to a delayed software project would likely increase its completion time. This was explained as a consequence of the following three effects: (i) new staff need training and they need to be trained by experienced staff; (ii) there is an assimilation delay for new staff to become productive in a project; (iii) adding new people increases the communication overhead among the members of a project team. The above phenomenon is known as Brooks' Law, credited to years of industry experience [3; 23; 24; 25; 26]. However, the non-Brooks' Law phenomenon has also been reported by some researchers and organisations by utilising certain techniques [11; 25; 27; 28].

The study by Abdel-Hamid *et al.* revealed that adding more people to a late project did not always cause it to complete later. He assumed in his study that there is, in general, a desire among project managers to change the composition of their workforce. However, if the perceived remaining time is shorter than the anticipated hiring and assimilation delay, then the project manager would not add new manpower. Although the comprehensive software project model he proposed has been examined through a case study, the effect of changing manpower has not been examined due to the limitations of the case study.

Stutzke [9] investigated the possible situations in which a project may benefit by adding new members to the workforce. Results derived from his mathematical model indicate that adding new staff would not necessarily have a negative impact on a project if certain constraints are taken into account. These constraints are: $r > (1 + m) * a \ \ and \ f \leq 1/m$ where f = fractional increase in the staff; r = remaining time to complete project, a = assimilation time, and m = mentoring cost (fraction of staff member's time spent mentoring one new hire). Stutzke’s model has one weakness, which is that is does not represent the communication overhead. His model also didn’t take into account how the levels of capability and experience of the new hires can affect assimilation time and training needed.

In recent years, the effects of pair programming on software development productivity have gained more attention [11; 22; 29; 30; 31]. Williams *et al.* utilised Stutzke's mathematical equations to explore the relationship between pair programming and Brooks’ Law. Their study implies that practising pair programming helps reduce assimilation delay and mentoring time. Although their result shows improvement in assimilation time and training overhead, the effect of pair programming hasn’t been investigated on a project’s overall communication overhead.

Through literature review, it is noted that human factors, such as lower programmer capability and insufficient experience, are the main causes of software project delays [32; 33]. For instance, the result obtained from analysing thirty-one causes of project delays by Ziya Ma [32] shows that “insufficient experience of developers” and “inappropriate access to development and test tools” are the most frequently reported causes of delay. Therefore, attempts have been made to further explore the effects of personnel factors, particularly their impact on the productivity levels of software projects.

In the field of software engineering, researchers have long acknowledged the importance of Brooks’ Law. Results obtained from recent literature review indicate the growing trend of issues that were initially pointed out by Brooks [24; 25]. In recent years, the reinvestigation of Brooks’ Law for projects of different types and characteristics has been suggested. For example, Callegari and Bastos [4] declared that “Brooks’ Law needs more investigation when considering a project’s characteristics, such as size and available time”. Meanwhile, recent studies show, in contrast to Brooks’ Law, that more programmers contributing to open-source software projects led to the resolution of difficult problems and the success of the software [13; 14; 15; 16; 17]. Schweik and English [13] are among those who argue that a higher number of people on free and open-source projects would contribute to the success of the project. The summary of approaches, factors and project types that have been studied before are presented in Tables 1 and 2.

Table 1. Summary of previous studies on adding new manpower to late software projects

Author	Approach	Main Parameters	Project Type
Abdel-Hamid and Madnick [3]	System Dynamics	Communication Overhead, Hiring Delay, Assimilation Time, Training Overhead	Commercial
Stutzke [9]	Mathematical Expression	Mentoring Time, Effective Assimilation Time	Commercial
Hsia *et al.* [10]	System Dynamics	Task Constraints, Overhead, Training Time, Assimilation Time,	Commercial
Williams *et al.* [11]	Mathematical Expression	Mentoring Time, Effective Assimilation Time, Pair Programming	Commercial
Madachy [12]	System Dynamics	Communication Overhead, Assimilation Time, Training Overhead	Commercial
Schweik and English [13]	Statistical	Inter-Personnel Communication Rate	Open-Source
Capiluppi and Adams [15]	Statistical	Inter-Personnel Communication Rate	Open-Source

Table 2. Main parameters of previous models on adding new manpower to late software projects

Author	New Personnel (NP) vs. Experienced Personnel (EP) Productivities	Assimilation time (workdays)	Training Time	Communication Overhead

Abdel-Hamid and Madnick [3]	EP = 1 task/ person-day
	NP = 0.5 task/ person-day
		80	20.2 % of the time of an experienced person during assimilation time	$0.06 \times n^2$
				Where n is the total manpower

Stutzke [9]	Not differentiated	33.4	25.0 % of the time of an experienced person during mentoring time	Not considered

Hsia *et al.* [10]	EP = 1 task/ person-day
	NP = 0.5 task/ person-day
		80	20.2 % of the time of an experienced person during assimilation time	$0.06 \times n^2$
				Where n is the total manpower

Williams *et al.* [11]	Not differentiated	27(with pair programming: 12)	37.0 % (with pair programming: 26.0 %) of the time of an experienced person during mentoring time	Not considered

Madachy [12]	EP = 1.2 function points/ person-day
	NP = 0.8 function points/ person-day
		20	25.0 % of the time of a fulltime, experienced person during assimilation time	$0.06 \times n^2$
				Where n is the total manpower

2.2. Personnel Factors

It was found through literature review that personnel factors, such as inadequate programmer capability and insufficient experience, are among the main causes of delays in software projects [32; 33]. Practitioners identify people as the main contributor of a project's success but project managers argue that although people are indeed key, their actions sometimes contradict their words [18]. Product complexity and personnel/team capability have been reported to have the highest influence on the development rate of a software project [20; 21; 22]. Trendowicz and Munch [21] identified 249 different factors that influence software development productivity through literature review. Their study shows that team capability and experience (personnel factors) were the most commonly cited factors. Their results are shown in Figure 1.

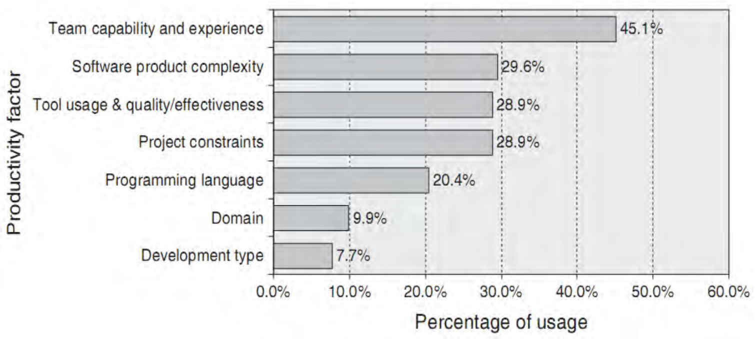


Fig. 1. Most common productivity factors reported in literature, adopted from [21]

Carpetz and Ahmed [34] emphasised the importance of skill diversity in the software engineering field. They argue that "the diversity of skills contributes to problem solving as different people see a problem from different perspectives". Various studies have reported the influence of a variety of these factors. For instance, the effect of the experience and ability of software developers on the comprehension of web applications has been studied by Ricca *et. al.* [35], and a study on the impact of developer personality on pair programming has been conducted by Hannay *et al.* [22].

The importance of people in the productivity of software development projects, as emphasised in literature, is the motivation for this study to focus on personnel attributes. Many studies have focused on and ranked the importance of different productivity factors, however, very few studies have provided numeric values for the productivity ranges of these factors. Among others, Boehm *et al.* proposed the productivity ranges of different effort multipliers for estimating the cost and schedule of software projects, using the Constructive Cost Model (COCOMO).

COCOMO is a very well-known cost/schedule estimation model for software projects. The COCOMO II.2000 is the latest version available for this model [20]. The complete Bayesian analysis on COCOMO II yields the

Productivity Ranges that provide an insight into identifying the high payoff areas to focus on in improving software productivity [36]. Therefore, personnel factor productivity multipliers of the COCOMO II model are employed for the purpose of this study.

3. The Proposed Model

The development of a System Dynamics model involves the identification of model parameters and variables, equations and relationships. These include the modelling of a software development process that comprises factors representing communication overhead, training overhead, and assimilation delay. Personnel capability factors should then be added to the model.

In order to construct and execute the model, a simulation environment with the capability to execute the model is needed. A simulation environment allows the experimentation of the developed model, which provides an opportunity to observe dynamic interactions and results. In this study, the IThink simulation environment was chosen to implement the proposed System Dynamics model and execute the simulation this was due to: (a) it is a robust and user-friendly model that covers all the needs of the current study; (b) its license was available; and (c) it has been used by others and reported to be a reliable tool [12; 37].

3.1. The System Dynamics Model

The main components (associated parameters that represent process flows, and the different factors that affect these process flows) of the base model along with extended sections are explained in this section. The main components of the model include: - Software Production Flow; - Personnel Allocation Flow; - Assimilation Time; - Training Overhead; -Communication Overhead; - Personnel Factors (all six factors).

Software Production Flow: The System Dynamics model of a software project simulates the high level of a software development process. It uses production rate and predicts the completion date of a project. The relationship between the amount of original (or remaining) work and the current development rate determines a project's scheduled completion date. The main contributors to the software production rate are effort and productivity.

Personnel Allocation Flow: A gap between the planned schedule and the actual progress of a software project would trigger the need to add new manpower. As shown in Figure 2, this trigger is represented in a model, with an auxiliary element called the personnel allocation trigger. We have also allocated a separate level for new personnel to differentiate their influence on software production from that of the initial personnel of a project..

Assimilation Time: New personnel need time to train in order to become a productive member of a project team. This training duration is called assimilation time. On average, once new personnel pass the assimilation time, they become as productive as experienced personnel. Thus, we added an assimilation rate to the personnel allocation flow chain. The values of assimilation delays reported in literature are listed in Table 2.

Training Overhead: The existing, experienced personnel need to spend portions of their effort for mentoring new personnel. This leads to an effort loss of experienced personnel as the result of mentoring. The reported values for training (mentoring) time in literature are listed in Table 2.

Communication Overhead: Communication overhead directly affects the software production rate. It is widely held that communication overhead increases in portion to n^2 , where n is the size of the team [3] [10; 12]. Therefore, the communication overhead is expressed using an auxiliary element in the model, as depicted in Figure 2, with a non-linear function as:

$$CO = 0.06 \times NP^2 , \qquad (1)$$

$$where, CO \ is \ Communication \ Overhead ,$$

$$and \ NP \ is \ the \ total \ Number \ of \ Personnel .$$

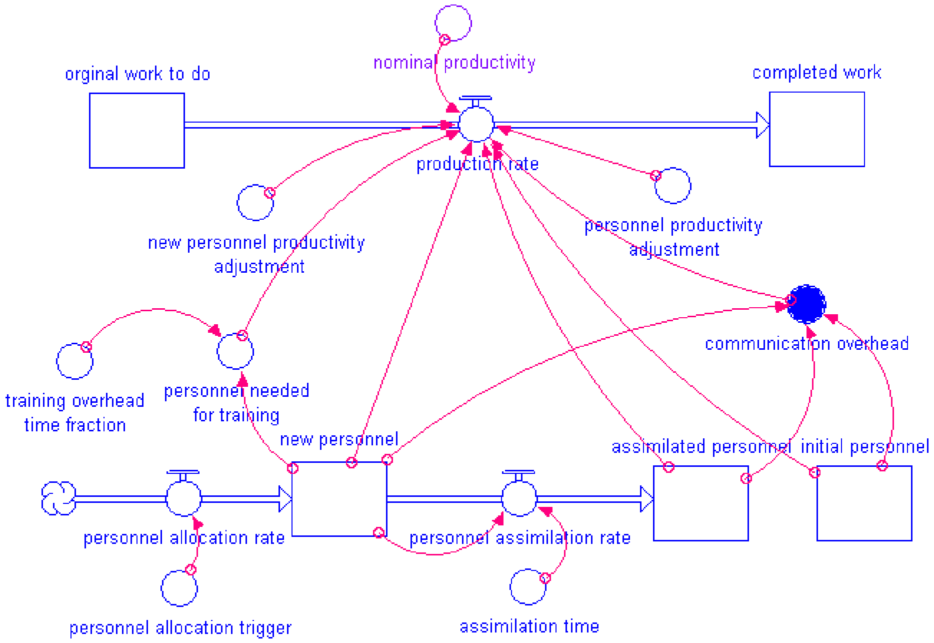

nominal productivity
orginal work to do
completed work
production rate
new personnel productivity adjustment
personnel productivity adjustment
communication overhead
training overhead time fraction
personnel needed for training
new personnel
assimilated personnel
initial personnel
personnel allocation rate
personnel assimilation rate
personnel allocation trigger
assimilation time

Fig. 2. Model representing training overhead, communication overhead and assimilation time

3.2. Representation of Personnel Factors in the Model

It has been stated in literature review that among related publications, the productivity drivers of COCOMO II.2000 [20] present the most appropriate values that can be employed as input variables of the personnel capabilities of our System Dynamics simulation model. The data range of productivity factors in COCOMO II shows that the combination of human factors provides the highest productivity range compared to all other types of factors, such as process, product, project and organisation. Six personnel factors from COCOMO.II(2000) [20] were employed for our study, namely, Analyst Capability (ACAP), Programmer Capability (PCAP), Personnel Continuity (PCON), Applications Experience (APEX), Platform Experience (PLEX), and Language and Tools Experience (LTEX). The productivity ranges of these human factors are shown in Figure 3.

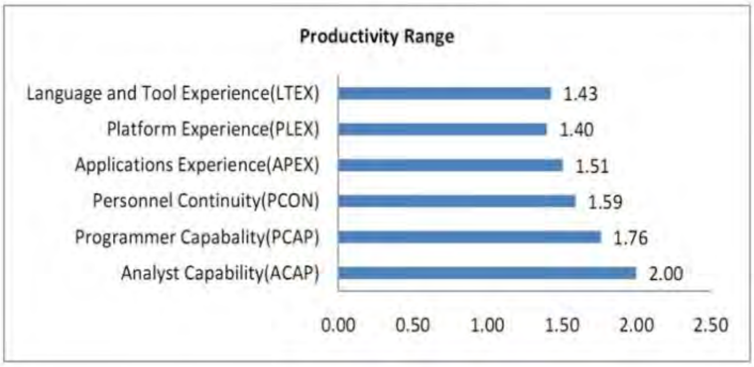


Fig. 3. Productivity ranges of COCOMO II personnel factors

Table 3. Productivity influence of COCOMO II personnel factors extracted for different levels

Personnel Factors	Very Low	Low	Nominal	High	Very High
Analyst Capability (ACAP)	0.7042	0.8403	1.00	1.1765	1.4085
Programmer Capability (PCAP)	0.7463	0.8696	1.00	1.1364	1.3158
Personnel Continuity (PCON)	0.7752	0.8929	1.00	1.1364	1.2346
Application Experience (APEX)	0.8197	0.9091	1.00	1.1364	1.2346
Platform Experience (PLEX)	0.8403	0.9174	1.00	1.0989	1.1765
Language and Tool Experience (LTEX)	0.8333	0.9174	1.00	1.0989	1.1905

We needed to make the model capable of simulating the productivity of new personnel with different levels of productivity ranges as personnel factors. The ratios between very low, low, high and very high productive parameter ratings, and the nominal productive parameter rating were extracted based on the effort multiplier values of personnel factors given in the COCOMO II cost/schedule estimation model. In other words, the productivity rate of a person with different levels of capability in comparison to a person with the nominal level is extracted as values are reported in Table 3.

In order to simulate the influence of new manpower with different capabilities, one auxiliary element was allocated in the model for each of the above personnel capabilities in the model, namely, LTEX, PLEX, APEX, PCON, PCAP, and ACAP, as shown in Figure 4.

For the next step, we had to identify the parameters that might be influenced by personnel factor productivity multipliers. Therefore, the relevant relationships were applied and necessary equations were updated in the model, as depicted in Figure 4.

So far, it has been assumed that nominal values for assimilation time and training overhead will be applied for all new personnel regardless of their personnel experience and capabilities. However, it is rational to assume that a more capable and more experienced person would need less training and would become assimilated sooner than a person who is less capable and less experienced. Therefore, we have attempted to represent the impact of personnel factors on training overhead and assimilation time.

Based on the above explanation, two cause-effect relationships were added to the model, one between 'personnel factor multipliers' and 'assimilation time', and another between 'personnel factor multipliers' and 'training overhead time fraction', as illustrated in Figure 4. The affected factors were then proportionally adjusted according to the values given for the personnel productivity multipliers.

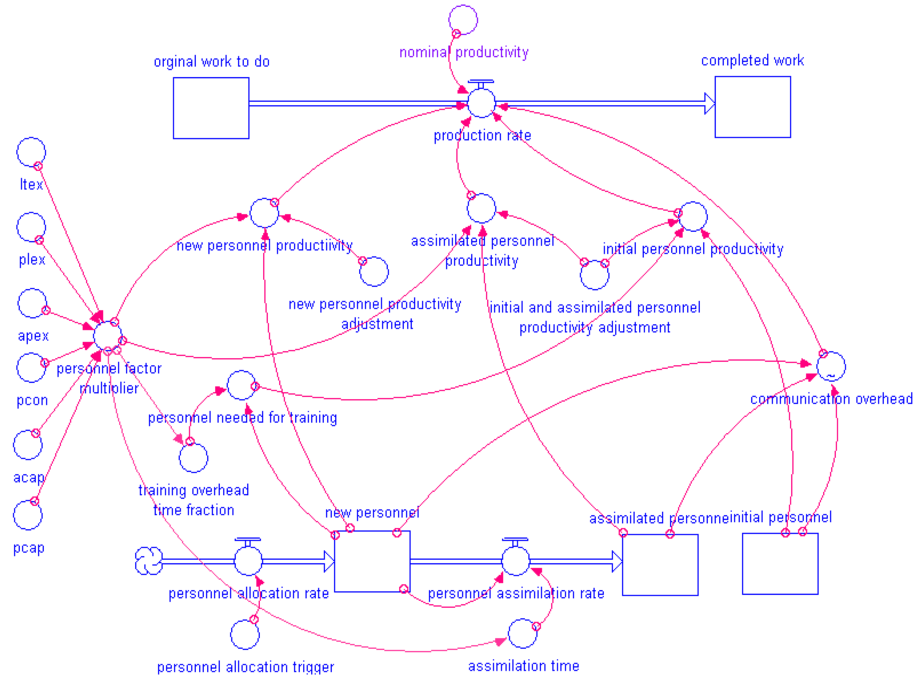

nominal productivity
orginal work to do
completed work
production rate
ltex
plex
apex
pcon
acap
pcap
personnel factor multiplier
new personnel productivity
assimilated personnel productivity
initial personnel productivity
new personnel productivity adjustment
initial and assimilated personnel productivity adjustment
personnel needed for training
communication overhead
training overhead time fraction
new personnel
assimilated personneinitial personnel
personnel allocation rate
personnel assimilation rate
personnel allocation trigger
assimilation time

Fig. 4. The System Dynamics Model

After including all necessary factors into the model, it was now ready for simulation experiments.

4. Model Experiment - Case Study Taken from Literature

For the current research, we applied two case studies for model experiment and simulation analysis of the proposed model: (1) A case study taken from literature, and (2) A case study from an industry software project.

As the proposed model extends Madachy's model structure, and to have a basis for comparison, his reported case study was used to execute the simulation model. The implementation of the same case study from literature provided us with the advantages of evaluating the validity of our model and observing whether it produces the same results as that of the base model.

4.1. Simulation Model Settings

Variables and parameters of a model need to be initialised before running a simulation. Table 5 shows the list of parameters that were initialised.

Table 5. Simulation model specifications from literature

Parameter Name	Value
Original work	500
Initial (Experienced) Personnel	20
Estimated completion duration	274
Assimilation time	20
Communication overhead	0.06 * n^2
Training overhead	0.25
New personnel	0, 5, 10, 15
Productivity	Nominal productivity = 0.1 FP
New vs. Initial (experienced) personnel productivity	Productivity adjustment for experienced personnel = 1.2 Productivity adjustment for new personnel = 0.8

4.2. Experiment with Base Model Setting

Once its parameters are initialised, the model is ready to execute the simulation and analyse simulation behaviours and results. To emulate the base model setting, all personnel factor productivity multipliers in the model were set to their nominal values. Therefore, we expect the constructed model to reproduce the same simulation behaviours as reported for the base model.

For the first experiment, the model simulates the "as-is" (or default) situation, with no new manpower addition to the project. Curve #1 in Figure 5 shows the software production rate and scheduled completion day without adding new manpower. Based on the specifications of the case study, the overall software production rate of the 20-member project team is consistently equal to “1.82” function points. Based on this production rate, the project was forecast to be completed in 274 days. In the next simulation, the personnel allocation trigger was set to add 5 new personnel at about Day 110. As can be observed in curve #2, when 5 new people were added, the software team’s production rate dropped suddenly. However, it recovered during the assimilation time and eventually exceeded the default production rate. This interesting effect resulted in the project finishing three days sooner, on Day 271. Although a 3-day reduction in project duration is hardly significant, it demonstrates that Brooks’ Law doesn’t hold in all situations. Curve #3 shows the simulated process behaviour with the addition of 10 new people to the software team. The drop in development rate caused by the negative effects of adding 10 people never did correct itself to surpass or even catch up to meet the default level and the project was only completed on Day 296.

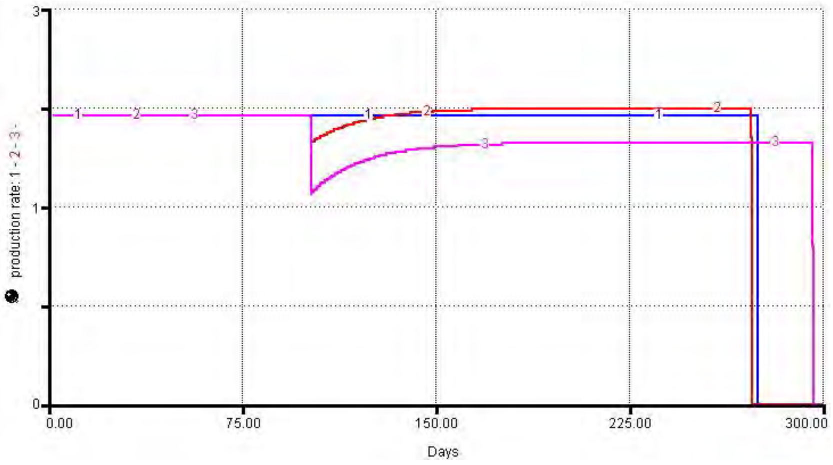


Fig. 5. Effects of adding manpower to a project at Day 110. Curve 1 - no new manpower added, curve 2 - five manpower added, curve 3 - ten manpower added

Above results demonstrates the proposed model reproduces the exact behaviour of that reported for the base model. By verifying our simulation result against a reference model, it shows that our model is reliable to use for further experiments.

4.3. Experiment of Adding New Manpower with High Levels of Capability and Experience

In this stage, we investigate how personnel attributes may impact the project completion schedule. To observe the behaviour of new changes to the model, the same case was used for this part as well. The hiring of new manpower with very high productivity rates is then simulated. Based on the actual conditions of the case study, it is very unlikely that a project manager would add analysts to a software team at about Day 110. Therefore, we did not include the effect of analyst capability in our simulation. In addition, we also set the value of personnel continuity at the nominal level in order to focus only on the impact of the new manpower's skills, experience and capabilities.

As demonstrated in Figure 6, curve #1 indicates the simulated behaviour of the default situation in which no one is added to the project. Next is the simulation of adding 5 new members with the highest level of capability to the project. Interestingly, the curve doesn't show any drop in production rate (relative to the base model), which leads to a significant schedule reduction (of around 36 work-days) and the project finishes on Day 238 (curve #2). The addition of 10 people to the project (Figure 6, curve #3) shows some schedule improvement, although slightly less effective than adding 5 people. However, adding 15 people (Figure 6, curve #4) causes a significant drop in production rate and the project eventually ends on Day 292.

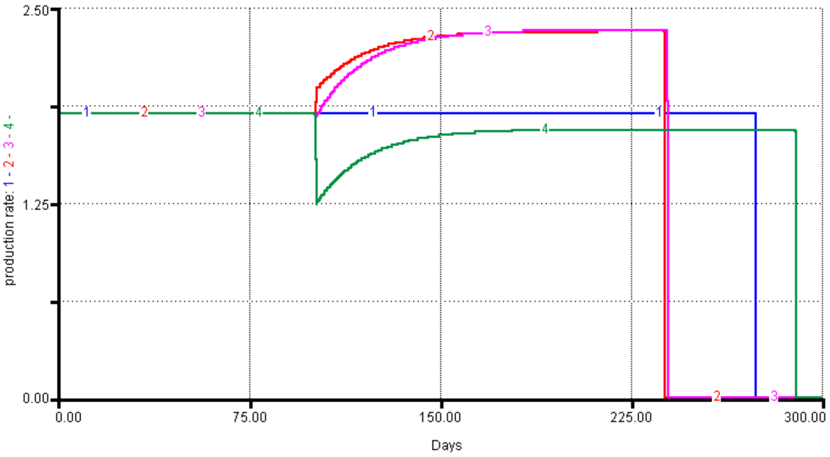


Fig. 6. Effect of adding new manpower with the best possible personnel factors to the project at Day 110. Curve 1 - no new manpower added, curve 2 - five new personnel added, curve 3 - ten new personnel added, curve 4 - fifteen new personnel added

The above result indicates that a significant schedule improvement can be achieved if people with a certain level of capability are added to ongoing software projects. The result shows that the schedule reduction was around 36 work-days. That is approximately two calendar months and can be significant enough to save a project from failure. Nevertheless, it was also found that adding a large number of people (Figure 6, curve #4), even those with high levels of capability and experience, can have a catastrophic result. Therefore, it is concluded that adding people with high levels of capability and experience can significantly minimise the negative effects of adding new manpower, and would accelerate the progress of a project only if the number of people added is reasonable.

5. Model Experiment - A Case Study from an Industry Software Project

A real-life case study was also adopted to validate our simulation model. We compared the results and calibrated the model using data obtained from a real-life case study. For the purpose of current research and according to the defined scope, we aimed to study the effects of adding new manpower to medium-sized, in-house, new development projects.

The above constraints are applied because the assumptions and model parameter values that have previously been identified are reported in project environments with such specifications. The information of case study is obtained from a well-known company in the production of on-demand enterprise software applications. The organisation is rated at Level 3 of the Capability Maturity Model-Integrated (CMMI) model for the Software Production division. The company's Project Office provided the information of a project that matched the needs and the constraints of our study. The Project Office is responsible for all the project planning and monitoring activities.

5.1. Simulation Model Settings

The Project Office and the project manager were asked through interviews to provide information on the project size, the number of initial personnel, the number of new people added to the project and the time they were added, the capability and experience of these new personnel, etc. Upon receiving these project specifications, we calibrated and customised the model according to the following characteristics of the previous case study.

Inconstancy of initial personnel: Based on the information obtained from the case study, we noticed a pattern in the Project Office's policy for personnel allocation. A project would usually begin with the minimum number of personnel required to perform the fundamental analysis and design activities. More personnel would progressively be added to the project for detailed design, development and testing activities. Gradually, after deploying and finalising individual modules, a few people will be released from the project. This is because sequential task constraints exist in different phases of a software project. For instance, in the construction phase, breaking tasks into smaller ones are more practicable than in the inception phase. Therefore, changes in the number of personnel were implemented into the model and the constant value of the initial personnel was converted into a variable that represents the number of personnel during the project lifecycle. Changes in the number of personnel wouldn't add overheads to the project as its effects had already been taken into account prior to project commencement.

Adding people with distinct capabilities: In the previous case study, the capabilities of new manpower were within the equivalent range. However, in this case people have different capabilities and experience. To manage such an issue, we designed a spreadsheet into which the personnel factors of individuals could be entered. The spreadsheet is then bound to its related representative factor of the model. This enabled the representation of new manpower with distinct attributes. The detailed list of people added, along with their capability and experience levels, is given Table 6.

Table 6. Capability and experience levels of new personnel added to the project (N: Nominal, VL: Very Low, L: Low, H: High, VH: Very High)

Personnel	PCAP	PCON	APEX	PLEX	LTEX
Person One	VH	N	H	VH	VH
Person Two	H	N	H	H	H
Person Three	H	N	N	H	H
Person Four	H	N	H	N	H
Person Five	N	N	H	N	N
Person Six	N	N	L	N	N

Based on the explanations given and the project specifications, the parameters of the model were initialised (as presented in Table 7), after which the simulation was ready to be executed and the simulated behaviour analysed using the results obtained.

Table 7. Simulation model specifications of the second case study

Parameter Name	Value	Description
Original work	760	Project Size in Function Points (FPs), equivalent to 152 main tasks, where each task, on average, is equal to 5 FPs.
Initial personnel	10 – 16	Equivalent number of fulltime personnel that varies during the project lifecycle.
Planned completion days	375	Estimated completion time in work-days
Assimilation time	40	Assimilation duration in work-days
Communication overhead	0.06 * n^2	n is the number of personnel who will be communicating with one another.
Training overhead	0.25	The portion (in this case, 25%) of a fulltime, experienced person’s time needed to train new personnel during the assimilation period.
New personnel	6	Number of personnel added from about Day 200.
Planned Nominal productivity	0.17	Expressed in function points/person-day.
Actual Nominal productivity	0.14	Expressed in function points/person-day.

Parameter Name	Value	Description
New vs. Initial(experienced) personnel productivity	New Personnel= 1 Exp Personnel= 0.5	New personnel productivity level is set to half that of experienced personnel

5.2. Simulation with Base Model Settings

The effects on the project's progress of adding people with the given capabilities are investigated in the following simulation run.

Curve #1 in Figure 7 plots the expected project progress (completed work in FPs) based on the project plan. This shows that the project was planned to be completed in 375 days. However, the project's actual progress did not match its planned progress. Curve #2 shows that the project progressed behind schedule. Based on the team's average production rate after 200 days, only 307 FPs were completed, which is 35 work-days (or 20%) behind the scheduled 387 FPs. The curves continue to diverge, widening the gap between the planned and actual schedules toward the end of the project, prompting the project management team to assign 6 more people.

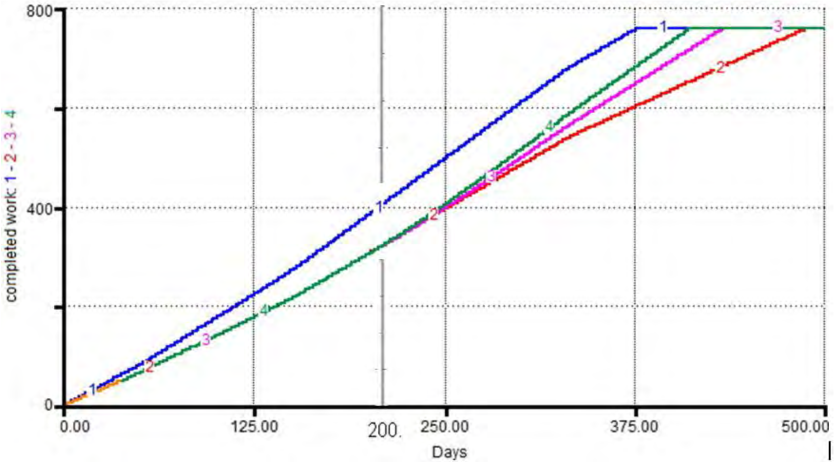


Fig. 7. Effect of adding new personnel to the project (without considering the impact of personnel factors). Curve 1 - planned, curve 2 - no new manpower added, curve 3 - six new personnel added on Day 200

To investigate the effect of adding six new people to the project, simulation was performed, first without consideration of their personnel attributes, as illustrated by curve #3. This shows that adding new people has a positive effect on the overall progress, but the project could still not sufficiently catch up and make up for the delay. Curve #4 simulates the same situation with consideration of personnel factors. Although the progress did not achieve the planned situation, the work was completed considerably faster (relative to curve #3) after adding people at around Day 200. This shows that adding new people with more than the nominal capability level has a positive impact on a project's progress.

Simulations of the team's production rates are shown in Figure 8, Curves #1 and #2 show the changing production rates of the software team throughout the project lifecycle, caused mainly by the changing number of personnel. In the project plan, team members were expected to achieve an average productivity rate of 3.48 FPs per month. However, as at Day 200, the actual average production rate was only 2.76 FPs per month. Curve #2 forecasts the project completion based on an average production rate of 2.76 FPs per month, giving an estimate that the project would be completed in 486 days. Unfortunately, this is more than 100 days later than planned.

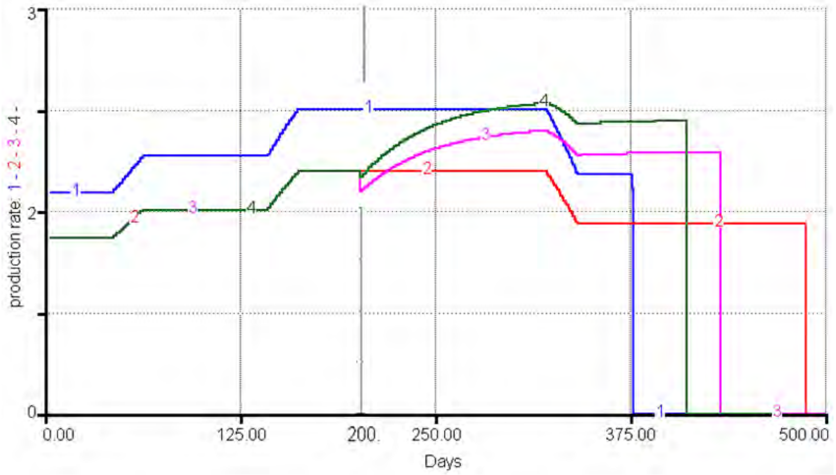


Fig. 8. Effect of adding new personnel to the project (with and without considering the impact of personnel facotrs). Curve 1 - planned, curve 2 - no new manpower added, curve 3 - six manpower added on Day 200, curve 4- six new personnel added on Day 200 with including the personnel capability and experience

Curve #3 in Figure 8 illustrates the effect of adding six people at about Day 200, all of whom with productivities equal to the average rate recorded during the first 200 days of the project. It shows a small drop in productivity once new personnel is added, picking up during the assimilation time, and finally achieving overall higher team productivity, completing the project on Day 434. In the next simulation run (curve #4), the effects of personnel capability and experience were included, showing that when new manpower was added to the project, the production rate almost did not drop, and instead, increased progressively over the assimilation period, whilst training overhead was gradually eliminated as new people become trained and assimilated. The project completes after approximately 406 days, which is considerably shorter than the 439 days it would take without consideration of the personnel factors of new manpower. Although it still failed to bring the project back to schedule, it completed much sooner than if the project had been continued without adding new manpower. But the question is whether or not the forecast progress and completion day would approximately represent the actual situation. This is discussed in the next section.

5.3. Analysis of Project Progress: Simulation against Actual

In this section we attempt to evaluate how accurately the proposed simulation model would represent an actual situation. Based on the values given by the Project Office, the actual progress of the project is plotted in Figures 9 and 10.

Figure 9 shows the cumulative work progress during the project lifecycle. Curve #1 shows the progress as predicted by the proposed simulation model, while curve #2 is the reported actual progress of the project. Although the simulated progress doesn't exactly mimic the real situation, it is a close approximation. Interestingly, as the figure shows, the actual project had completed only slightly sooner than predicted by the simulation.

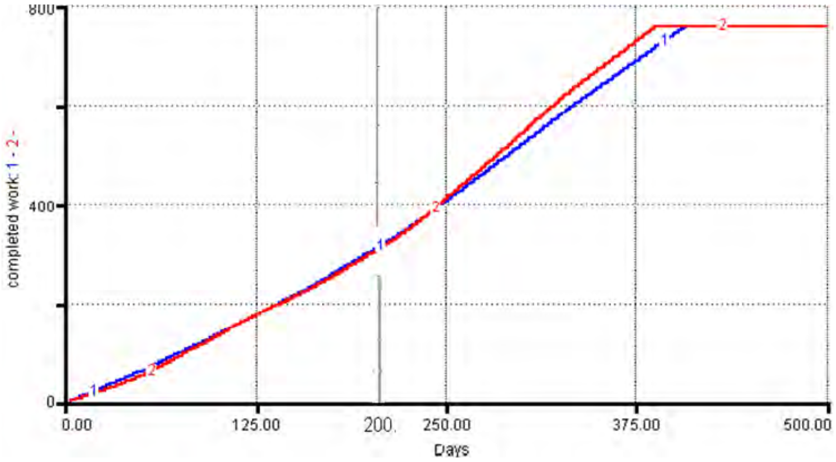


Fig. 9. Project Progress: Curve 1 - simulation experiment, curve 2 - actual situation

Figure 10 shows the difference in production rate between the simulation results and the actual situation. Curve #1 shows the production rate and completion date of the project as predicted by the simulation, while Curve #2 shows the actual production rate as recorded monthly. The simulation forecast that the project would be completed at around Day 406. This was based on the average productivity rate from project commencement to Day 200, and took into account the addition of six people with predicted personnel capabilities. In the actual situation (curve #2), the software team's productivity rate follows almost the same pattern, but with some fluctuation, as that predicted by the simulation. From around Day 220, curve #2 gradually surpasses curve #1 and continues to maintain its productivity difference from curve #1, eventually leading to the project completing sooner at around Day 387.

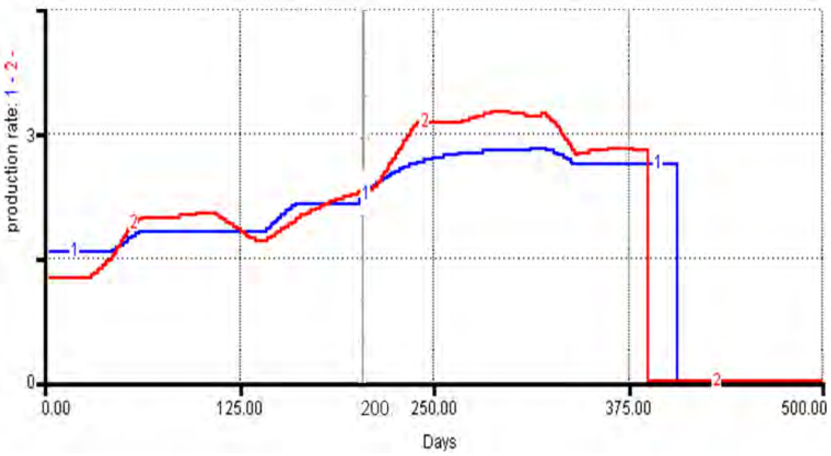


Fig. 10. Software team productivity: Simulation experiment against real situation. Curve 1 - simulation experiment, curve 2 - actual situation

To investigate how accurately the simulation result predicts an actual situation, we measured the percentage error using the absolute differences between the simulated and experimented values. As the project completion day was forecast at Day 200, the baseline for calculation was set at Day 200. Therefore, we have:

$$\text{Error \%} = \frac{|\,\text{Experimental} - \text{Theoretical}|}{\text{Theoretical}} \times 100 \qquad (\mathbf{2})$$

$$= \frac{|\,(387-200) - (406-200)|}{406-200} \times 100 = \frac{19}{206} \times 100 = 9.23\ \%$$

The proposed System Dynamics model with consideration of personnel factors demonstrated 90.77% accuracy in the investigated case study. The result gives us confidence that the proposed model is capable of approximately simulating an actual situation, and can be employed for decision trade-off analysis by software project managers.

5.4. Comparison of Experiment Results

In earlier sections, the simulation results of different scenarios were presented and discussed. Here, we summarise the results gained from the simulation experiments and compare them with a case study. Table 8 shows the project's scheduled completion days in each scenario, and compares the predicted completion days in the different scenarios with the actual completion days.

Table 8. Comparison of scheduled completion days in a real-life case study

Scenario	Project completion days	Differences in schedule compared with actual situation	Prediction accuracy in comparison with actual situation
(1) Plan based on expected productivity	375	-12 days	Not applicable
(2) Estimation based on actual productivity	486	+99 days	Not applicable
(3) Estimation based on actual productivity and new manpower	434	+47 days	79.91%
(4) Estimation based on actual productivity, new manpower, personnel factors, relative assimilation time and training overhead	406	+19 days	90.23%
(5)Actual situation	387	0	100%

The above comparisons show that when personnel factors are taken into account, the prediction achieves better accuracy. Furthermore, it should be noted that although adding people did not help bring the project back on schedule, it was effective in preventing further delays. The above findings affirm our confidence that the proposed simulation model can approximately represent an actual outcome. The results indicate the importance of personnel capability and experience to productivity and, consequently, to an improvement in project schedule.

6. Trade-off Analysis

As previously mentioned, an aim of this study is to enable project managers foresee the possible consequences of their decisions and to perform various trade-off analyses. In experiments for the first case study (section 4.1), it was observed (Figure 6) that too many people, despite having very high levels of capability and experience, would not be beneficial to a project and would be a waste of time and resources. In this section, we will explore how the proposed simulation model can help project managers find the optimal number of personnel to add to a project at a given time, and examine at which point on the time scale would adding more manpower be a clearly wrong decision.

In an actual situation, project managers have to trade-off the number of personnel that can be added to a project with the additional cost it would incur. A project manager can take advantage of the proposed simulation model to perform various sensitivity analyses using simulation experiments to find the best possible solution for his project. These analyses can factor in a range of different variables, such as the number of team members, the levels of capability and experience of the manpower, and the time at which new manpower could be added to the project. Due to space constraints, it is not possible to show the result of each sensitivity analysis on paper of this size. However, a summary of the simulation results using three different numbers of additional manpower (3, 6 and 9 new personnel) at four different points on the time scale (Days 200, 250, 300 and 350) in an industry case, as explained in section 4.2, is provided in Table 9.

Based on the above settings, the results (Table 9) show that adding more new people leads to a bigger reduction in schedule up to a certain point. However, after this point, as the number of additional manpower increases, the effect on reducing the schedule decreases. For instance, the difference between adding 3 and 6 new personnel on Day 200 is 26 days, but the difference is less than 6 days between adding 6 and 9 new team members.

Table 9. Trade-off analysis of the effect of adding different numbers of new personnel at four different points on the time scale.

Day	Item	3 New Personnel	6 New Personnel	9 New Personnel
200	Completion date (day)	432	406	400
	Cumulative added effort (day)	858	1716	2574
	Cost of adding manpower	128700	257400	386100
	Added value	216000	320000	344000
	Trade-off Cost	87300	62600	-42100
250	Completion date (day)	441	420	415
	Cumulative added effort (day)	708	1416	2124
	Cost of adding manpower	106200	212400	318600
	Added value	180000	264000	284000
	Trade-off Cost	73800	51600	-34600
300	Completion date (day)	450	433	429
	Cumulative added effort (day)	558	1116	1674
	Cost of adding manpower	83700	167400	251100
	Added value	144000	212000	228000
	Trade-off Cost	60300	44600	-23100
350	Completion date (day)	460	448	444
	Cumulative added effort (day)	408	816	1224
	Cost of adding manpower	61200	122400	183600
	Added value	104000	152000	168000
	Trade-off Cost	42800	29600	-15600

In a real-life situation, a project manager needs to work out the best point on a time scale at which to add new people and the ideal number of new personnel to add for his project. The trade-off between the cost of adding new manpower and the added value they bring to the project (as the result of a reduction in project duration) would answer the above questions. For example, if we determine that a delay in project completion would cost $4,000 per day, and the cost of adding new manpower is $150 per day per person, as shown in Table 9, then the trade-off of adding 9 new personnel is negative in all cases. The figures show that the earlier the new staff is added, the higher the trade-off value obtained. In this case, adding 3 new team members would be the best fit for the project in terms of trade-off cost at each of the four points on the time scale.

7. Model Evaluation

It is important to note that a model is seldom intended to be a complete and accurate representation of a real system. Usually, the aim of evaluating a System Dynamics model is not to validate its accuracy; instead, it's about making sure that the model provides greater insight into and better understanding of a phenomenon. Given the above, the verification and validation of this study are focused on building confidence in the model as a reasonable representation of the system and in its usefulness in producing results. Hence, the assessment of the proposed system dynamics simulation model has been performed mainly in the following stages: model verification, and model validation.

During model verification, we examined whether its implementation is error-free and properly represents the intended logical behaviour. For this purpose, the model components were first reviewed and analysed during its incremental development stage. Second, the model was verified by comparing its simulated behaviour with that of the base model.

In model validation, we examined whether the model addresses the problem defined in the current study. Two experiments, one with values reported from the case study, and the other one with data taken from a real-life situation, were chosen to perform the examination.

Comparing the simulation experiments with the results obtained from a case study reported in literature assures that the model has behavioural consistencies with the previously reported results. Furthermore, validating the results of the simulation against those obtained from the actual reported progress of a project in a real-life environment strengthened our confidence that the model can be effectively employed by project managers.

8. Conclusion

In order to fulfil the objectives of the study, a System Dynamics model was designed and constructed. Experiments were performed to verify and validate the model, and to explore the effects of adding new manpower with different personnel capabilities, skills and experiences to delayed software projects.

Experiments with two case studies demonstrated that the negative effects of adding new manpower to the progress of a project can be considerably minimised when new manpower with certain levels of personnel factors is added. The results obtained show a significant improvement in project schedule for both the case studies.

In the first case study, which was taken from literature, the objectives of doing the experiments were, firstly, to verify the accuracy of the model and, secondly, to have a basis of comparison for the different experiment's results. The outcome of performing the experiments with base work simulation settings assures that the model correctly replicates the previously reported results.

In the second case study, which was taken from a real-life project, the aims of performing the experiments were, firstly, to identify the limitations of the model when applied to a real-life environment and, secondly, to assess the accuracy of the forecast results against the actual situation. The results of the experiment were most promising, demonstrating that the model is capable of approximately forecasting actual progress and the actual completion date of a project.

The proposed System Dynamics model can be employed for trade-off analysis of a software project and to forecast the impact of different decisions on changing manpower in a project. It has significant benefits for the software project managers, namely providing better clarification on the status of a project, forecasting the optimum staff level required to meet a deadline, and reducing the risk of decision making. The above advantages help project managers become less doubtful and more confident about their decisions as the simulation analysis reveals the likely consequences of their decisions.

The research and experiment in this study focused on the impact of personnel skills, capability and experience, thus one of the good future extensions of this work could be the study of how psychological and social characteristics of personnel could contribute to faster fitting to a software production working team.

References

[1] It-cortex. Statistics on IT Projects Failure Rates. October 2009, from http://www.it-cortex.com/Stat_Failure_Rate.htm .2002

[2] Standish-Group. New Standish Group Report Shows More Projects Failing and Less Successful Projects. Retrieved October 2009, from http://www1.standishgroup.com/newsroom/chaos_2009.php. 2009

[3] Abdel-Hamid, T. and S. E. Madnick. Software Project Dynamics: An Integrated Approach. Upper Saddle River, Prentice-Hall.(1991)

[4] Callegari, D. and R. Bastos. A Systematic Review of Dynamic Reconfiguration of Software Projects. Information technology and management 1(3): 103-113.(2008)

[5] Eden, C., T. Williams and F. Ackermann. Dismantling the Learning Curve: The Role of Disruptions on the Planning of Development Projects. International Journal of Project Management 16(3): 131-138.(1998)

[6] Ford, D. and J. Sterman. Dynamic Modeling of Product Development Processes. System Dynamics Review 14(1): 31-68.(1998)

[7] Lyneis, J. and D. Ford. System Dynamics Applied to Project Management: A Survey, Assessment, and Directions for Future Research. System Dynamics Review 23(2-3): 157-189.(2007)

[8] Brooks, F. P. The Mythical Man-Month (Anniversary Ed.), Addison-Wesley Longman Publishing Co., Inc.(1995)

[9] Stutzke, R. A Mathematical Expression of Brooks' Law. 9th International Forum on COCOMO and Cost Modeling, Los Angeles, CA, 1-24.(1994)

[10] Hsia, P., C. Hsu and D. Kung. Brooks' Law Revisited: A System Dynamics Approach. Proceedings of 23rd Annual IEEE Computer Software and Applications Conference, COMPSAC, Phoenix, USA, IEEE Computer Society: 370-376.(1999)

[11] Williams, L., Shukla A., and Anton A.I. An Initial Exploration of the Relationship between Pair Programming and Brooks' Law. Proceedings of the Agile Development Conference, IEEE Computer Society.(2004)

[12] Madachy, R. Software Process Dynamics. Chichester, Wiley-IEEE Press.(2008)

[13] Schweik, C. M., R. C. English, M. Kitsing and S. Haire. Brooks' Versus Linus' Law: An Empirical Test of Open Source Projects. Proceedings of the international conference on Digital government research, Montreal, Canada, Digital Government Society of North America.(2008)

[14] Adams, P., A. Capiluppi and C. Boldyreff. Coordination and Productivity Issues in Free Software: The Role of Brooks' Law. Proceedings of ICSM 2009, Edmonton, Canada: 319-328.(2009)

[15] Capiluppi, A. and P. J. Adams Reassessing Brooks' Law for the Free Software Community. In. Open Source Ecosystems: Diverse Communities Interacting, Springer Boston. 299: 274-283.(2009)

[16] Neus, A. and P. Scherf. Opening Minds: Cultural Change with the Introduction of Open-Source Collaboration Methods. IBM Systems Journal 44(2): 215-225.(2010)

[17] Samoladas, I., L. Angelis and I. Stamelos. Survival Analysis on the Duration of Open Source Projects. Information and Software Technology 52(9): 902-922.(2010)

[18] Pressman, R. Software Engineering: A Practioner's Approach Sixth Edition. New York, McGraw-Hill.(2006)

[19] Finnie, G., G. Wittig and D. Petkov. Prioritizing Software Development Productivity Factors Using the Analytic Hierarchy Process. Journal of Systems and Software 22(2): 129-139.(1993)

[20] Boehm, B. W., Clark, Horowitz, Brown, Reifer, Chulani, R. Madachy and B. Steece. Software Cost Estimation with Cocomo Ii with Cdrom. Upper Saddle River, Prentice Hall.(2000)

[21] Trendowicz, A. and J. Münch. Factors Influencing Software Development Productivity-State-of-the-Art and Industrial Experiences. Advances in Computers 77: 185-241.(2009)

[22] Hannay, J. E., E. Arisholm, H. Engvik and D. I. K. Sjoberg. Effects of Personality on Pair Programming. IEEE Transactions on Software Engineering 36(1): 61-80.(2010)

[23] Putnam, L. A General Empirical Solution to the Macro Software Sizing and Estimating Problem. IEEE Trans. Software Eng. 4(4): 345-361.(1978)

[24] McCain, K. and L. Salvucci. How Influential Is Brooks' Law? A Longitudinal Citation Context Analysis of Frederick Brooks' the Mythical Man-Month. Journal of Information Science 32(3): 277.(2006)

[25] Verner, J., S. Overmyer and K. McCain. In the 25 Years since the Mythical Man-Month What Have We Learned About Project Management? Information and Software Technology 41(14): 1021-1026.(1999)

[26] Abdel-Hamid, T. The Dynamics of Software Project Staffing: A System Dynamics Based Simulation Approach. IEEE Transactions on Software Engineering 15(2): 109-119.(1989)

[27] Opelt, K. Overcoming Brooks' Law. In. Agile Processes in Software Engineering and Extreme Programming. Berlin Heidelberg, Springer-Verlag: 175-178.(2007)

[28] Wu, M. and H. Yan. Simulation in Software Engineering with System Dynamics: A Case Study. Journal of Software 4(10): 1127.(2009)

[29] Arisholm, E., H. Gallis, T. Dyba and D. I. K. Sjoberg. Evaluating Pair Programming with Respect to System Complexity and Programmer Expertise. IEEE Transactions on Software Engineering 33(2): 65-86.(2007)

[30] Nosek, J. T. The Case for Collaborative Programming. Communication of ACM 41(3): 105-108.(1998)

[31] Salleh, N., E. Mendes, J. Grundy and G. Burch. An Empirical Study of the Effects of Conscientiousness in Pair Programming Using the Five-Factor Personality Model. Proceedings of ICSE '10 The 32nd ACM/IEEE International Conference on Software Engineering, Cape Town, South Africa, ACM: 577-586.(2010)

[32] Ma, Z. Intelligent System Dynamics Modeling and Decision Analysis for Software Schedule Slippage Recovery. Doctoral Dissertation. Arizona State University. (2000)

[33] Maxwell, K., L. Van Wassenhove and S. Dutta. Software Development Productivity of European Space, Military, and Industrial Applications. IEEE Transactions on Software Engineering 22(10): 706-718.(1996)

[34] Capretz, L. F. and F. Ahmed. Why Do We Need Personality Diversity in Software Engineering? ACM SIGSOFT Software Engineering Notes 35(2): 1-11.(2010)

[35] Ricca, F., M. Di Penta, M. Torchiano, P. Tonella and M. Ceccato. How Developers' Experience and Ability Influence Web Application Comprehension Tasks Supported by Uml Stereotypes: A Series of Four Experiments. IEEE Transactions on Software Engineering 36(1): 96-118.(2010)

[36] Chulani, S., B. Boehm and B. Steece. Bayesian Analysis of Empirical Software Engineering Cost Models. IEEE Transactions on Software Engineering 25(4): 573-583.(1999)

[37] Madachy, R. A Software Project Dynamics Model for Process Cost, Schedule and Risk Assessment. Doctoral Dissertation. University of Southern California. 1994)

Masrah Azrifah Azmi Murad received her PhD in Artificial Intelligence from the University of Bristol, UK in 2005. She is currently an associate professor at the Department of Information Systems, Universiti Putra Malaysia. She is a member of IEEE and IEEE Computer Society. Her current research interests include text mining, applied informatics, and automated software engineering.

Mostafa Farshchi recived his Master of Science in Software Engineering from Universiti Putra Malaysia (UPM) in 2011. His research interest includes software engineering, system dynamics, and agent technology. He has more than 6 years experience on enterprise software development and works as a senior software engineer in enterprise software projects.

Yusmadi Yah Jusoh received the B.Econs. and M.IT. degrees from Universiti Kebangsaan Malaysia (UKM) in 1996 and 1998, respectively. She received the PhD from the same university in 2008. She is a lecturer at Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM) since 1998. Her research interest includes information systems, information technology strategic planning, software engineering and management information system.

Received: May 25, 2011; Accepted: March 08, 2012.

DOI: 10.2298/CSIS111201013P

A New Strategic Tool for Internal Audit of the Company Based on Fuzzy Logic

Aleksandar Pešić[1], Duška Pešić[2], and Andreja Tepavčević[3]

[1] Business and Industrial Management, Union University,
Andre Nikolića 29, 11000 Belgrade, Serbia
pesic@mpk.edu.rs
[2] Information Technology School, ComTrade Technology Centre,
Savski nasip 7, Novi Beograd, Serbia
duska.pesic@its.edu.rs
[3] Faculty of Science, University of Novi Sad,
Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia,
andreja@dmi.uns.ac.rs

Abstract. Since the internal audit of the company is essential to supply information needed for the effective management and improvement of the competitive position, the purpose of this paper is to introduce an innovative strategic management tool for the assessment of internal organizational factors that overcomes some limitations of traditional appraisal methods, and also enables more comprehensive evaluation of the company's internal environment. Although the classical Internal Factor Evaluation matrix (IFE matrix) is widely used, it has some constraints, such are lack of considering the ambiguity and vagueness of the internal factors. An original method – FSIF (Fuzzy Synthesis of Internal Factors) represents a systematic approach that incorporates fuzzy logic in order to better describe real situation. Proposed FSIF method well serves the needs of modern Management information system because it provides monitoring of internal development of an organization through time and also comparing different organizations taking into account various factors and weights.

Keywords: Internal factors, fuzzy sets, IFE, FSIF.

1. Introduction

Effective performance of any organization in a modern competitive environment significantly depends on a detailed analysis of internal factors. This is a complex and very important activity for management, since the information obtained in this way might give insight into the current state of resources and overall capabilities of the organization. It represents the basis of any effective strategy and a way of understanding organizational capabilities. The main step in conducting function-by-function analysis is to construct a relevant strategic tool for assessment of major internal factors in the functional areas of business (e.g. finance, marketing, production/services, human recourse management, research and development, information systems).

In connection to this, a strategy for an organization has to be built from what that organization is. “Neglect of this basic step can result in strategies which appear well founded in relation to the market and the assessments of the future environment, but which cannot be implemented because of ill-founded implicit assumptions about organization itself” (Jenster, Hussey 2001, p. 11 [7]).

Since the internal factors analysis is a vital requirement of strategic thinking, the question of choice of the appropriate method is inevitable. As a qualitative assessment of internal factors cannot give a satisfactory degree of precision, management of the organization should pay a special attention to methods that provide a comparative, quantitative measurement; such that the objects of assessment that are expressed in the form of attributes can be transformed into a quantitative form, expressed as a quantitative variable.

This article is organized as follows. In Section 2, related work is elaborated. In Section 3 original FSIF model is presented. In Section 4 a case illustration is given on which FSIF method and IFE matrix method are compared and discussed and in Section 5 a conclusion is given.

2. Related Work

There are various methods for organizational appraisal aimed to ease some of the subjectivity by introducing weights and ratings into the model. One of the most popular and useful strategic management tool for auditing key internal factors in functional areas of a business is the Internal Factor Evaluation matrix (IFE matrix) proposed by Fred R. David [2,3].

IFE method summarizes internal factors (strengths and weaknesses of the organization) and also provides a basis for identifying and evaluating relationships among the functional areas of a business.

According to David [2,3], IFE matrix answers four major questions about organization:

- What are the organization’s key strengths and weaknesses?
- What is the relative importance of each strength and weakness to the organization’s overall performance?
- Does each factor represent: a major weakness, a minor weakness, a minor strength, a major strength?
- What is organization’s total weighted score resulting from the analysis of the IFE?

IFE matrix can be developed in five steps:

1. Major strengths and weaknesses are listed (all factors should be stated objectively and their number should be from 10 to 20).
2. A weight that indicates relative importance of the factor is assigned (weight should ranges from 0.0 – not important, to 1.0 – all important, and the sum of all weights must equal 1.0).
3. A 1 to 4 rating to each factor is assigned (a major weakness (rating = 1), a minor weakness (rating = 2), a minor strength (rating = 3), a major strength (rating = 4)).
4. Each factor's weight is multiplied by its rating.
5. The weighted scores to determine total weighted score for the organization is calculated (total weighted score well below 2.5 indicate internally weak position of the organization, and scores significantly above 2.5 indicate a strong internal position).

It is important to notice that a factor could present both: a strength and a weakness. In that case, the factor should be included twice in the IFE matrix.

Similar to the IFE matrix is IFAS (Internal Factor Analysis Summary) approach introduced by Wheelen and Hunger [12]. As they pointed out "IFAS table is one way to organize the internal factors into generally accepted categories of strengths and weaknesses as well as to analyze how well a particular company's management is responding to these specific factors in light of the perceived importance of these factors to the company" (Wheelen and Hunger, 2007, p. 129 [12]).

Still, there are limitations of these methods. Assigned four degrees rating to each factor is not always appropriate to real situation (a finer rating is often necessary). Besides, vagueness of the factors is not considered. In addition, the total number of internal factors is limited to 10-20.

In this paper, a new method based on fuzzy logic, for analysis of corporate internal factors is developed to overcome these limitations. Proposed FSIF method illustrates some of the possibilities of fuzzy logic through a simple and easy understandable model that does not require advanced mathematical knowledge.

Although fuzzy set theory and fuzzy logic, as useful tools to grasp uncertainty, have already been exploited in several papers in solving various problems of strategic management ([4], [5], [8], [9], [13]), in this paper we propose original approach to the internal scanning process that can be easily implemented and is adjustable by the user.

The main contribution is that FSIF method enables managers to be better informed about every important aspect of the company's internal environment which leads to the increased effectiveness of the decision making process.

3. Fuzzy synthesis of internal factors - model FSIF

The first step in applying the proposed FSIF method is, similarly as in the classical IFE method, determination of all the relevant internal factors of the organization.

Identification of the most important factors is based on the Company audit questionnaire proposed by Jenster and Hussey [7]. Their questionnaire covers six key functional areas of the organisation, and that is reason why the questions in the questionnaire are listed in the six broad categories: finance, marketing, production, human recourse management and management effectiveness, research and development, and information systems.

In fuzzy model for the internal assessment, questions from the Company audit questionnaire are used as a basis of which we have derived the same number of factors. It stems that derived factors serves only to provide relevant basic inputs for proposed fuzzy appraisal method which enables completely different process of assessment.

It is valuable to notice that importance of the factors can immensely vary between different types of industry, and different sizes of organization within each industry, so it is recommended that management of the organization adjust the list of factors to fit their business.

An advantage of the FSIF method for assessment of the internal factors (in comparison with classical IFE and IFAS methods) is that the number of factors is not restricted, so that managers can use factors according to the situation of their business and evaluate those factors that have strategic influence to the company.

In FSIF method, we propose the use of interval [0,5] in order to keep a certain similarity with the existing techniques of assessment of the organization (IFE), in which it is possible to choose one of four options: 1, 2, 3 or 4. Moreover, the proposed fuzzy model does not limit the total number of internal factors [which is 10-20 in the IFE method], because it does not request the additivity condition.

Let $F_i, i=1,...,n$ be factors that are determined as relevant factors for an organization *O*. Their estimated value is $v_i, i=1,...,n$. This value belongs to interval [0,5]. Four fuzzy sets[1] are defined:

$\overline{MJW}$ – „Major weakness“,

[1] A fuzzy set is a mapping from a set to [0,1] real interval. A function value represents degree of membership of an element to the fuzzy set.

$\overline{MNW}$ – „Minor weakness“,
$\overline{MNS}$ – „Minor strength“ or
$\overline{MJS}$ – „Major strength“.

Depending on the values of factors, the degree of membership of every factor to each of the four fuzzy sets is calculated.

Graphic presentations of the defined fuzzy sets are given in Figure 1. As in the standard method, we can notice the fact that a factor may be at the same time a factor of minor weakness and a factor of minor strength (in different degrees). Factor F_i in Fig. 1 belongs to the set $\overline{MJW}$ with the degree of membership μ_{1i}, while it belongs to set $\overline{MNW}$ with the degree μ_{2i}.

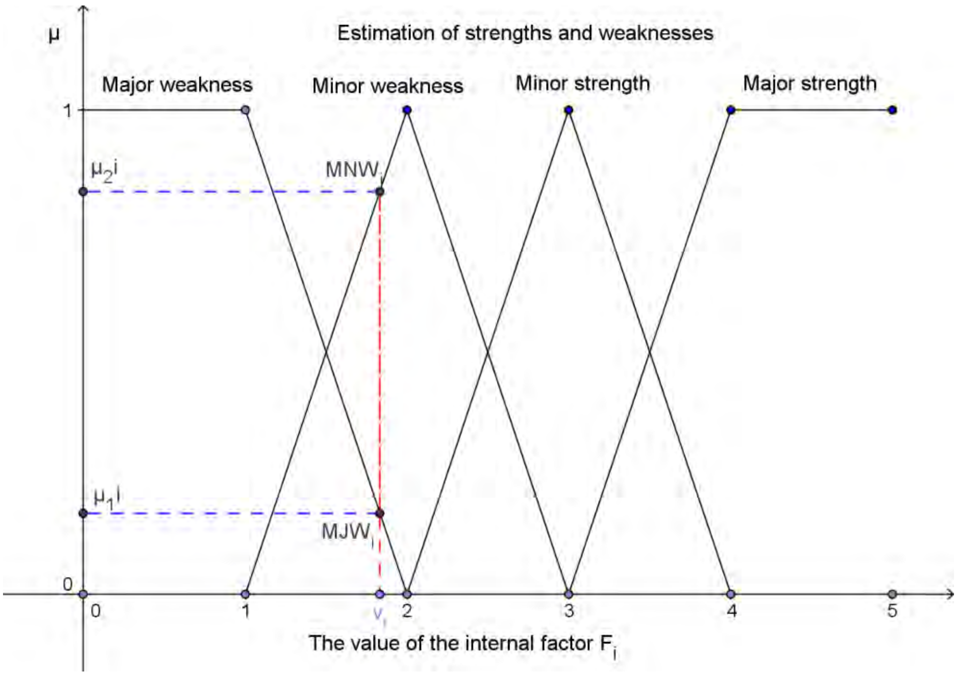


Fig. 1. Estimation of strengths and weaknesses of an organization

Theorem 1. The sum of all the obtained degrees of membership for every factor must be equal to 1. For every factor F_i

$$\sum_{k=1}^{4} \mu_{ki} = 1 \qquad (1)$$

.

Proof. Since we use the triangular fuzzy sets, for a factor F_i with a value *x* membership function μ_{1i} is equal to 1 for x belonging to the interval [0,1]. Further, μ_{1i}(*x*) = -*x*+2 is valid for *x* in interval [1,2], and it is equal to 0 in interval [2,4]. Similarly, μ_{2i}(*x*) = *x*-1 is valid for *x* in interval [1,2], and also μ_{2i}(*x*) = -*x*+3 for x in interval [2,3]. It is equal to 0 for *x* outside interval [1,3]. Similarly we obtain the formulas for other two membership functions.

If the value of the factor is from 0 to 1, the degree of membership to fuzzy set $\overline{MJW}$ is 1, and to all the other fuzzy sets it is 0. If the value belongs to the interval [1,2], then by the symmetry of graphs of functions representing fuzzy sets $\overline{MNW}$ and $\overline{MJW}$ with respect to the line μ=0.5, the degree of membership to fuzzy set $\overline{MJW}$ is equal to 1 minus the degree of membership to fuzzy set $\overline{MNW}$, and the degree related to the remaining two fuzzy sets is equal to 0. We can prove this fact also using the formulas above for fuzzy sets representing the $\overline{MJW}$ andand $\overline{MJW}$. In example, for a factor F_i with the value x belonging to interval [1,2], we obtain:

$$\mu_{1i}(x) + \mu_{2i}(x) = (-x+2)+(x-1) = 1.$$

Analogously, we notice that in the interval [2,3], only two degrees are different from 0 ($\overline{MNW}$ and $\overline{MNS}$) and their sum is equal to 1. In the interval [3,4], the graphs of functions $\overline{MNS}$ and $\overline{MJS}$ are symmetric with respect to the line μ=0.5, therefore sum of related two degrees is equal to 1 and remaining two degrees are equal to 0. We can prove the later two statement also using the formulas for triangular fuzzy sets. Finally, in the interval [4,5] only $\overline{MJS}$ degree is equal to 1 and all others to 0. Therefore over the whole domain, the sum of the degrees is 1. ■

The same procedure is applied to all the relevant factors $F_i, i = 1,...,n$. Each of the factors belongs with a certain degree of membership to $\overline{MJW}$, $\overline{MNW}$, $\overline{MNS}$ and $\overline{MJS}$.

The importance of relevant factors is also taken into account. Here we propose weight to be taken from [0,2] interval, obtaining the ponder that increase or decrease the influence of some of the factors on the overall level of state of the internal environment of the organization. The influence of a factor is decreased if the weight is from the interval [0,1), and it is increased if the weight is from the interval (1,2]. If the weight is 1, it has no impact on the value of the factor. The obtained data are presented in Table 1.

Table 1. Degrees of membership of internal factors to fuzzy sets

Internal factor	$\overline{MJW}$	$\overline{MNW}$	$\overline{MNS}$	$\overline{MJS}$	Weight
F_1	μ_{11}	μ_{21}	μ_{31}	μ_{41}	α_1
F_2	μ_{12}	μ_{22}	μ_{32}	μ_{42}	α_2
F_3	μ_{13}	μ_{23}	μ_{33}	μ_{43}	α_3
…	…	…	…	…	…
F_n	μ_{1n}	μ_{2n}	μ_{3n}	μ_{4n}	α_n

Multiplying the corresponding degrees of membership of factors F_i to the observed fuzzy sets with the weights of these factors, we obtain the weighted fuzzy membership degrees (Table 2).

Table 2. Weighted fuzzy membership degree

Internal factor	$\overline{MJW}$	$\overline{MNW}$	$\overline{MNS}$	$\overline{MJS}$
F_1	$\alpha_1\mu_{11}$	$\alpha_1\mu_{21}$	$\alpha_1\mu_{31}$	$\alpha_1\mu_{41}$
F_2	$\alpha_2\mu_{12}$	$\alpha_2\mu_{22}$	$\alpha_2\mu_{32}$	$\alpha_2\mu_{42}$
F_3	$\alpha_3\mu_{13}$	$\alpha_3\mu_{23}$	$\alpha_3\mu_{33}$	$\alpha_3\mu_{43}$
…	…	…	…	…
F_n	$\alpha_n\mu_{1n}$	$\alpha_n\mu_{2n}$	$\alpha_n\mu_{3n}$	$\alpha_n\mu_{4n}$

In order to determine the state of the internal environment of the organization as a whole, we find the weighted arithmetic mean of the degrees of membership of factors to fuzzy sets $\overline{MJW}$, $\overline{MNW}$, $\overline{MNS}$ and $\overline{MJS}$. For the first two sets, the degree is taken with a negative sign [because it is the weaknesses of the organization] and for the other two fuzzy sets with a positive sign [because it is the strength of the organization]. Moreover, the degrees of membership to $\overline{MJW}$ and $\overline{MJS}$ are multiplied with 1.5, increasing their influence in the total sum, because it shows a great weakness and a great strength (Table 3). The factor 1.5 is an empirical value that is chosen to emphasize the major weakness and strength of the factors.

Summing up the values in the last row, we obtain a number I^O that indicates the current state of the internal environment of the organization O:

$$I^O = -(1.5)\sum_{i=1}^{n}\frac{\alpha_i\mu_{1i}}{n} - \sum_{i=1}^{n}\frac{\alpha_i\mu_{2i}}{n} + \sum_{i=1}^{n}\frac{\alpha_i\mu_{3i}}{n} + (1.5)\sum_{i=1}^{n}\frac{\alpha_i\mu_{4i}}{n} \qquad (2)$$

Value I^0 is a real number from the interval [-3,3]. If the resulting number is positive, the organization has more strengths than weaknesses, and as the number is greater, the state of the organization is better. If the resulting number is negative, the organization has more weakness than strengths, and less number means that the condition of the organization is worse.

Table 3. Weighted arithmetic mean of the degrees of membership

Internal factor	$\overline{MJW}$	$\overline{MNW}$	$\overline{MNS}$	$\overline{MJS}$
F_1	$\alpha_1\mu_{11}$	$\alpha_1\mu_{21}$	$\alpha_1\mu_{31}$	$\alpha_1\mu_{41}$
F_2	$\alpha_2\mu_{12}$	$\alpha_2\mu_{22}$	$\alpha_2\mu_{32}$	$\alpha_2\mu_{42}$
F_3	$\alpha_3\mu_{13}$	$\alpha_3\mu_{23}$	$\alpha_3\mu_{33}$	$\alpha_3\mu_{43}$
…	…	…	…	…
F_n	$\alpha_n\mu_{1n}$	$\alpha_n\mu_{2n}$	$\alpha_n\mu_{3n}$	$\alpha_n\mu_{4n}$
	$-(1.5)\sum_{i=1}^{n}\frac{\alpha_i\mu_{1i}}{n}$	$-\sum_{i=1}^{n}\frac{\alpha_i\mu_{2i}}{n}$	$\sum_{i=1}^{n}\frac{\alpha_i\mu_{3i}}{n}$	$(1.5)\sum_{i=1}^{n}\frac{\alpha_i\mu_{4i}}{n}$

4. A case illustration and discussion

In an observed organization, management have used the list of 29 factors proposed by Jenster and Hussey, but only 15 factors were extracted as relevant for the condition of their specific internal environment. In this phase, experiences and competencies of managers were vital for selection of the most important factors.

For the fuzzy assessment of the internal factors specially designed questionnaires (based on the company audit questionnaire, by Jenster and Hussey) were used.

Managers could use examples in the questionnaire as an aid in assessing the questions. It is important to notice that these are only sample responses, i.e. examples derived from various organizations and therefore, they may not be applicable to every organization. Part of the questionnaire for the fuzzy analysis used in this study is shown in Figure 2.

Questionnaire for the fuzzy analysis of the internal organizational factors

– Method FSIF–

Marketing

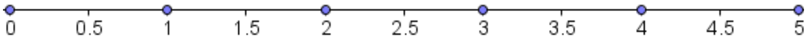
Factor: Marketing budget

In light of your assessment, how do you evaluate its competitive impact on the organisation?

Significant competitive disadvantage

Neither disadvantage, nor advantage

Significant competitive advantage



Examples:

- **0 - 1.5**
  We don’t have separate budget for marketing efforts.
- **1.5 - 2.5**
  When production costs are covered and depreciation is determined, we decide how much marketing we can afford.
- **2.5 - 3.5**
  For old products we decide a fixed percentage for marketing. For new products, we have separate budget.
- **3.5 - 5**
  The budget is determined by relative profit margins of products, position in the product life cycle, competitive activity and market potential for each product.

Please, rate the importance of this internal factor to your organisation, from 0 - „low importance“ to 2 - „high importance“

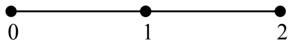


Fig. 2. Questionnaire for the fuzzy analysis

Applying these questionnaires, the following information on the factors identified by the management of the organization were obtained (Table 4).

In Table 5, IFE matrix for the observed case is presented.

Table 4. Values of internal factors, membership degrees and weights of the factors in application of FSIF model

Internal factor	Value	$\overline{MJW}$	$\overline{MNW}$	$\overline{MSN}$	$\overline{MNS}$	Weight
Financial resources	2.84	0	0.16	0.84	0	0.5
Long-range financial planning	2.37	0	0.63	0.37	0	1.3
Accounting system	1.83	0.17	0.83	0	0	0.4
Quality of products/services	4.2	0	0	0	1	1.7
Corporate reputation	3.23	0	0	0.77	0.23	1.1
Pricing policy	3	0	0	1	0	1.9
Systematic market analysis	2.86	0	0	0.86	0.14	2
Distribution channels	1.58	0.42	0.58	0	0	2
Marketing budget	2.16	0	0.84	0.16	0	0.5
Outsourcing prospective	0.5	1	0	0	0	0.3
Material work flow	1.51	0.49	0.51	0	0	0.8
Internal communication/Information systems	3.33	0	0	0.67	0.33	1
HRM procedures	3.59	0	0	0.41	0.59	0.9
Management effectiveness	3.72	0	0	0.28	0.72	1.8
Organization’s R&D activity	3.87	0	0	0.13	0.87	1

Table 5. IFE matrix

Internal factor	Rating	Weight	Weighted score
Financial resources	3	0.03	0.09
Long-range financial planning	2	0.08	0.16
Accounting system	2	0.02	0.04
Quality of products/services	4	0.10	0.40
Corporate reputation	3	0.06	0.18
Pricing policy	3	0.11	0.33
Systematic market analysis	3	0.12	0.36
Distribution channels	2	0.12	0.24
Marketing budget	2	0.03	0.06
Outsourcing prospective	1	0.02	0.02
Material work flow	2	0.04	0.08
Internal communication/Information systems	3	0.06	0.18
HRM procedures	4	0.05	0.20
Management effectiveness	4	0.10	0.40
Organization’s R&D activity	4	0.06	0.24
	Total:	1	2.98

In this case illustration, we can notice that use of [0,5] interval enable better assessment of factors and thus also a better evaluation of the degree of their weakness and strength, than the use of four integers 1, 2, 3 or 4. By the interval approximation, we obtain additional information about the degree of membership of factors to each of the sets „Major weakness“, „Minor weakness“, „Major strength“ „Minor strength“ while in IFE matrix only a strict membership to sets is determined (belonging in classical sense: it belong to a set or it does not belong to a set).

Applying the weights in FSIF model, the data in Table 6 have been obtained.

Table 6. Weighted fuzzy membership degree

Value of Internal factor	$\overline{MJW}$	$\overline{MNW}$	$\overline{MSN}$	$\overline{MNS}$
2.84	0	0.08	0.42	0
2.37	0	0.819	0.481	0
1.83	0.068	0.332	0	0
4.2	0	0	0	1.7
3.23	0	0	0.847	0.253
3	0			0
2.86	0	0	1.72	0.28
1.58	0.84	1.16	0	0
2.16	0	0.42	0.08	0
0.5	0.3	0	0	0
1.51	0.392	0.408	0	0
3.33	0	0	0.67	0.33
3.59	0	0	0.369	0.531
3.72	0	0	0.504	1.296
3.87	0	0	0.13	0.87

The results in Table 7 are obtained by calculation of the weighted arithmetic mean.

Table 7. Weighted arithmetic mean

$\overline{MJW}$	$\overline{MNW}$	$\overline{MSN}$	$\overline{MNS}$
-0.16	-0.22993	0.372929	0.526

In case of IFE method, regardless of whether a factor is of major or minor weakness or strength, it has the same influence on the final result. In case of FSIF method, a factor is further weighted by multiplying the degree of membership of internal factors to major weakness to -1.5, and to major strength to 1.5. This increases the impact of these factors on the overall state of internal functional areas of the organization. Moreover, by displaying the obtained values that represent the weakness of organization by negative numbers, it is more evident which of the sets has more influence to the final result.

Finally, by summing up the obtained weighted arithmetic mean, in this case we obtain:

$$I^O = 0.509 .$$

Since the obtained value is positive, the observed organization has more advantages than disadvantages. However, as the value is relatively low (in comparison with maximum value 3) , it is possible to introduce some changes to improve its condition.

In the case of IFE method, the resulting value of 2.98 (in comparison with maximum value 4) would indicate that the state of organization is better then it is shown by FSIF method. First of the reasons is that the influence of factors representing minor and major weakness as well as minor and major strength to the final result is equal. Second reason is the fact that during rounding, some numbers between major weakness and minor weakness got value 2 (minor weakness) and others, which are between minor strength and major strength got value 4 (major strength). Some factors that in some degree have both roles: of minor strength and of minor weakness got values of minor strength (value 3). In case of FSIF method this difference is clearly indicated and therefore the final result is more precise and accurate.

5. Conclusion

This paper is focused on internal aspects of corporation, and further on a new measure, which is developed by using an original method of an internal assessment based on fuzzy logic.

Classical mathematical disciplines which are based on two valued logic could not be satisfactorily used in investigation of human behavior, which is an argument for the implementation of fuzzy logic in solving specific problems in the internal organizational environment.

Fuzzy set theory and fuzzy logic, as a mathematical approach to solving problems of analysis of internal functional areas of the organization allows overcoming some of the problems faced by managers of organizations. The paper points to some limitations of IFE matrix method and propose a way to eliminate them. Simple features of symmetric triangular fuzzy sets are used as well as graphics and elementary arithmetic operations.

An advantage of fuzzy-analytical methods in quantification of the organization's internal factors proposed in this paper, compared with IFE and other commonly used methods is in using fuzzy sets by which vague information is better assessed and described.

Besides, the scales used in this paper are not graduated, as is the scale used for measuring internal factors in IFE-matrix method. The scales considered here can take any value from an interval, which better describes real situation. Grades of weakness or strength of factors are represented by four fuzzy sets and our choice of symmetric triangular fuzzy sets enables easy calculations and understanding of the new tool. The importance of relevant factors is also taken into account using a weight from a real interval obtaining the ponder that increase or decrease the influence of factors.

Another advantage is that the additivity condition required in the mentioned IFE and IFAS methods is overcome in this fuzzy framework. The disadvantage of additivity setting is that if the importance of one internal factor is increased, then automatically importance of other internal factors must be reduced. Another disadvantage of known methods is the fact that a number of factors should be limited which is also overcome by this method.

The data obtained in a process of internal screening can be clearly presented and analyzed using the method proposed in this paper. For each organization a new fuzzy matrix can be produced taking into account variation of internal factors and difference in degree of influence of factors to the performance of organization.

The information obtained can be easily compared and interpreted, regardless of the organization observed, thereby increasing the diagnostic value of the proposed method.

Therefore, application of the FSIF method enables implementing more precise and more up-to date information system that assist managers in reaching a better understanding of the overall company’s competitiveness.

Acknowledgement. We are grateful to the referees for the comments that improved the paper. The research of the third author is supported by the Ministry of Education and Science of the Republic of Serbia, Grant No. 174013 and by the Provincial Secretariat for Science and Technological Development, Autonomous Province of Vojvodina, grant "Ordered structures and applications".

References

1. G. Bojadziev, M. Bojadziev, Fuzzy Logic for Business, Finance and Management, Advances in Fuzzy Systems – Applications and Theory, Vol. 23, World Scientific Publishing Co., Singapore, (2007)
2. F. David, The Strategic Planning Matrix – A Quantitative Approach, Long Range Planning 19, no.5, (1986)
3. F. David, Strategic Management: Concepts and Cases, (11th Edition), Prentice Hall, (2006)
4. S. Ghazinoory, A. Esmail Zadeh, A. Memariani, Fuzzy SWOT analysis, Journal of Intelligent & Fuzzy Systems 18, (2007)
5. S. Ghazinoory, A. Esmail Zadeh, A. Memariani, Application of fuzzy calculations for improving portfolio matrices, Information Sciences 180, 1582–1590 (2010)

6. J. Gil-Aluja, Fuzzy Sets in the Management of Uncertainty, Springer-Verlag, Heidelberg, (2004)
7. P.V. Janster, D.E. Hussey, Company Analysis: Determining Strategic Capability, Wiley, New York, (2001)
8. E. Pap, Z. Bošnjak, S. Bošnjak, Application of fuzzy sets with different t-norms in the interpretation of portfolio matrices in strategic management, Fuzzy Sets and Systems 114, 123-131, (2000)
9. A. Tepavčević, A. Pešić, Special Lattice Intuitionistic Fuzzy Sets and Applications in Management in Non-profit Organization , Fuzzy Economic Review, Volume X, Number 1, (2005)
10. A. Tepavčević, A. Vujic, On an application of fuzzy relations in biogeography, Information Sciences, 89, 77-94 (1996)
11. A. Tepavčević, G. Trajkovski, L-fuzzy lattices: an introduction, Fuzzy Sets and Systems 123, 206-219 (2001)
12. T.L. Wheelen, D.L. Hunger, Strategic Management and Business Policy, Eleventh Edition, Prentice Hall Inc., (2007)
13. Fuzzy Sets in Management, Economics and Marketing, C. Zopounidis, P. Pardalos, G. Baourakis (Eds), World Scientific, (2001)

Dr Aleksandar Pešić received a PhD degree in Industrial Engineering and Management at the Faculty of Technical Sciences, University of Novi Sad, Serbia. He is an assistant professor at the Faculty of Business and Industrial Management, Union University, Belgrade. His research interests include marketing research and quantitative assessments of the organizational effectiveness. He is an author of about 20 research papers and three textbooks.

Dr Duška Pešić got her PhD at Faculty of Science, University of Novi Sad, Serbia. She is a professor at Information Technology School in Belgrade, Serbia. She has been working on applying mathematical methods in management and economics and on teaching methods and techniques in mathematics. She is an author of about 30 research papers.

Prof. dr Andreja Tepavčević got her PhD in Mathematics 1993 at University of Novi Sad, Serbia. She is a professor at Faculty of Science in Novi Sad since 2003. Her research interests are in fuzzy set theory, universal algebra, lattice theory and applications. She is an author of about 90 research papers and five university textbooks. She participated at about 60 conferences and gave invited lectures at about 15 universities and was a member of organizing and program committees for several seminars and conferences. She is a member of AMS and IEEE.

Received: December 01, 2011; Accepted: May 23, 2012.

DOI: 10.2298/CSIS110720007K

Experimental investigation of the quality and productivity of Software Factories based development

Andrej Krajnc[1], Marjan Heričko[1], Črt Gerlec[1], Uroš Goljat[1] and Gregor Polančič[1]

[1] University of Maribor,
Faculty of Electrical Engineering and Computer Science,
Smetanova ulica 17, SI-2000 Maribor, Slovenia
{andrej.krajnc1, marjan.hericko, crt.gerlec, uros.goljat, gregor.polancic}@ uni-mb.si

Abstract. Software organizations are always looking for approaches that help improve the quality and productivity of developed software products. Quality software is easy to maintain and reduces the cost of software development. The Software Factories (SF) approach is one of the approaches to provide such benefits. In this paper, the quality and productivity benefits of the SF approach were examined and evaluated with an experiment involving two treatments - the traditional and the SF approach. For the purposes of this experiment, the Goal – Question – Metric (GQM) approach was used. Participants were grouped into thirty-two teams. There were sixteen projects available. The results were evaluated and presented through quality and productivity criteria, which were used for the experimental study. The results showed that the Software Factories approach was significantly better than the traditional approach.

Keywords: software factories approach, benefits, quality, productivity, experiment.

1. Introduction

A continuous objective in software engineering is to develop high quality solutions within a short time [3, 6, 15]. This can be achieved with the use of known software development methods, or approaches, where the quality of solutions is provided [38]. In most cases the project stakeholders would like to evaluate the software development outcomes as well as the effectiveness and efficiency of the underlying software development approach.

Several approaches that help to decrease time and effort in software development activities exist [4, 5, 8, 9], whereas the most efficient way of creating software is not to develop it, but rather reuse it. The biggest motivation for reusing software assets is to decrease software development costs and reduce the time and effort needed for their development. Software quality can also be improved with software reuse [36]. When reusing software parts, it also improves maintainability, because of the use of already well tested software parts. When we discuss software reuse, we need to look at two different aspects of software reuse: developing for reuse and developing with reuse [36]. The first is important when something is developed for reuse, like a component or some software part, while the latter is important when such a component or part of the software code is reused. Over time, a large amount of different approaches and techniques for software reuse have been developed, including: software frameworks, software libraries, software generators, design patterns and software product lines.

Within these approaches, software frameworks and software product lines (abbreviated as SPL) have been established as one of the most successful approaches for software reuse, because their reuse is based on product families rather than on individual reuses [2, 18]. In relation to them, a new approach for successful software reuse has evolved over the past few years: Software factories (abbreviated as SF).

A SF is a pattern for an approach to software system development, in which instances of those systems share features, functionality and architecture [2, 3]. The underlying four concepts of SF are: SPL, architecture framework, automated guidance and Model-Driven Development (abbreviated as MDD). Leveraging these concepts, SFs provide knowledge in the following forms: asset-like architectural frameworks with common features, models to create parts of software patterns, and recipes and tools for helping the developer. These assets help to automate the delivery of members of an SPL. In other words, an SF can produce software solutions in a way analogous to the way an airplane factory produces airplanes. A certain SF can produce software products in a specific domain. If we have an SF for mobile applications and we want to develop a web portal, we have to use another SF.

The SF approach provides the following benefits [2, 3, 4, 6]: an increase in productivity, a decrease in the time to market, an increase in the level of reuse, the providing of automatic guidance, a higher level of abstraction and an increase in product quality. Some of these benefits can also be adopted from the SPL approach [6, 7].

1.1. Motivation for the study

As Lord Kelvin (1824-1907) noted [40] “To measure is to know” and “If you cannot measure it, you cannot improve it.” By applying this idea to the SF domain, we believe that the measurement of development approaches helps to control, estimate and improve development processes and consequently the organization.

One of the main objectives of software engineering is to continuously improve the quality of outcomes as well as the efficiency of engineering activities [38]. As stated in the previous section, several approaches that leverage these objectives do exist; however, if the effectiveness of these approaches is not objectively analyzed, we cannot generalize assumptions out of them.

The above statements are also valid in the SF domain, with many stated benefits (see the previous section) and researchers have reported that there is a lack of empirical investigations [41]. There have been some studies related to measuring SF benefits [4, 5, 18], but their results have been primarily focused on quality characteristic, like number of defects, and reusability. Another motivation for our study was to test the theoretically [2, 3] stated quality and productivity benefits of the SF approach in an empirical way. Doing such a study was also a test to see if the SF approach as such was mature enough to be used later on in real projects in the industry. Another motivation for this study was to motivate participants to use more advanced development approaches to develop solutions with higher quality.

The goal of this study was to empirically evaluate the SF approach and to investigate if it is more effective and efficient when compared to traditional development.

In our research, we have addressed and evaluated SF in terms of their quality and productivity, compared to a “traditional” software development approach. The traditional approach has been defined as *“a software development approach where the whole software product is built by developers from scratch.”* This means that no additional tools that help generate source code and no explicit design patterns are used. There is also no reuse of any already available code.

For our study we set up a following research question: “*Does the SF approach deliver better quality code and does it increase the productivity of the development team compared to a traditional approach?*”

According to the research question, we organized the paper as follows. This section further investigates software quality and productivity. Section 2 describes research foundation for this work; section 3 describes work related to the object of the investigation; section 4 describes the goals of our study, the hypotheses associated with the study and the design of the experiment. The results are presented and interpreted in section 5. In section 6, we have listed our conclusions together with the limitations, as well as the theoretical and practical implications.

2. Research foundation

In the following subsections, research foundation regarding software quality and productivity is presented.

2.1. Software quality

The primary objective of software development projects is to fulfill either the stated or implied user requirements, which are commonly conceptualized in terms of software quality [38, 39]. To achieve high quality, it is important to use proper measurements for software solutions [38]. International standards, like ISO 9126, emphasize the need for measurements for assuring product quality [30]. It is important to measure software from the beginning of development until the end of the product's lifecycle. Software quality is divided into internal and external quality [30]. Internal quality is the totality of characteristics of the software product from an internal view [30]. Internal quality is measured and evaluated against the internal quality requirements. The details of software product quality can be improved during code implementation, reviewing and testing [30]. External quality is the totality of characteristics of the software product from an external view [30]. It is the quality when the software is executed, which is typically measured and evaluated while testing in a simulated environment with simulated data using external metrics [30]. Despite the fact that new software development methods and approaches have been applied in software development, there are still problems with the quality of software products [38]. Quality is judged according to different characteristics [38], which is given different significance for different stakeholders. Therefore, as previously mentioned, different views of software quality can be observed and analyzed. There are differences in analyzing quality from the point of view of customers or users on one hand, or from the point of view of the development team on the other [30]. As mentioned, we are firstly interested in external quality and subsequently in internal quality (Fig. 1).

When measuring internal quality, we use different software metrics [38] that cover large aspects of object-oriented development, like complexity, inheritance, coupling, and cohesion [10, 16]. It is important to have goal-oriented measurement; there you can define clear objectives that you want to achieve with a measurement [12, 38].

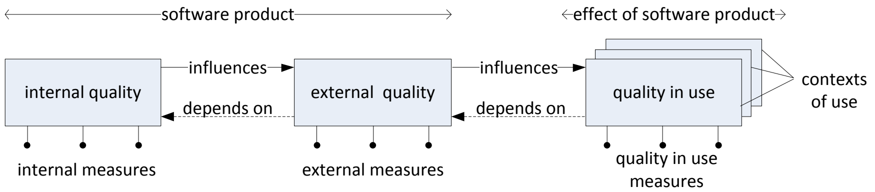


Fig. 1 Relationships between types of software quality [30]

In research about the SPL and SF approach, there are many cases regarding improved quality [3, 4, 6, 29, 37]. These cases are primarily related to external quality measurement (like the number of defects) or they theoretically state the quality benefits of such an approach. However, researchers should also focus on the internal quality of the SF approach and its derived software code, as well as its quality.

2.2. Software productivity

Productivity is one of the most important benefits, when considering the evaluation of a new software approach. Due to the ever-increasing demand for new software, it is important to use an approach that increases the productivity of the development team and shortens the time that is necessary for solutions to get to the market [35]. Productivity also depends heavily on different aspects, such as the experience of the developers, the approach used, the development environment, and what domain the solutions will be developed in (as well as the knowledge the developers have of such a domain), etc [1, 5, 29]. Productivity has been widely researched and analyzed in different contexts and approaches.

Measuring software productivity has been discussed in different ways [1, 15, 32, 39]. Different metrics have been proposed and used [39], like using size related metrics - Lines of Code (LOC) and Number of Classes (NOC), when using object-oriented development. The most important measure of software productivity is to measure the effort needed for the development of a certain project, product or functionality [38, 39]. Effort is mostly presented as the time, in hours or minutes, required for development [15].

Another view on the SPL or SF productivity benefit deals with the economics of such an approach and the economics of reusability. There are several models that discuss software product line economics. Most decisions about which products to include in the SPL and how to organize and structure the development of the products are economic decisions. One of the most known models is the model SIMPLE [31]. The model provides a set of functions that account for the expenses and benefits of building the product line and operating the product line organization [31]. It helps an organization decide if it should adopt the software product line strategy to build products through the costs and benefits related with the use of the SPL approach. Poulin [35] presented a model for estimating the financial benefits of software development with SPL. The model was used to calculate the “Product Line Return on Investment (ROI)” metric. In this analysis, Poulin also used an LOC metric and the percentage of the reuse code in each project. The authors in [32, 33] present the economic impact on adoption of an SPL approach. Their findings were related to the increase of quality, and improved productivity. With their model, they present a top-down approach to evaluate the SPL process of development. An important part of their study was to study the reuse effort. When talking about reuse effort, there are different approaches to measuring reusability, such as the level of the reuse in organizations, reuse Return on Investment (ROI) metrics and the effort needed for the development for reuse [34]. These approaches are focused mostly on improving productivity and better quality when reusing software code or components.

3. Related work

In this section, related work will be presented. Related work is considered from different areas of software development, from software factories, software product lines and software frameworks.

3.1. Software factories

Software factories were largely presented by Greenfield, Short [2] and Lenz, Wienands [3]. The authors addressed and stated the benefits that this approach delivers with its adoption.

Menendez [17] performed a study of SF-based projects in the aerospace industry. He demonstrated a 35% productivity improvement in his research. In our research, we want to show that improvements can also be made for smaller projects. A limitation of this research is that the study was performed for demonstration purposes only and was only applicable for the relatively small domain of navigation in the aerospace industry.

Aoyama [19] presented an evolution of the SF approach at Fujitsu. A model for using the SF approach was adopted. Their productivity improvement was about 30% higher than the development process that was used before the SF approach. Some other benefits, including higher productivity, have also been gained, such as the incremental delivery of products, lower total costs and a shorter development cycle.

Matsumoto [28, 29] presented SF which was established at Toshiba. In Japan, several companies have used the SF approach, which helped to reduce the cost of software development and increased the quality of software. Each of them, especially Toshiba, achieved high levels of productivity and improved quality of software. For example, Toshiba achieved 0.2 detected errors per 1000 LOC.

3.2. Software product lines

Knauber et al. [11] defined several hypotheses related to SPL, where for the scope of our research, the following hypotheses are interesting:

- “SPL decrease the development effort per product.” We will adopt this hypothesis and change it a little bit to use with our productivity criteria.
- “SPL decrease the time to market per product”, which will also be adopted into our productivity hypotheses.

A limitation of their work is that their findings are based on theoretical conclusions, rather than on empirical data.

Ajila and Dumitrescu [13] conducted research about SPL evolution in the form of changes. The original goal of this research was to study the economic impact of market repositions on the product line and the identification of metrics that can be used to record changes in the product line. One of the goals was to measure the lines of code (LOC) metric in each period of development. They also measured the efficiency of the product line in the case of developing products. The results showed that the use of the product line approach eases the integration of a new product. In addition to this, the efficiency also increased. The limitation of this research was that the authors only used the Lines of Code (LOC) metric. No metrics for code complexity were used.

In [20], Chen presented an SPL process simulator. He simulated the process of the development life cycle in terms of time-to-market. The simulation results showed that time-to-market can be reduced. The data gathered in the simulation was compared to theoretical data in [11]. According to our research, where the real projects' data was used, their data was gathered from a simulator and the data input in a simulator was selected from randomly distributed numbers within a certain range.

3.3. Software frameworks

Object-oriented frameworks have been largely researched. Because of their important relation with SPL and SF, research work made on the productivity and quality of object-oriented frameworks is relevant.

Polančič et al. [21] presented an empirical examination of application frameworks success. They did a survey regarding several important factors. One of the factors also covered productivity and quality. In a survey with 389 participants, the average answer with regard to productivity improvement using frameworks was “agree” while the same amount were in agreement with quality improvement. Both marks are on Likert scales of 1-7 with end points of “strongly agree” and “strongly disagree”. The limitation of this paper was its research method compared to our research. The use of a survey can sometimes get objective answers from participants.

Basili et al. [15] did an experiment to better understand the benefits of reuse in an object-oriented framework. They conducted a four-month long experiment, in which a new project was developed. The results showed an approximately 34% reuse rate. Their findings were that productivity improved with the increase of the reuse rate. Productivity was presented as an equation between size and effort time. In the paper, no data about quality was presented.

Morisio et al. [22] presented an empirical study in an industrial context on the production of software using a framework. They tested hypotheses regarding productivity and quality. They made a direct comparison with the traditional approach. The limitation of this research is the development process, where all projects were developed by the same programmer. Also, quality was measured in the relationship between development effort and the rework effort required to correct the code. Productivity was markedly better at about 50%. It should be emphasized that really small projects were used; the largest had 2,673 lines of code. The limitation of this work is also that all projects were developed by a single developer.

Mamrak and Sinha [23] did a case study of productivity and quality gains using an object-oriented framework. They implemented several input forms using the framework. The effort they needed to generate and implement a new application using a framework was reduced by 23% in terms of the average lines of code written. By comparison, only the LOC metric was used and the quality factor was presented through reused code.

In [25] authors presented a study to assess the impact of experience and maturity on productivity in software development. Two projects were measured, one using initial and one using subsequent development. First project was developed using new platform. For the quality measurement software metrics have been used. The project developed with new platform had about four times higher development effort. On the other hand, quality measures were both, more or less equal and there was no significant difference. A limitation of the study was that the participants were not familiar with developing with a new platform.

Table 1. Summary of related work.

Author	Research area	Methods	Factors	Results
Menendez[17]	SF	Case study	Productivity	35% productivity improvement
Aoyama [19]	SF	Case study	productivity, shortened development time	productivity improved by about 30%
Matsumoto [28, 29]	SF	Case study	productivity, quality	higher level of productivity achieved, quality presented as less number of defects
Knauber et al.[11]	SPL	Theoretical study	productivity, shorten time to market	hypothetical graphs of improved productivity
Ajila and Dumitrescu [13]	SPL	Case study	size of product code (productivity), changes on the product line	lower time-to-market for products, integration of new product is easier and more efficient

Chen [20]	SPL	Simulation	effort reduction, time-to-market reduction	improved development effort after a number of products developed
Polančič et al. [21]	Software Frameworks	Survey	different factors for the acceptance of object-oriented frameworks, including productivity and quality	participants agreed with productivity and quality improvement when using software frameworks
Basili [15]	Software Frameworks	Experiment	productivity, level of reuse	as reuse rate in projects increases, productivity increases
Morisio et al.[22]	Software Frameworks	Experiment	productivity, quality	productivity improved by about 50%
Mamrak and Sinha [23]	Software Frameworks	Case study	productivity, quality	23% less LOC written
Tomaszewski and Lundberg [25]	Software Frameworks	Experiment	productivity, quality	productivity was improved about four times higher using new approach, quality measures show no significant difference
Our study	SF	Experiment	productivity, quality	Improves productivity and effort for about 14%, better quality of code

4. The empirical study

The objective of our study was to investigate the impact of the SF approach on software quality and software development productivity, as compared to the traditional approach. To achieve the objective, we used between-subjects based on an experimental method. The experimental research as described in this article was a part of a larger study in which we investigated different effects of the SF approach on the software development process. For the purpose of this part of the research, we have narrowed it down and focused on the quality and productivity of the code.

4.1. Experimental variables and hypotheses

As previously mentioned, for the purposes of this research, we defined the criteria that are included in the literature as the advantages/benefits of SF [2, 3] and SPL [6, 7]. The GQM approach [12] (Table 2) for defining factors and corresponding metrics from the stated objective of the research was used.

Table 2. GQM approach [12] for our research.

Goal	Purpose	Experimental investigation
	Issue	Quality and productivity benefit
	Object	Software Factories based approach
	Viewpoint	From the developers' viewpoint
Question	Does the SF approach increases *productivity* and decreases the time of development?	Does the SF approach deliver a better *quality* of code?
Metric	Effort Time, LOC, NOC	Quality index (QI)

4.2. Productivity hypothesis

One of the core benefits of SF is in alignment with the following statement: “The Software Factories approach increases the productivity of the development team and decreases the time needed for development [3]”. As mentioned in Section 2.2, for productivity, it is also important that the development be done with software reuse, because it shortens the development time. As stated in the motivation for the study, the goal was to measure if the development effort would be improved with the use of the SF approach. In our study, productivity was defined as the amount of work indicated with the sum of hours needed for the implementation of a project, Lines of Code (LOC) and the Number of Classes (NOC) metrics. According to this, we declared the following hypothesis:

H1: The time needed for the development of software projects using a traditional approach is greater than the time and effort needed for the development of software projects when using the Software Factories approach.

The corresponding null hypothesis states that there is no difference between the two groups of approaches (SF versus traditional approach) in light of the productivity.

4.3. Quality hypothesis

As introduced in Section 2.1, quality can be measured internally and externally. In our study, we have chosen to measure internal quality. With internal quality we can cover different aspects of software code and its quality.

Several software development quality metrics exist [24]. In our study we have decided to measure the quality of code with software metrics sets presented in [14].

We tested the code quality with both approaches (SF and traditional). The metrics used in this criterion provided numerical values. These values represent objective metrics, because the input in the metric function is data and the output from the function is a single numerical value. In this case, there is no impact on the results. On the other hand, if one looks at the subjective metrics, there can be some influence on the results. This is the reason why we decided to take software product metrics [10, 24], which were already tested and are statistically proven. Size-related metrics were used in the productivity phase of the measurement. Quality measurement was realized with object-oriented class-related metrics [1, 14]. The chosen metrics cover the coupling, complexity and maintainability of classes [16].

The Quality Index (QI) was proposed in [1, 14]. The authors defined the method (equation) that expresses the quality of code with different metric sets. In [14], interchangeable metric sets were evaluated to calculate QI. We used those interchangeable metric sets to evaluate the quality of code in both approaches.

The Quality Index is defined as $QI=\frac{\sum_{i=1}^{n}PMQ_iR}{n}$

$$PMQ_iR=f_i(m\gamma)$$

where *PMQR* is the product metric quality rating, which is between 0 and 5 value, *n* is the number of code metrics used in the calculation, *mv* is a code metric value and *f* is the function that transforms the metric value for the metric *i* to the product metric quality rating [1]. The quality rating transformation function is defined for each metric individually. The QI is composed of *n* product metrics. The number of metrics and its type should be defined according to the project and environment characteristics. Each product metric has its threshold values [1, 14]. For the original quality index QI(0), the following metrics have been chosen: the Depth of Inheritance Tree (DIT), Coupling between Objects (CBO), Lack of Cohesion in Methods (LCOM), and the Maintainability Index (MI). According to these statements regarding quality, the following hypothesis was declared:

H2: The Software Factories approach delivers code that has higher quality than code delivered with a traditional approach.

The corresponding null hypothesis states that there is no difference between the traditional and SF approach.

4.4. Experimental participants

We were aware that our ideal candidates for experimental subjects would be a group of people, who already had prior knowledge in the field of SF and approximately equal experience and familiarity with the SF approach. For practical reasons we searched for candidates among undergraduate students of the same course, which we previously trained to have the same amount of training with each approach available (SF and traditional). In this way, we minimized the effect that different prior knowledge or experience could have on the experimental results.

Because advanced development technologies were presented, the students needed to have experience with object-oriented technology, including an object-oriented programming language.

Participants were in their final year of their study, so they had a sufficient development and programming experience.

The reason for selecting final year students was, because research was based on a complex technology, such as SF approach is. We provide participants with extra help during courses, as mentioned before; training of each approach was available for them and a short questionnaire about their programming experiences was made. All that was provided for this, that relation between university students and subjects in other context (like more experienced developers) can be drawn.

4.5. Experimental objects

In this part, we will present the objects that the subjects are going to examine or work with. The objects are applications or projects, developed as web applications in different domains using the SF and traditional approach. In Table 3, the projects and their domains are presented. In the third cell, a percentage of the Core Asset Base (CAB) is presented when using the SF approach. With the creation of a project based on the SF approach, a common part for each project developed in such a way is created. Measured in the LOC metric, this CAB part contains 1,072 lines of code. Everything else represents the variabilities of each project. For the SF approach, a Microsoft Web Client Software Factory was used, which is a package for developing web applications using the SF approach in Microsoft Visual Studio Environment. As previously mentioned, this package creates a core project for developing web applications (ASP.NET) and uses known patterns for development, such as the Model-View-Presenter and Model-View-Controller. On the other hand, subjects who developed objects with a traditional approach, use the object-oriented development paradigm and the ASP.NET technology for developing web-based applications on the Microsoft platform. Both groups used the object–oriented programming language C#. Projects were developed in the Microsoft Visual Studio environment.

Table 3. Objects examined by subjects in study

Project	Domain	CAB (%)
Project1	Accounting (web application)	13.05
Project2	Warehousing (web application)	8.45
Project3	Banking (web application)	9.95
Project4	Warehousing (web application)	12.95
Project5	Accounting (web application)	13.46
Project6	Warehousing (web application)	10.68
Project7	Warehousing (web application)	26.08
Project8	Other services (cinema services – web application)	15.61
Project9	Other services (restaurant services – web application)	9.40
Project10	Other services (taxi services – web application)	15.54
Project11	Banking (web application)	18.35
Project12	Accounting (web application)	19.93
Project13	Accounting (web application)	10.74
Project14	Banking (web application)	10.17
Project15	Warehousing (web application)	11.91
Project16	Warehousing (web application)	10.16

4.6. Experimental environment and instruments

The experiment was conducted at a university setting, within the laboratory work of a subject related to SF. In this laboratory work we taught students how to develop multi-tier applications over a four-month long course. We used the process of this laboratory work for our experiment in such a way that the students had to build web applications using SF approach and traditional approach.

The experimental process started with a short questionnaire about the students' programming knowledge and experiences.

There were sixteen projects available and two different approaches used: the SF approach and the traditional approach. The first thing the students did was randomly choose their project and the approach they were going to use. Every project was developed in two different ways: one student developed it with a traditional approach and one with an SF approach.

Table 4. Experimental study.

R	**G_{SF}**	O_1	X_{SF}	O_Q	O_P
	G_{TR}	O_1	X_{TR}	O_Q	O_P

Notes: R… randomization process, X_{SF} … **treatment** SF approach, X_{TR} … **treatment** Traditional approach, **G_{SF} … Group** SF approach, **G_{SF} … Group** Traditional approach**, O … observation, O_Q … observation** quality**, O_P … observation** productivity

The development process was divided into iterations.

In the first two iterations, the students had to be grouped together by project and had to complete the requirements. They also conducted a design phase for projects. For every project, quality design specifications were prepared and provided by the supervisor. It was necessary to check and analyze the specifications together with students per project, because both needed to have the same requirements and design. That was for the purpose of the SF benefits evaluation. Students filled out a document about their working status and productivity. After the design phase, the document was examined by a supervisor, who provided comments on the requirements.

After that, the education of developing web applications was turned over to the students. Also, tutorial implementations were added on the course site for them. Complete documentation was also provided. The supervisor was also available during courses to answer questions about the use of developing web applications and developing web applications using the SF approach.

Then the implementation phase began. The first step for students was to set up a project solution. Their task was to write down the time they spent learning the technology or use of the SF approach. The next step was the implementation of the project. For the implementation phase, a Microsoft

Web Client Software Factory was used for the SF approach while students who developed traditional approach used the ASP.NET technology for developing web-based applications on the Microsoft platform.

5. Experimental results

5.1. Descriptive statistics

All participants had some experience with object-oriented programming languages and relational databases, and therefore had the basic skills necessary for such a study (Table 5).

Table 5. Descriptive statistics.

Variable	Values	Freq.	Valid percent (%)
Gender	Female	3	9.4%
	Male	29	90.6%
Programming experience	Basic	26	81.3%
	Advanced	5	15.6%
	Expert	1	3.1%
Years of programming experiences (besides study)	0 years	16	50.0%
	< 1 year	11	34.4%
	1 – 2 years	3	9.4%
	>2 years	2	6.2%
Programming knowledge of .NET environment	Basic	26	81.3%
	Advanced	2	6.2%
	Expert	3	9.4%
	I don’t use .NET	1	3.1%
Knowledge of developing web applications in a .NET environment	Basic	20	62.5%
	Advanced	4	12.5%
	Expert	2	6.2%
	Don’t know	6	18.8%

Table 5 shows the descriptive statistics of the experiment's participants and their previous experience with programming in a .NET environment. This experience was self-reported. As noted in Table 5, we analyzed 32 out of a total 32 responses. As anticipated, the typical participant was a male who was introduced to the object-oriented development during the course of their college studies and had some programming experience.

5.2. Productivity hypothesis testing

In the GQM model (Table 2) [12], we set out to measure productivity with actual hours and the size-related metrics Lines of Code (LOC) and Number of Classes (NOC). The projects developed with the SF approach were measured together with part of the code, which represented the percentage of CAB code, as presented in Section 4.5.

Table 6. Actual hours spent on each project.

Actual spent hours				LOC		NOC	
	TR	SF	%	TR	SF	TR	SF
Project1	115	91	20.87	3289	8214	28	221
Project2	108	112	-3.70	2381	12691	33	276
Project3	124	104	16.13	5971	10771	46	140
Project4	106	89	16.04	5585	8281	53	190
Project5	112	107	4.46	3004	7965	37	239
Project6	124	99	20.16	6420	10038	36	219
Project7	108	91	15.74	2163	4111	28	129
Project8	116	94	18.97	1577	6867	13	158
Project9	122	96	21.31	7970	11409	84	297
Project10	121	104	14.05	2647	6899	34	151
Project11	107	96	10.28	1758	5842	26	154
Project12	116	97	16.38	3672	5378	48	118
Project13	116	98	15.52	2027	9983	25	217
Project14	108	92	14.81	4486	10542	36	249
Project15	104	88	15.38	2652	9002	18	227
Project16	123	103	16.26	5396	10552	48	291
Sum	1830	1561					
Mean	114.38	97.56	14.70	3812.38	8659.06	37.06	204.75
STDEV	6.97	6.89		1926.57	2390.23	16.69	58.08
p (H_0)	0.00			0.00		0.00	
df	30			30		30	
t	6.86			-6.31		-11.10	

The actual hours spent were compared in a pair. The null hypothesis H_{01} was tested, which stated that mean values in actual spent hours in both approaches are equal. The alternative hypothesis H_{A1} states that mean values in actual spent hours are not equal. Both tests were also made with LOC and NOC results.

According to the summarized data, there is a difference in the hours needed for the development phase in both approaches. Participants using the SF approach needed, on average, 14.70% less time to develop projects.

Table 6 shows the actual hours spent on each project. In the data, we can see that projects using the SF approach needed less time. This difference with other projects is also seen in the result of the size-related metrics, Lines of Code (LOC) and Number of Classes (NOC). This difference is especially visible in the traditional approach. The LOC value in projects using the traditional approach is high. The effort results are in favor of SF approach. Table 6 also shows statistics for the tested pairs. For all three factors - spent hours, LOC and NOC value - the difference is significant at $p<0.05$. The evaluation of results for the productivity criteria shows that the time and effort needed for the development of a SF project is less than the time and effort needed for the development of projects with a traditional approach. Therefore, the null hypothesis (H_{01}) is rejected in favor of the alternative hypothesis (H_{A1}).

5.3. Quality hypothesis testing

Our research was based on the quality of the code, which was measured with selected software product metrics. Also, the projects here developed with the SF approach were measured together with part of the code which represented the percentage of CAB code, as presented in Section 4.5. Based on [1, 14], the quality index was tested on data. The QI measure was compared in a pair. The null hypothesis H_{02} was tested, which stated that mean values in QI measure are equal. The alternative hypothesis H_{A2} states that mean values in QI measure are not equal.

An analysis of the results (Table 7) shows that the SF approach does deliver more quality code, as can be seen with the QI measurement. Normally, good code is expected to be QI > 3 [1, 14]. For the measurements, different metric sets were used. In all sets, the QI of the SF approach was better. The code developed with a traditional approach delivered less quality code, because it used more complex code and had a smaller set of classes. The SF approach brought a more controlled environment, which helped with the maintainability and complexity of projects. Table 7 also shows statistics for the tested pairs. For all QI measurement sets the difference is significant at $p<0.05$. Therefore, the null hypothesis (H_{02}) is rejected in favor of the alternative hypothesis (H_{A2}). All results point to better software quality code when using the SF approach.

Table 7. Quality index measurement

	QI(0)		QI(1)		QI(2)		QI(4)		QI(10)	
	TR	SF	TR	SF	TR	SF	TR	SF	TR	SF
Project1	2.75	4.00	2.25	3.00	2.50	3.75	2.25	2.75	3.00	3.75
Project2	3.00	3.75	2.25	2.75	2.75	3.50	2.25	2.75	3.25	3.50
Project3	3.50	4.00	2.50	3.00	2.75	3.50	1.75	2.75	2.75	3.50
Project4	2.50	3.75	2.00	3.00	2.50	3.50	2.00	2.75	3.00	3.50
Project5	3.00	4.00	2.25	3.00	3.00	3.75	2.25	3.00	3.25	3.75
Project6	3.50	4.00	2.50	3.00	2.75	3.75	1.75	2.75	2.75	3.50
Project7	3.00	4.00	2.25	3.00	3.00	3.75	2.25	2.75	3.00	3.50
Project8	3.00	3.75	2.75	3.00	2.00	3.25	1.75	2.75	2.75	3.00
Project9	3.00	4.00	2.25	3.00	3.00	3.75	2.25	2.75	3.00	3.75
Project10	3.00	3.75	2.50	3.00	2.75	3.50	2.25	2.75	3.00	3.50
Project11	3.00	4.00	2.50	3.00	3.00	3.50	2.25	3.00	3.00	3.50
Project12	3.00	3.75	2.50	3.00	3.00	3.75	2.25	2.75	3.00	3.50
Project13	3.25	3.75	2.50	3.00	3.00	3.50	2.25	2.75	3.00	3.50
Project14	2.75	4.00	2.25	3.00	2.75	3.50	2.25	3.00	3.00	3.50
Project15	2.75	4.00	2.50	3.00	2.75	3.75	2.25	2.75	3.00	3.75
Project16	3.00	4.00	2.25	3.00	2.75	3.75	2.25	3.00	3.00	3.75
Mean	3.00	3.91	2.38	2.98	2.77	3.61	2.14	2.81	2.98	3.55
STDEV	0.25	0.12	0.18	0.06	0.27	0.16	0.20	0.11	0.14	0.18
p (H_0)	0.00		0.00		0.00		0.00		0.00	
df	30		30		30		30		30	
t	-12.64		-12.63		-10.93		-11.57		-9.53	

6. Discussion

6.1. Threats to validity and research limitations

By threats, we are referring to threats to internal and external validity [27]. The first threat to validity of this study is the participants' background. It was the first time that many participants developed web applications, especially in the chosen technology. The concept of the SF approach was also new to them. Second, the experimental setting alone is a threat to validity, because it was the participants' random choice to select what approach and which project they would work on. Third, when dealing with university students, it was also difficult to validate the accuracy of the provided effort data and have confidence in them to actually fill out the data for the actual spent hours correctly.

External validity refers to the approximate truth of conclusions involving generalizations within different contexts [26]. External validity threats are always present when experiments are performed with students. However, last year students have a sufficient development and programming experience, thus we can consider them as a less experienced software engineers. Consequently, we regard them as representatives of the context where we would like to generalize the achieved results. Another possible threat to external validity is that our projects were small, suggesting that their complexity and functionality may be limited when compared to large software projects. Nevertheless, replications should be performed with different subjects in different contexts to confirm or contradict the results.

Conclusion validity concerns the issues that affect the ability of drawing a correct conclusion. A definition of conclusion validity could be the degree to which conclusions we reach about relationships in our data are reasonable [26]. The conclusion validity threats were mitigated by the experiment design and by the properly selection of the population. Regarding the recruited subjects, we drew a fair sample from that population and conducted our experiment with subjects belonging to this sample. Moreover, proper tests were performed to statistically reject null hypotheses.

Readers should also interpret our results while considering the following limitations. The metrics thresholds values for quality indexes were used from [14] and these values are programming-language and design-approach dependent. For development, the Microsoft development environment and technology was used.

6.2. Theoretical and practical implications

Several theoretical and practical implications of our work can be foreseen. First, we defined a model based on the GQM approach for the evaluation of an SF approach within the context of quality and productivity. For researchers and practitioners, the evaluation model can also be applied to other research areas, like software frameworks, software product lines and the adoption of design patterns. Second, some researchers and practitioners have already proposed some of the benefits gained when using the SF approach. But only the productivity factor was evaluated and researched. In this research, we added the quality of code factor and investigated its impact. Product development managers can, through our research, gain an idea on the economic benefits of software development, because less effort is needed and there is improved quality for the products developed with the SF approach.

For the SF approach, it was known that its implementation provides benefits, such as quality and productivity. Papers in related work have shown benefits being achieved, albeit using different metrics and variables. Our empirical study on the other hand contributed to the collective knowledge of the SF approach with regard to the quality of the developed code.

6.3. Future work and conclusions

In our study we followed an empirical approach in evaluating engineering techniques to gain transferrable insights about them. As previously mentioned, this is to rarely done in our field, but is still necessary. We presented an empirical study to assess the impact of the Software Factories (SF) approach on a set of product and code quality indicators. The study was conducted in a university setting on 16 projects developed using “traditional” and SF approaches during an experiment. We specifically studied the impact of an SF approach on quality of code and productivity. The results showed that the use of the SF approach in software development has a statistically significant impact on the quality of code and productivity. In the experiment, small projects were used and a difference in quality and productivity could already be seen. One can conclude that this will also hold true for larger projects, which are greater and more complex. With the SF approach, developed projects are more maintainable and have better quality.

The achieved results can be considered relevant as we tried to minimize the gap between the university setting and industry environment. Indeed, the selected participants are not far from actual stakeholders since they were familiar with object-oriented programming and development.

However, despite the significance of the achieved results, we are going to replicate the experiment in different contexts. In particular, we plan to perform the replication with industry subjects (more experienced software developers).

Our future direction aims to investigate the impact of the SF approach on documentation, reusability, maintainability and modularity in software development.

References

1. M. Heričko et al., A method for calculating acknowledged project effort using a quality index. Informatica (Ljublj.), (2007), vol. 31, no. 4, pp. 431-436.
2. J. Greenfield, K. Short, Software Factories: Assembling Applications with Patterns, Models, Frameworks, and Tools, John Wiley & Sons, Indianapolis, 2004.
3. G. Lenz, C. Wienands, Practical Software Factories in .NET, Apress, 2006
4. J.S.Her et al., A framework for evaluating reusability of core asset in product line engineering, Information and Software Technology 49 (2007) 740-760.
5. D. Zubrow, G. Chastek, Measures for software product lines, Technical Notes CMU/SEI-2003-TN-031, 2003
6. P. Clements, L.M. Northrop, Software Product Lines - Practices and Patterns, Addison-Wesley, Boston, 2001.
7. H. Gomaa, Designing Software Product Lines with UML, George Mason University, Addison-Wesley, Boston, 2004.
8. Etzkorn et al., Automated reusability quality analysis of OO legacy software, Information and Software Technology 43 (2001) 295-308.

9. Chatzigeorgiou et al., An empirical study on students' ability to comprehend design patterns, Computers & Education 51 (2008) 1007–1016.
10. N. Fenton, S. L. Pfleeger, "Software Metrics: A Rigorous and Practical Approach", second edition, International Thomson Computer Press, London, UK, 1997.
11. P. Knauber et al., Quantifying Product Line Benefits, PFE-4 2001, LCNS 2290, pp. 155-163, 2002.
12. V. Basili, (1994). "The Goal Question Metric Approach" (PDF). ftp://ftp.cs.umd.edu/pub/sel/papers/gqm.pdf, Retrieved on 22.01.2011.
13. S.Ajila, R.Dumitrescu, Experimental use of code delta, code churn, and rate of change to understand software product line evolution, The Journal of Systems and Software 80 (2007) 74-91.
14. A. Živkovič, U. Goljat, M .Heričko, Improving the usability of the source code quality index with interchangeable metrics sets. *Inf. process. lett.*, Feb. 2010, vol. 110, iss. 6, p. 236-240
15. V. R. Basili , L. C. Briand , W. L. Melo, How reuse influences productivity in object-oriented systems, Communications of the ACM, v.39 n.10, p.104-116, Oct. 1996
16. S.R. Chidamber, C.F. Kemerer, A metrics suite for object oriented design, IEEE Transactions on Software Engineering 20 (6) (Jun. 1994) 476–493.
17. J. Menendez, The Software Factory: Integrating CASE technologies to improve productivity, Report – Lean 96-02, July 1996, MIT
18. K. Pohl, G. Böckle, and F. van der Linden, Software Product Line Engineering: Foundations, Principles, and Techniques, 1st ed.: Springer-Verlag, 2005
19. M. Aoyama, Beyond Software Factories, concurrent-development process and an evolution of sofware process technology in Japan, Information and Software Technology 38 (1996) 133-143
20. Y. Chen, G. Gannod, J. Collofello, A software product line process simulator, Soft. Process Improve. Pract. 11(4), 385-409 (2006)
21. G. Polančič, M. Heričko, I. Rozman, An empirical examination of application frameworks success based on technology acceptance model, The Journal of Systems and Software 83 (2010) 574-584.
22. M. Morisio, D. Romano, I. Stamelos, Quality, Productivity and Learning in Framework-Based development: An Exploratory Case Study, IEEE Transactions on Software Engineering, vol.28,no.9, 2002
23. A. Mamrak, S. Sinha, A Case Study: Productivity and Quality Gains Using an Object-Oriented Framework, Soft. Pract. Exper. 29(6), 501-518 (1999)
24. V.R. Basili, L.C. Briand, W.L. Melo, A validation of object-oriented design metrics as quality indicators, IEEE Transactions on Software Engineering 22 (10) (Oct. 1996) 751–761.
25. P. Tomaszewski, L. Lundberg, The increase of productivity over time – an industrial case study, Information and Software Technology 48 (2006) 915-927
26. Shaddish,W., Cook T. & Campbell, D., Experimental and Quasi-Experimental Design for Generalized Causal Inference, Houghton Mifflin Co., 2002
27. Jedlitschka, A., Pfahl, D., Reporting guidelines for controlled experiments in software engineering, in Proc. ACM/IEEE International Symposium on Empirical Software Engineering (ISESE) 2005, IEEE Computer Society Press, pp. 95-195
28. Yoshihiro Matsumoto. "Toshiba Fuchu Software Factory," *Modern Software Engineering*, pp. 479-501, Van Nostrand Reinhold, New York (1990).

29. Yoshihiro Matsumoto. "A Software Factory, An Overall Approach to Software Production," *Software Reusability* ed. by P. Freeman, IEEE Computer Society, March 1987.
30. ISO/IEC 9126. Software Product Evaluation – Quality Characteristics and Guidelines for the User, International Organization for Standardization, Geneva, 2001.
31. J. McGregor, Qualitative SIMPLE, Journal of Object technology, vol. 7, no. 7, September-October 2008.
32. K. Schmid , M. Verlage, The Economic Impact of Product Line Adoption and Evolution, IEEE Software, v.19 n.4, p.50-57, July 2002
33. K. Schmid, An Economic Perspective on Product Line Software Development, First Workshop on Economics-Driven Software Engineering Research, 1999
34. J. Poulin, Measuring Software Reuse, Addison Wesley, 1996
35. J. Poulin, The Economics of Product Line Development, International Journal of Applied Software Technology, vol. 3, 20-34, March 1997
36. E.A Karlsson, Software Reuse: A Holistic approach, John Wiley, 1995
37. F. van der Linden, K. Schmid, E. Rommes, Software Product Lines in Action, The best industrial Practice in Product Line Engineering, Springer, 2007
38. C. Ebert, R. Dumke, Software Measurement, Springer, 2007
39. D. Galorath, M. Evans, Software sizing, estimation and risk management, Auerbach Publications, 2006
40. Lord Kelvin Quotations, http://zapatopi.net/kelvin/quotes/, Retrieved on 28.5.2012.
41. Frakes, W. B., & Kang, K. (2005). Software reuses research: Status and future. IEEE Transactions on Software Engineering, 31(7), 529–536.

Andrej Krajnc is a researcher and PhD student at the University of Maribor, Faculty of EE&CS, Institute of Informatics. He received his B.Sc. in computer science from the University of Maribor in 2006. His research work covers different aspect of Software Product Lines, Software Factories, .NET platform, object technology and software metrics. He worked in several industry projects.

Marjan Hericko is a Full Professor at the University of Maribor, Faculty of EE&CS, Institute of Informatics. He received his M.Sc. (1993) and Ph.D. (1998) in computer science from the University of Maribor. His research interests include all aspects of IS development with emphasis on metrics, software patterns, process models and modeling.

Uros Goljat is a teaching assistant at the University of Maribor. His research work covers agile software development methodologies (Scrum, XP, etc.), different aspects of object-oriented technology, internal software quality, and .NET platform. He gained his practical experiences in cooperation with industry on several projects. Uros received his Computer Science B.Sc. in 2007 from University of Maribor.

Črt Gerlec is a researcher and PhD student associated with the Faculty of Electrical Engineering and Computer Science, Institute of Informatics at the University of Maribor. His research interests are mining software repositories, software evolution, software quality, software metrics, information systems and more. He is experienced software developer on Microsoft.NET platform and expert for software architecture, design patterns and best practices.

Dr. Gregor Polančič is an assistant professor at the Institute of Informatics. He received his PhD in Computer Science from the University of Maribor in 2008. His main research interests are: (1) analysis and design of software, (2) web applications, communities, architectures and patterns, (3) FLOSS software, projects and development models, (4) business process modeling, informatization and re-engineering and (5) computer mediated communication and collaboration. Dr. Polančič has been a work co-ordinator/member of several applied projects and work co-ordinator/member in several international research projects. Dr. Polančič has appeared as an author or co-author in more than 10 peer-reviewed scientific journals. In all, his bibliography contains more than 130 records.

Received: July 20, 2011; Accepted: June 04, 2012.

DOI: 10.2298/CSIS111211014T

Nearest Neighbor Voting in High Dimensional Data: Learning from Past Occurrences

Nenad Tomašev and Dunja Mladenić

Artificial Intelligence Laboratory, Jožef Stefan Institute and
Jožef Stefan International Postgraduate School
1000 Ljubljana, Slovenia
nenad.tomasev@ijs.si, dunja.mladenic@ijs.si

Abstract. *Hubness* is a recently described aspect of the *curse of dimensionality* inherent to nearest-neighbor methods. This paper describes a new approach for exploiting the hubness phenomenon in k-nearest neighbor classification. We argue that some of the neighbor occurrences carry more information than others, by the virtue of being less frequent events. This observation is related to the hubness phenomenon and we explore how it affects high-dimensional k-nearest neighbor classification. We propose a new algorithm, Hubness Information k-Nearest Neighbor (HIKNN), which introduces the k-occurrence informativeness into the hubness-aware k-nearest neighbor voting framework. The algorithm successfully overcomes some of the issues with the previous hubness-aware approaches, which is shown by performing an extensive evaluation on several types of high-dimensional data.

1. Introduction

Supervised learning (classification) is one of the most fundamental machine learning tasks, often encountered in various practical applications. It involves assigning a label to a new piece of input data, where the label is one out of several predefined categories. Many algorithmic approaches to performing automatic classification have been explored in the past. This includes, among others, Bayesian learning methods, support vector machines (SVM), decision trees and nearest neighbor methods [1].

The k-nearest neighbor algorithm is one of the simplest pattern classification algorithms. It is based on a notion that instances which are judged to be similar in the feature space often share common properties in other attributes, one of them being the instance label itself. The basic algorithm was first proposed in [2]. The label of a new instance is determined by a majority vote of its k-nearest neighbors (kNN) from the training set. This simple rule has some surprising properties which go in its favor. For instance, when there is no overlap between the classes, 1-nearest neighbor is asymptotically optimal [3] [4]. As for the kNN rule, it has been shown to be universally consistent under some strong assumptions, namely $k \to \infty$ and $k/n \to 0$ [5] [6].

Let $D = (x_1, y_1), (x_2, y_2), ..(x_n, y_n)$ be the data set, where each $x_i \in R^d$. The x_i are feature vectors which reside in some high-dimensional Euclidean space, and $y_i \in c_1, c_2, ..c_C$ are the labels. It can be shown that in the hypothetical case of an infinite data sample, the probability of a nearest neighbor of x_i having label c is asymptotically equal to the posterior class probability in point x_i, namely $p(c|x_i) = \lim_{n\to\infty} p(c|\text{NN}(x_i))$. Real-world data is usually very sparse, so the point probability estimates achieved by kNN in practice are much less reliable. However, this is merely one aspect of the well known *curse of dimensionality*.

Concentration of distances [7,8] is another phenomenon of interest, since all nearest-neighbor approaches require a similarity measure. In high-dimensional spaces, it is very difficult to distinguish between relevant and irrelevant points and the very concept of nearest neighbors becomes much less meaningful.

Hubness is a recently described aspect of the dimensionality curse, related specifically to nearest neighbor methods [9] [10]. The term is coined to reflect the emergence of *hubs*, very frequent nearest neighbors. As such, these points exhibit a substantial influence on the classification outcome. Two types of hubs can be distinguished: *good hubs* and *bad hubs*, based on the proportion of label matches/mismatches in their k-occurrences. The phenomenon of *hubness* will be explained in more detail in Section 2.2, and the previous approaches for exploiting hubness in kNN classification will be outlined in Section 2.3.

The issue of data dimensionality needs to be emphasized because most real world data sets are in fact high-dimensional, for example: textual documents, images, audio files, data streams, medical histories, etc.

1.1. Contributions

This paper aims at further clarifying the consequences of hubness in high dimensional kNN classification, by focusing on one specific aspect of the phenomenon - the difference in the information content of the individual k-occurrences. Here we summarize the main contributions of the paper:

- When there is hubness, some points occur much more frequently in k-neighbor sets. We claim that some occurrences hence become much less informative than others, and are consequently of much lower value for the kNN classification process.
- We propose a new hubness-aware approach to k-nearest neighbor classification, Hubness Information k-Nearest Neighbor (HIKNN). The algorithm exploits the notion of occurrence informativeness, which leads to a more robust voting scheme.
- We provide a thorough experimental evaluation for the approach, by testing it both on low-to-medium hubness data and also high-hubness data from two different domains: images and text. The experiments are discussed in Section 5, while Section 7 takes a deeper look into the class probabilities which the algorithm returns.

2. Related work

2.1. kNN classification

The k-nearest neighbor method is among the most influential approaches in machine learning, due to its simplicity and effectiveness. Many extensions to the basic method have been proposed, dealing with various different aspects - including attribute weighting [11], adaptive distances [12] [13], fuzzy labels [14] [15] [16], evidence-theoretic approaches [17], and many more. Some advanced algorithms have been proposed recently, including the large margin kNN classifier which learns the Mahalanobis distance matrices via semidefinite programming [18] [19].

2.2. Hubs, frequent nearest neighbors

The emergence of *hubs* as prominent points in k-nearest neighbor methods had first been noted in analyzing music collections [20] [21]. The researchers discovered some songs which were similar to many other songs (i.e. frequent neighbors). The conceptual similarity, however, did not reflect the expected perceptual similarity.

The phenomenon of *hubness* was further explored in [9] [22], where it was shown that hubness is a natural property of many inherently high-dimensional data sets. Not only do some very frequent points emerge, but the entire distribution of k-occurrences exhibits very high *skewness*. In other words, most points occur very rarely in k-neighbor sets, less often than what would otherwise have been expected. We refer to these rarely occurring points as *anti-hubs*. [23]

Denote by $N_k(x_i)$ the number of k-occurrences of x_i and by $N_{k,c}(x_i)$ the number of such occurrences in neighborhoods of elements from class c. The latter will also be referred to as the *class hubness* of instance x_i. A k-neighborhood of x_i is denoted by $D_k(x_i)$.

The skewness of the $N_k(x)$ distribution in high dimensional data can sometimes be very severe [22]. Let us illustrate this point by plotting the $N_k(x)$ distribution for one of the datasets which we used for the experiments, namely the Acquis data. This is shown in Figure 1, for $k = 5$. Such a drastic shift in the distribution shape must certainly be taken into account when designing kNN algorithms for high dimensional data.

Hubness-aware algorithms have recently been proposed for clustering [10], instance selection [24], outlier and anomaly detection [22] [25] and classification [9] [26] [27] [28], which we will discuss below.

2.3. Hubness-aware classification

Hubs, as frequent neighbors, can exhibit both *good* and *bad* influence on kNN classification, based on the number of label *matches* and *mismatches* in the respective k-occurrences.The number of good occurrences will be denoted by

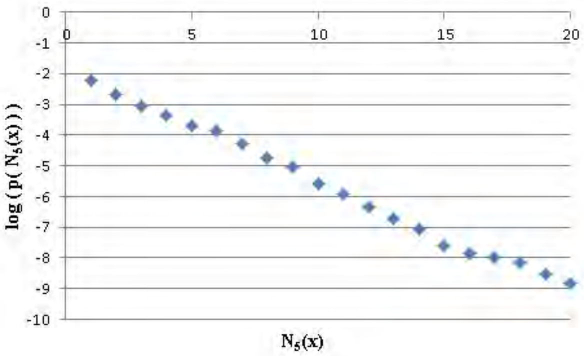


Fig. 1. The hubness distribution of the Acquis data is given for the 5-occurrence probabilities of $N_5(x) \in \{1..20\}$. We see that the distribution apparently forms a straight line on the logarithmic scale, so it is in fact exponential.

$GN_k(x_i)$ and the number of bad ones by $BN_k(x_i)$, so that $N_k(x_i) = GN_k(x_i) + BN_k(x_i)$.

All three previously proposed approaches deal with bad hubs in seemingly similar, but radically different ways. We will refer to these algorithms as hubness-weighted kNN (hw-kNN) [9], hubness fuzzy kNN (h-FNN) [26] and naive hubness Bayesian kNN (NHBNN) [27]. We discuss these ideas below, outlining their respective strengths and weaknesses.

hw-kNN

- **Idea:** When a point exhibits bad hubness, give its vote lesser weight. This has been achieved by calculating the standardized bad hubness as $h_B(x_i) = \frac{BN_k(x_i) - \mu_{BN_k}}{\sigma_{BN_k}}$, where μ_{BN_k} and σ_{BN_k} denote the mean and standard deviation of bad hubness, respectively. Each x_i is then assigned a voting weight of $w_i = e^{-h_B(x_i)}$.
- **Strengths:**
  - Reduces the influence of bad hubs
  - Very simple and easy to implement
- **Weaknesses:**
  - Each element still votes by its own label, which means that bad hubs still exhibit some detrimental influence
  - Some information is left unexploited, since class hubness is ignored
  - It is equivalent to kNN for $k = 1$

h-FNN

- **Idea:** Decompose bad hubness into fuzzy class-specific hubness-based votes as $u_c(x_i) \propto N_{k,c}(x_i)/N_k(x_i)$. This is only possible for points with

$N_k(x_i) > 0$ and only sensible for points with $N_k(x_i) > \theta$, where θ is some predefined threshold parameter. Anti-hubs are thus considered to be special cases. Their fuzzy votes are approximated by average class-to-class fuzzy votes. This algorithm is otherwise based on the fuzzy nearest neighbor (FNN) framework [14], with distance weighting included.

– **Strengths:**
  - Generalizes the hw-kNN approach by taking class hubness into account
  - Combines fuzzy votes with distance weighting

– **Weaknesses:**
  - No clear way of dealing with anti-hubs, approximations need to be used instead
  - Uses a threshold parameter θ for determining anti-hubs, which is difficult to set in practice. If learned automatically from the data, it can lead to over-fitting.

NHBNN

– **Idea:** Observe each k-occurrence as a random event and use the Naive Bayes rule to calculate the posterior class affiliation probabilities, as shown in Equation 1. The $x_{it}, t = \{1, 2..k\}$ represent the k nearest neighbors of x_i. As in h-FNN, anti-hubs are a special case and one needs to estimate their class hubness scores via local or global approximative approaches.

$$p(y_i = c|D_k(x_i)) \approx \frac{p(y_i = c) \prod_{t=1}^{k} p(x_{it} \in D_k(x_i)|y_i = c)}{\sum_{c \in C} p(y_i = c) \prod_{t=1}^{k} p(x_{it} \in D_k(x_i)|y_i = c)} \quad (1)$$

– **Strengths:**
  - Generalizes the hw-kNN approach by taking class hubness into account
  - Rephrasing the problem in Bayesian terms allows for further improvements and extensions based on the known ways for improving Bayesian classifiers

– **Weaknesses:**
  - Strong dependencies between occurrences in the same k-neighbor set greatly restrict the applicability of the approach in larger k-neighborhoods
  - Due to these dependencies, class affiliation probabilities tend to be close to 0 or 1 in many cases.
  - Additionally, both weaknesses of h-FNN hold for NHBNN as well

3. The motivation

3.1. Casting a vote: label vs class hubness

Before we delve into the specific ideas behind our proposed approach, the reasons for using the class hubness scores need to be further elucidated. For simplicity, let us begin by focusing on the 1-NN rule. It was already mentioned in the introduction that $p(c|x_i) = \lim_{n\to\infty} p(c|\text{NN}(x_i))$. If the data were not sparse and if there was no overlap between the classes and no noise, 1-NN would work really well. Of course, none of these conditions are met in real world data.

So, what happens is that nearest neighbors sometimes have different labels and this can already be seen on the training set. Observe an illustrative low-dimensional example displayed in Figure 2.

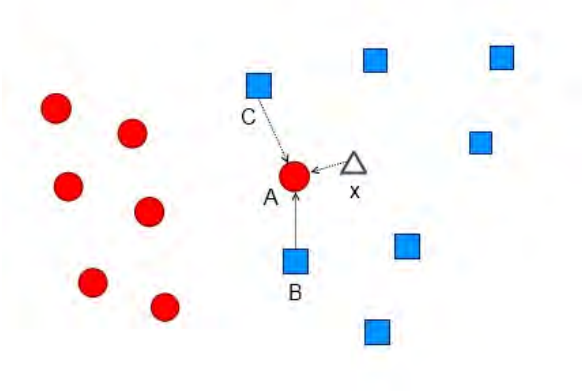


Fig. 2. Illustrative example of a binary classification case. The first class is given by the red circles, the second by the blue squares. A triangle represents an instance yet to be classified. An arrow is drawn from an instance to its nearest neighbor.

The point x is about to be classified. Let's say that the circles represent class 0 and squares represent class 1. According to the 1-NN rule, x would be assigned to class 0, since this is the label of NN$(x) = A$. But, we also have NN$(B) = A$ and NN$(C) = A$, and points B and C are of class 1. If we were to try approximating $p(y = c|A \in D_1(x)) \approx \frac{N_{1,c}(A)}{N_1(A)}$ for $c = 0, 1$, we would get $p(y = 0|A \in D_1(x)) = 0$ and $p(y = 1|A \in D_1(x)) = 1$. So, according to class hubness, x should be assigned to class 1, which seems more plausible when looking at the data.

Two-dimensional data does not exhibit hubness, so Figure 2 can only serve as a simplified model. A more general case is presented in Figure 3. Two examples are given, with class hubness scores shown on the right. In both examples, the label of x is 0 (the red circle).

In the first example, $N_{1,0}(x) = 3$ and $N_{1,1}(x) = 21$, which indicates high bad hubness. Therefore, if x is a neighbor to the point of interest, it is certainly beneficial to base the vote on the class hubness scores, instead of its label. It would reduce the probability of error.

In the second example, $N_{1,0}(x) = 3$ and $N_{1,1}(x) = 4$, which makes for a very small difference in class hubness scores. Even though $N_{1,1}(x) > N_{1,0}(x)$, the label of x is 0, so it is not entirely clear how x should vote. What needs to be evaluated is how much trust should be placed in the neighbor's label and how

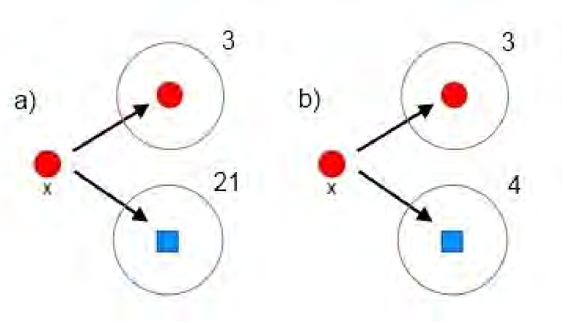


Fig. 3. A more general binary classification case. Class hubness is shown for point x towards both classes. Two examples are depicted, example 'a' where there is a big difference in previous k-occurrences, and example 'b' where there is nearly no observable difference.

much in the occurrence information. If there had been no previous occurrences of x on the training data (an anti-hub), there would be no choice but to use the label. On the other hand, for high hubness points we should probably rely more on their occurrence tendencies. It is precisely the points in between which need to be handled more carefully.

Anti-hubs While discussing the relevance of hubness for kNN classification, we must keep in mind that most points are in fact anti-hubs, when the inherent dimensionality of the data is high. This is illustrated in Figure 4, where the percentage of points exceeding certain k-occurrence thresholds is given. The Dexter data (from the UCI repository) exhibits some hubness, so that even for k as large as 10, there is still around 15% of instances that never occur as neighbors.

Both previously proposed class-hubness based approaches (h-FNN and NHBNN) have failed to provide an easy and consistent way of handling anti-hubs, which is probably their most pronounced weakness. In Section 4 we propose a new way of dealing with such low hubness points.

3.2. Informativeness

The basics What is the information content of an observed event? Intuitively, the more surprised we are about the outcome, the more information the outcome carries. We're all quite used to the sun coming up every morning and by observing this over and over again we don't gain any novel insights. If, however, the sun fails to appear on the sky someday, such a peculiar event would be much more informative, though unfortunate.

This is where information theory comes in. The event self-information is equal to the negative of the logarithm of its probability (i.e. the logarithm of the

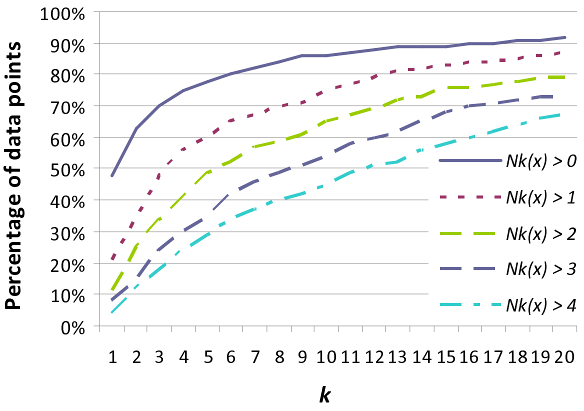


Fig. 4. Percentage of elements with hubness over a certain threshold, for $k = 1$ to $k = 20$ on Dexter data. Each line corresponds to one threshold.

inverse of the probability) [29]. It is often possible to estimate the event probabilities directly by observing the frequencies in previous occurrences, which is what we will be doing with the neighbor points.

Hubs Suppose that there is a data point $x_i \in D$ which appears in *all* k-neighbor sets of other $x_j \in D$. Assume then that we are trying to determine a label of a new data point x and that x_i also appears in this neighborhood, $D_k(x)$. This would not be surprising at all, since x_i appears in *all* other previously observed neighborhoods. Since such an occurrence *carries no information*, x_i should not be allowed to cast a vote. By going one step further, it is easy to see that less frequent occurrences may in fact be more informative and that such neighbors might be more local to the point of interest. This is exploited in our proposed approach.

Going back to the *always-a-neighbor* example, we can see that both the traditional kNN voting scheme and the fuzzy scheme proposed in the h-FNN algorithm fail to handle the extreme case properly. The fact is that whichever point x we observe, $x_i \in D_k(x)$, so there is no correlation between x_i being in $D_k(x)$ and the class affiliation of x. In case of the original kNN procedure, x_i would vote by its label, y_i. If, on the other hand, we were to vote by the class hubness induced fuzziness as in h-FNN, we would in fact be voting by class priors. This is, of course, the lesser evil, but it is still the wrong thing to do. Since there is no information that can be derived from the occurrence of x_i, its vote should be equal to zero.

This scenario does seem quite far-fetched. When reviewing the experimental results, though, it will become clear that such pathological cases are not only theoretically possible - they occasionally take place in real world data.

Anti-hubs Most high-dimensional points are anti-hubs and suppose that x_i is one such point that never occurs in k-neighborhoods on D, i.e. $N_k(x_i) = 0$. Let us say that we are trying to determine the label of a new point x and x_i is found among the neighbors, i.e. $x_i \in D_k(x)$. Such an occurrence would be highly informative. We could be fairly certain that the point x_i carries some important *local* information to the point of interest, since it is not a shared neighbor with many other points.

Of course, not all points are hubs and anti-hubs, as many points will fall somewhere between the two extremes. Any approach designed to handle the informativeness hubs and anti-hubs needs to be applicable to the entire spectrum of possible occurrence frequencies, so that these medium-hubness points are processed in an appropriate way.

It is quite surprising that these simple observations have before gone unnoticed. Previous kNN algorithms have not been taking occurrence informativeness explicitly into consideration.

This is very significant for high dimensional data, where hubs appear. The skewness in the $N_k(x)$ distribution induces the skewness in the distribution of self-information among individual neighbor occurrences. In the following Section we will propose an information-based voting procedure which exploits this fact.

4. The algorithm

Let now x_i be the point of interest, to which we wish to assign a label. Let x_{it}, $t = \{1, 2..k\}$ be its k nearest neighbors. We calculate the informativeness of the occurrences according to Equation 2. In all our calculations, we assume each data point to be its own 0^{th} nearest neighbor, thereby making all $N_k(x_i) \geq 1$. Not only does this give us some additional data, but since it makes all k-occurrence frequencies non-zero, we thereby avoid any pathological cases in our calculations.

$$\begin{aligned} p(x_{it} \in D_k(x_i)) &\approx \frac{N_k(x_{it})}{n} \\ I_{x_{it}} &= \log \frac{1}{p(x_{it} \in D_k(x_i))} \end{aligned} \tag{2}$$

We proceed by defining relative and absolute normalized informativeness. We will also refer to them as *surprise values*.

$$\alpha(x_{it}) = \frac{I_{x_{it}} - \min_{x_j \in D} I_{x_j}}{\log n - \min_{x_j \in D} I_{x_j}}, \quad \beta(x_{it}) = \frac{I_{x_{it}}}{\log n} \tag{3}$$

As we have been discussing, one of the things we wish to achieve is to combine the class information from neighbor labels and their previous occurrences. In order to do this, we need to make one more small observation. Namely, as the number of previous occurrences ($N_k(x_{it})$) increases, two things happen simultaneously. First of all, the informativeness of the current occurrence of x_{it} drops. Secondly, class hubness gives us a more accurate estimate of $p_k(y_i = c|x_{it} \in D_k(x_i))$. Therefore, when the hubness of a point is high, more information is contained in the class hubness scores. Also, when the hubness of a point is low, more information is contained in its label.

$$\bar{p}_k(y_i = c|x_{it} \in D_k(x_i)) = \frac{N_{k,c}(x_{it})}{N_k(x_{it})} = \bar{p}_{k,c}(x_{it})$$

$$p_k(y_i = c|x_{it}) \approx \begin{cases} \alpha(x_{it}) + (1 - \alpha(x_{it})) \cdot \bar{p}_{k,c}(x_{it}), & y_{it} = c \\ (1 - \alpha(x_{it})) \cdot \bar{p}_{k,c}(x_{it}), & y_{it} \neq c \end{cases} \tag{4}$$

The α factor controls how much information is contributed to the vote by the instance label and how much by its previous occurrences. If x_{it} never appeared in a k-neighbor set apart from its own, i.e. $N_k(x_{it}) = 1$, then it votes by its label. If, on the other hand, $N_k(x_{it}) = \max_{x_j \in D} N_k(x_j)$, then the vote is cast entirely according to the class hubness scores.

The fuzzy votes are based on the $p_k(y_i = c|x_{it})$, which are approximated according to Equation 4. These probabilities are then weighted by the absolute normalized informativeness $\beta(x_{it})$. This is shown in Equation 5.

$$u_c(x_i) \propto \sum_{t=1}^{k} \beta(x_{it}) \cdot d_w(x_{it}) \cdot p_k(y_i = c|x_{it}) \tag{5}$$

Additional distance weighting has been introduced for purposes of later comparison with the h-FNN algorithm [26], since it also employs distance weighting. It is not an essential part of the algorithm. We opted for the same distance weighting scheme used in h-FNN, which was in turn first proposed in FNN [14]. It is given in Equation 6.

$$d_w(x_{it}) = \frac{\|x_i - x_{it}\|^{-2}}{\sum_{t=1}^{k} (\|x_i - x_{it}\|^{-2})} \tag{6}$$

Equations 2, 3, 4 and 5 represent our proposed solution for exploiting the information contained in the past k-occurrences on the training data and we will refer to this new algorithm as Hubness Information k-Nearest Neighbor (HIKNN). It embodies some major improvements over the previous approaches:

- Unlike h-FNN and NHBNN, it is essentially parameter-free, one only needs to set the neighborhood size (k).
- Anti-hubs are no longer a special case. They are, however, handled appropriately via the information-based framework.
- Label information is combined with information from the previous k-occurrences, so that both sources of information are exploited for the voting.
- Total occurrence informativeness is taken into account

Algorithm 1 HIKNN: Training

```
Input: ($X$,$Y$,$k$)
training set $T = (X, Y) \subset R^{d\times 1}$
number of neighbors $k \in \{1, 2 \ldots n-1\}$
Train:
kNeighbors = findNeighborSets($T$, $k$)
for all $(x_i, y_i) \in (X, Y)$ do
  $N_k(x_i) = 0$
  for all $c = 1 \ldots C$ do
    count $N_{k,c}(x_i)$
    $N_k(x_i) += N_{k,c}(x_i)$
  end for
  calculate $\alpha(x_i)$ and $\beta(x_i)$ by Eq. 3
  for all $c = 1 \ldots C$ do
    calculate $p_k(y = c|x_i)$ by Eq. 4
  end for
end for
```

The training phase of the algorithm is summarized in (1). The voting is simply done according to (5) and requires no further explanations.

The time complexity of HIKNN, as with all other hubness-based approaches, is asymptotically the same as constructing a kNN graph. Fast algorithms for constructing approximate kNN graphs exist, like the algorithm by [30]. This particular procedure runs in $\Theta(dn^{1+\tau})$ time, where $\tau \in (0, 1]$ is a parameter which is used to set a trade off between speed and graph construction accuracy.

5. Experiments

We compared our proposed HIKNN algorithm to the existing related algorithms: kNN, hw-kNN, h-FNN and NHBNN - on 32 classification problems. We had three test cases: low-to-medium hubness data of lower intrinsic dimensionality, high-hubness textual data and high-hubness image data. In all cases, 10-times 10-fold cross validation was performed. Corrected resampled t-test was used to check for statistical significance. All experiments were performed for $k = 5$, which is a standard choice. Default values described in the respective papers were used for the parameters in h-FNN and NHBNN. The detailed results are given in Table 3 and the basic properties of the datasets are discussed in Table 2.

5.1. The data

Low and medium hubness data Datasets from the well known UCI data repository (http://archive.ics.uci.edu/ml/datasets.html) are usually of low or medium hubness. Since such datasets are less interesting from the perspective of hubness-aware algorithms, we present here the results on a sample of 10 UCI datasets.

The datasets were picked so that they correctly reflect the results on the entire repository. The Manhattan distance was used for this data, as well as for the image data. All features were normalized prior to classification.

Text The Acquis aligned corpus data (http://langtech.jrc.it/JRC-Acquis.html) represents a set of more than 20000 documents in several different languages. In the experiments we only used the English documents. The data was preprocessed and represented as a bag-of-words (term frequencies). On top of this data, 14 different binary classification problems were examined. We used the cosine similarity.

Images We used several datasets in the experiments which were subsets taken from the ImageNet online repository (http://www.image-net.org/). These datasets were selected to match some of the ones used in [28]. All datasets are quantized feature representations. Representations iNet3-iNet7 are based on SIFT features and were also appended color information.

On the other hand, in case of iNet3Err100, iNet3Err150 and iNet3Err1000 - Haar wavelet features were used. These three representations have one interesting property. Due to an I/O error during feature extraction, 5 images were accidentally assigned empty representations (zero vectors). Normally, this would have probably gone unnoticed. In this case, however, the hubness of zero vectors increased drastically with the representation dimensionality. Since all 5 of these points were of the minority class, the classification results were affected greatly and this a prime example of how bad the bad hubness can get in high dimensional data.

5.2. The results

The results in Table 3 show that the hubness-aware algorithms clearly outperform the basic kNN algorithm. Also, HIKNN seems to be the overall best approach, with a clear edge on the textual and UCI data, while performing more or less equal as h-FNN and NHBNN on image datasets. The detailed comparison between the algorithms is shown in Table 1. By comparing both the total number of wins and also the number of wins between pairs of algorithms, we see that HIKNN is to be preferred to the second-best algorithm in the experiments, h-FNN - since it beats it quite convincingly 27(9) : 5(1) in direct comparison and 115(74) : 79(53) overall.

In further comparisons on the image data, we examined the entire range of k-values to see how the algorithms are influenced by neighborhood size. The results on iNet6 are shown in Figure 5. We see that an increase in k separates the algorithms and makes distinctions easier. HIKNN achieves the best results for $k > 5$, where the highest accuracies are achieved. It is not surprising that the accuracy gain over h-FNN increases with k, since the number of large hubs also increases - and the payoff from taking their informativeness into account becomes more substantial. Also, we see that NHBNN simply fails to work

Table 1. Pairwise comparison of classifiers: number of wins (with statistically significant ones in parenthesis)

	kNN	hw-kNN	h-FNN	NHBNN	HIKNN	Total
kNN	–	0 (0)	1 (0)	10 (8)	0 (0)	11 (8)
hw-kNN	32 (27)	–	14 (1)	18 (13)	1 (0)	65 (41)
h-FNN	31 (27)	17 (9)	–	25 (16)	5 (1)	79 (53)
NHBNN	22 (19)	14 (9)	4 (1)	–	3 (1)	43 (30)
HIKNN	**32 (31)**	**29 (14)**	**27 (9)**	**27 (20)**	–	**115 (74)**

when the dependencies between neighbors become too strong, in this case for $k > 15$. The accuracy graphs for the other datasets depict the same general tendencies.

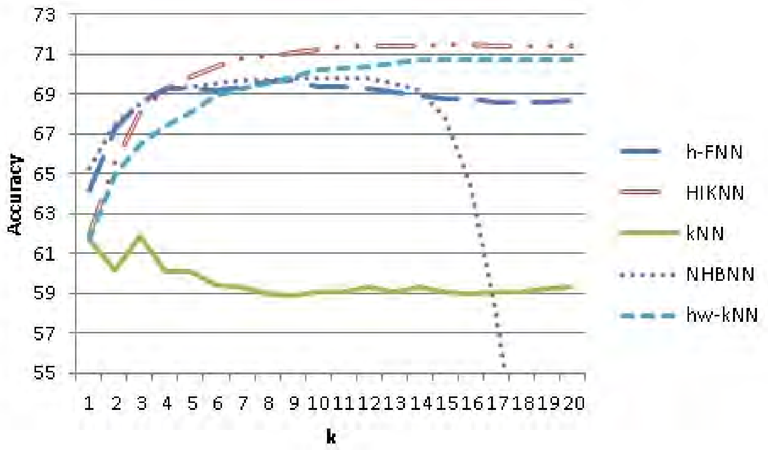


Fig. 5. Classifier accuracies over a range of neighborhood sizes $k \in 1..20$ on iNet6 dataset.

The results for the three iNet3Err representations require special attention. As mentioned in the data description, 5 points in the dataset ended up being zero vectors representing the minority class. We see how an increase in the representation dimensionality causes an amazing increase in bad hubness, which in turn completely disables the basic kNN classifier, as well as the hw-kNN approach. On this 3-category dataset kNN ends up being worse than zero-rule! Keep in mind that such a great drop in accuracy was caused by no more than 5 erroneous instances, out of 2731 total. In the end, 80% of the 5-occurrences were label mismatches. On the other hand, the algorithms based on class hubness: h-FNN, NHBNN and HIKNN - even though affected, retained a much more decent accuracy: 60% compared to the mere 21% by kNN. These five

Table 2. Overview of the datasets. Each dataset is described by its size, dimensionality, the number of categories, skewness of the N_5 distribution (S_{N_5}), proportion of bad 5-occurrences BN_5, as well as the maximal achieved number of occurrences on the dataset.

Data set	size	d	C	S_{N_5}	BN_5	$\max N_5$
dexter	300	20000	2	6.64	30.5%	219
diabetes	768	8	2	0.19	32.3%	14
glass	214	9	6	0.26	35.0%	13
ionosphere	351	34	2	2.06	12.5%	34
isolet1	1560	617	26	1.23	28.7%	30
page-blocks	5473	10	5	0.31	5.0%	16
segment	2310	19	7	0.33	5.3%	15
sonar	208	60	2	1.28	21.3%	22
vehicle	846	18	4	0.64	36.0%	14
vowel	990	10	11	0.60	9.7%	16
Acquis1	23412	254963	2	62.97	19.2%	4778
Acquis2	23412	254963	2	62.97	8.7%	4778
Acquis3	23412	254963	2	62.97	27.3%	4778
Acquis4	23412	254963	2	62.97	12.2%	4778
Acquis5	23412	254963	2	62.97	5.7%	4778
Acquis6	23412	254963	2	62.97	7.6%	4778
Acquis7	23412	254963	2	62.97	18.1%	4778
Acquis8	23412	254963	2	62.97	9.3%	4778
Acquis9	23412	254963	2	62.97	7.6%	4778
Acquis10	23412	254963	2	62.97	21.4%	4778
Acquis11	23412	254963	2	62.97	23.4%	4778
Acquis12	23412	254963	2	62.97	9.8%	4778
Acquis13	23412	254963	2	62.97	16.4%	4778
Acquis14	23412	254963	2	62.97	6.9%	4778
iNet3Err100	2731	100	3	20.56	10.2%	375
iNet3Err150	2731	150	3	25.1	34.8%	1280
iNet3Err1000	2731	1000	3	23.3	79.7%	2363
iNet3	2731	416	3	8.38	21.0%	213
iNet4	6054	416	4	7.69	40.3%	204
iNet5	6555	416	5	14.72	44.6%	469
iNet6	6010	416	6	8.42	43.4%	275
iNet7	10544	416	7	7.65	46.2%	268

Table 3. Overview of the experiments. Classification accuracy is given for kNN, hubness-weighted kNN (hw-kNN), hubness-based fuzzy nearest neighbor (h-FNN) and hubness information k-nearest neighbor (HIKNN). All experiments were performed for $k = 5$. The symbols •/○ denote statistically significant worse/better performance ($p < 0.05$) compared to HIKNN. The best result in each line is in bold.

Data set	kNN	hw-kNN	h-FNN	NHBNN	**HIKNN**
dexter	57.2 ± 7.0 •	67.7 ± 5.4	67.6 ± 4.9	**68.0** ± **4.9**	**68.0** ± **5.3**
diabetes	67.8 ± 3.7 •	75.6 ± 3.7	75.4 ± 3.2	73.9 ± 3.4	**75.8** ± **3.6**
glass	61.5 ± 7.3 •	65.8 ± 6.7	67.2 ± 7.0	59.1 ± 7.5 •	**67.9** ± **6.7**
ionosphere	80.8 ± 4.5 •	87.9 ± 3.6	90.3 ± 3.6 ○	**92.2** ± **3.2** ○	87.3 ± 3.8
isolet1	75.2 ± 2.5 •	82.5 ± 2.1 •	83.8 ± 1.8 •	83.0 ± 2.0 •	**86.8** ± **1.5**
page-blocks	95.1 ± 0.6 •	95.8 ± 0.6 •	96.0 ± 0.6	92.6 ± 0.6 •	**96.2** ± **0.6**
segment	87.6 ± 1.5 •	88.2 ± 1.3 •	88.8 ± 1.3 •	87.8 ± 1.3 •	**91.2** ± **1.1**
sonar	82.7 ± 5.5	83.4 ± 5.3	82.0 ± 5.8	81.1 ± 5.6 •	**85.3** ± **5.5**
vehicle	62.5 ± 3.8 •	65.9 ± 3.2	64.9 ± 3.6 •	63.7 ± 3.5 •	**67.2** ± **3.6**
vowel	87.8 ± 2.2 •	88.2 ± 1.9 •	91.0 ± 1.8 •	88.1 ± 2.2 •	**93.6** ± **1.6**
Acquis1	78.7 ± 1.0 •	87.5 ± 0.8 •	88.8 ± 0.7	88.4 ± 0.7 •	**89.4** ± **0.6**
Acquis2	92.4 ± 0.5 •	93.6 ± 0.5	93.3 ± 0.5	92.5 ± 0.5 •	**93.7** ± **0.5**
Acquis3	72.7 ± 0.9 •	78.7 ± 0.9	79.5 ± 0.9	78.9 ± 0.9	**79.6** ± **0.9**
Acquis4	89.8 ± 0.6 •	90.6 ± 0.6	90.5 ± 0.6	87.4 ± 0.7 •	**91.0** ± **0.5**
Acquis5	97.3 ± 0.3 •	97.6 ± 0.3	97.5 ± 0.3	95.1 ± 0.4 •	**97.7** ± **0.3**
Acquis6	93.6 ± 0.4 •	94.4 ± 0.5	94.0 ± 0.5 •	92.5 ± 0.5 •	**94.6** ± **0.5**
Acquis7	82.9 ± 0.8 •	86.3 ± 0.7 •	86.1 ± 0.6 •	85.7 ± 0.7 •	**87.0** ± **0.7**
Acquis8	92.3 ± 0.5 •	93.0 ± 0.5	93.1 ± 0.5	91.0 ± 0.5 •	**93.5** ± **0.5**
Acquis9	93.0 ± 0.5 •	**94.8** ± **0.4**	94.2 ± 0.5 •	93.4 ± 0.5 •	**94.8** ± **0.4**
Acquis10	83.1 ± 1.6 •	88.8 ± 0.7 •	88.7 ± 0.6 •	87.4 ± 0.7 •	**89.7** ± **0.5**
Acquis11	77.7 ± 0.9 •	81.8 ± 0.8	82.4 ± 0.6	81.9 ± 0.7	**82.5** ± **0.5**
Acquis12	91.9 ± 0.6 •	**92.8** ± **0.5**	92.6 ± 0.5	90.7 ± 0.6 •	**92.8** ± **0.5**
Acquis13	85.6 ± 0.7 •	87.5 ± 0.6	87.1 ± 0.7 •	85.2 ± 0.7 •	**88.0** ± **0.7**
Acquis14	94.2 ± 0.4 •	94.9 ± 0.4	94.6 ± 0.5	92.5 ± 0.5 •	**95.0** ± **0.5**
iNet3Err100	92.4 ± 0.9 •	93.6 ± 0.9 •	97.5 ± 0.9	97.5 ± 0.9	**97.6** ± **0.9**
iNet3Err150	80.0 ± 2.0 •	88.7 ± 2.0 •	94.6 ± 0.9	94.6 ± 0.9	**94.8** ± **0.9**
iNet3Err1000	21.2 ± 2.0 •	27.1 ± 11.2 •	59.5 ± 3.2	**59.6** ± **0.9**	**59.6** ± **3.2**
iNet3	72.0 ± 2.7 •	80.8 ± 2.3	**82.4** ± **2.2**	81.8 ± 2.3	82.2 ± 2.0
iNet4	56.2 ± 2.0 •	63.3 ± 1.9 •	**65.2** ± **1.7**	64.6 ± 1.9	64.7 ± 1.9
iNet5	46.6 ± 2.0 •	56.3 ± 1.7 •	**61.9** ± **1.7**	61.8 ± 1.9	60.8 ± 1.9
iNet6	60.1 ± 2.2 •	68.1 ± 1.6 •	69.3 ± 1.7	69.4 ± 1.7	**69.9** ± **1.9**
iNet7	43.4 ± 1.7 •	55.1 ± 1.5 •	**59.2** ± **1.5**	58.2 ± 1.5	56.9 ± 1.6
AVG	76.72	81.13	83.09	81.86	**83.60**

points occur in nearly all k-neighborhoods and this dataset shows how some pathological cases of very bad hubness also occasionally emerge in practical situations. Even if the erroneous points were not of the minority class, they would still have caused significant misclassification. Also, note that the major hub in the 1000-dimensional case appears in 86.5% of all k-neighbor sets. Its occurrence is, therefore, not very informative - and this further justifies the discussion presented in Section 3.2.

Bad hubness of the data is closely linked to the error of the kNN classification. The Pearson correlation coefficient comparing the kNN error with bad hubness percentages on the datasets in our experiments gives 0.94, which indicates strong positive correlation between the two quantities. HIKNN bases its votes on expectations derived from the previous k-occurrences, so it is encouraging that the correlation between the accuracy gain over kNN and bad hubness of the data is also very strong: 0.87 according to the Pearson coefficient.

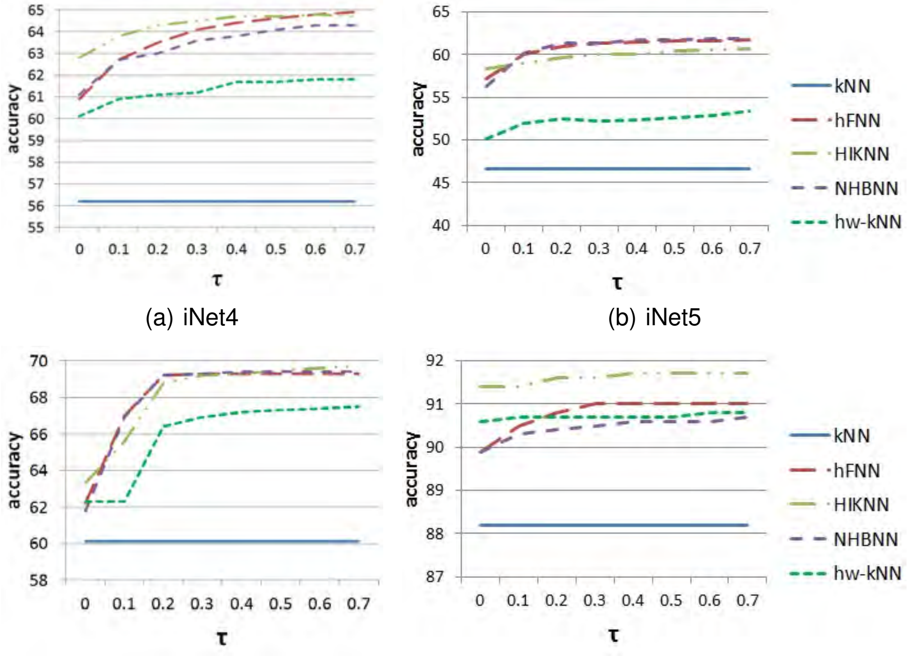
6. The approximate implementation

Computing all the k-neighbor sets on the training data in order to build an occurrence model could be overly time-consuming in large-scale data collections. Hubness-aware approaches would be applicable in large-scale scenarios only if it were possible to retain the previously observed improvements while working with some sort of approximate kNN sets.

Many approximate kNN algorithms have been proposed in the literature, either for speeding-up individual queries or constructing an entire kNN graph. It is the latter that is of interest for building an occurrence model. Many of these procedures had been proposed specifically for handling high-dimensional data, which is where hubness-aware classification has been shown to be useful.

In our experiments we focused on one such approach [30]. It is a divide and conquer method based on recursive Lanczos bisection. The time complexity of the procedure is $\Theta(dn^{1+\tau})$, where $\tau \in (0, 1]$ reflects the quality of the approximation. There are two ways to implement the recursive division and we have chosen the GLUE method, as it has proven to be significantly faster than the OVERLAP method, though the quality of the resulting graph is only slightly inferior in comparison [30]. The question that we would like to answer is: *for which values of τ are we able to retain the improvements observed on actual kNN sets*?

Re-running all the experiments for all τ values would be beyond the scope of this paper. We did, however, examine the full spectrum of τ-values for the four datasets previously used in the experiments. We report the results for the iNet4, iNet5, iNet6 and Acquis1 datasets in Figure 6. The original Acquis data had too many features for our approximate kNN graph implementation to be able to handle it properly in reasonable time, so we considered a projection onto a 400-dimensional feature space. The data was projected via canonical correlation analysis procedure onto a common semantic space obtained by correlating the



(a) iNet4

(b) iNet5

(c) iNet6

(d) Acquis1 projection

Fig. 6. The accuracy of the hubness-aware approaches when the occurrence model is inferred from the approximate kNN graph generated by [30]. We see that there are significant improvements even for $\tau = 0$.

English and French aligned versions of documents from the dataset [31] [32]. It is one of the standard dimensionality reduction techniques used in text mining and its details are beyond the scope of this paper.

The results shown in Figure 6 are indeed very encouraging. They suggest that significant improvements over the kNN baseline are possible even when the graph is constructed in linear time (w.r.t. number of instances). Moreover, the quality level of $\tau = 0.2$ or $\tau = 0.3$ already seems good enough to capture most of the original occurrence information, as the resulting accuracies are quite close to the ones achieved in the original experiments.

The accuracy curves for different algorithms sometimes intersect. This can be seen for iNet5, iNet6 and Acquis1 in Figure 6. In general, the approximate results correspond rather well to the non-approximate results, but the correlation between the two can vary depending on the particular choice of τ.

In these initial findings HIKNN appears to be quite robust to the employed approximate kNN graph construction method for $\tau = 0$. This is a very nice property, as it allows for obtaining usable results in reasonable time. If better approximations are required, $\tau = 0.3$ should suffice.

A comparison between the results shown in Figure 6(d) and those previously summarized in Table 3 reveals that dimensionality reduction may sometimes significantly affect the classification process and improve the overall classification accuracy. Even though hubness is practically unavoidable in most high-dimensional data mining tasks, its severity does depend on the particular choice of feature representation and/or similarity measure. It is, therefore, not surprising that the dimensionality reduction of the Acquis data helped the kNN classifiers by reducing data hubness. The hubness was not entirely eliminated and this is why all the hubness-aware classification methods still managed to outperform the kNN baseline for all the τ values.

These initial experiments suggest that hubness-aware methods are applicable even to large datasets, as the scalable, approximate kNN graph construction methods are able to deliver good hubness estimates. More experiments are needed to reach the final verdict, on different types of high-dimensional data.

7. Estimating class probabilities

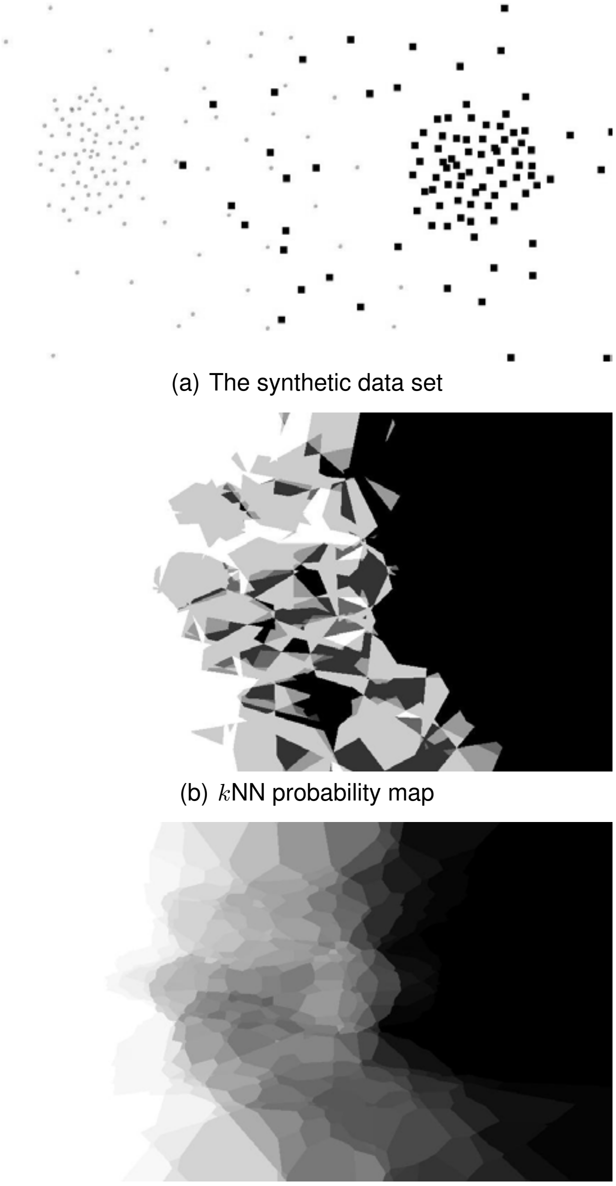
Most frequently, in classification, we are simply interested in assigning a label to a point of interest. What this label suggests is that we are entirely certain that a point belongs to a given class. However, this is just a special case of a more general problem. We would in fact like to be able to assign a 'fuzzy' label to each object, so that it belongs to several classes at the same time. This 'belonging' marks our confidence in any particular atomic label choice.

There are cases, however, when the classes overlap. This happens very frequently in real-world data. There exist points then, in these overlapping regions, that could belong to either of the neighboring categories. In such cases it is meaningless to assign a simple 'crisp' label to each point - what we would like to be able to do is to predict the actual class probability at each point, for every given class. This probability reflects the relative density of each class probability distribution at that point.

The HIKNN algorithm was made to be fuzzy and in the following experiments we wished to determine how well the predicted class probabilities reflect our intuition about the data. The basic kNN algorithm can also be used for point class probability estimates and it is a useful baseline for comparison.

In order to check if the predicted values make sense or not, we examined the algorithm output on synthetic 2D data. The fact that data has only 2 dimensions allows us to draw a *probability map*, where each pixel is 'classified' by the examined algorithms and assigned a probability of belonging to each class. We have generated several such datasets and here we discuss one of them. The dataset is simple, representing 2 categories with overlapping border regions. We have used HIKNN without the distance weighting. The resulting probability maps can be seen in Figure 7.

We see that the probability map generated by HIKNN looks much more natural in the overlapping region. The gradient between the classes should be more or less smooth if the model is able to generalize well. kNN produces a fractured



(a) The synthetic data set

(b) kNN probability map

(c) HIKNN probability map

Fig. 7. Probability maps inferred from kNN and HIKNN on synthetic data, for $k = 5$. Each pixel was classified by the algorithms and assigned a probability value of belonging to each of the two classes.

landscape, essentially over-fitting on the training data. These maps suggest that the votes based on previous occurrences may offer better estimates of the underlying class probabilities, which we intend to explore more thoroughly in our future work.

8. Conclusion

In this paper we presented a novel approach for handling high-dimensional data in k-NN classification, Hubness Information k-Nearest Neighbor (HIKNN). It is a hubness-aware approach which is based on evaluating the informativeness of individual neighbor occurrences. Rare neighbors (anti-hubs) are shown to carry valuable information which can be well exploited for classification.

The algorithm is parameter-free, unlike the previous class-hubness based hubness-aware classification algorithms. The danger of over-fitting is thereby greatly reduced.

The algorithm was compared to the three recently proposed hubness-aware approaches (hw-kNN, h-FNN, NHBNN), as well as the kNN baseline on 32 classification problems. Our proposed approach had an overall best performance in the experiments.

Since HIKNN modifies only the voting, it is easily extensible and could be combined with various sorts of metric learning or dynamic k-neighbor sets. We intend to explore these directions thoroughly in our future work.

Acknowledgments. This work was supported by the Slovenian Research Agency, the IST Programme of the EC under PASCAL2 (IST-NoE-216886) and the Serbian Ministry of Education and Science project no. OI174023.

References

1. I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann Publishers, 2005.
2. E. Fix and J. Hodges, “Discriminatory analysis, nonparametric discrimination: consistency properties,” USAF School of Aviation Medicine, Randolph Field, Texas, Tech. Rep., 1951.
3. T.M.Cover and P.E.Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. IT-13, no. 1, pp. 21–27, 1967.
4. L. Devroye, “On the inequality of cover and hart,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 3, pp. 75–78, 1981.
5. C.J.Stone, “Consistent nonparametric regression,” *Annals of Statistics*, vol. 5, pp. 595–645, 1977.
6. A. K. L. Devroye, L. Gyorfi and G. Lugosi, “On the strong universal consistency of nearest neighbor regression function estimates,” *Annals of Statistics*, vol. 22, pp. 1371–1385, 1994.
7. D. François, V. Wertz, and M. Verleysen, “The concentration of fractional distances,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 7, pp. 873–886, 2007.
8. C. C. Aggarwal, A. Hinneburg, and D. A. Keim, “On the surprising behavior of distance metrics in high dimensional spaces,” in *Proc. 8th Int. Conf. on Database Theory (ICDT)*, 2001, pp. 420–434.

9. M. Radovanović, A. Nanopoulos, and M. Ivanović, “Nearest neighbors in high-dimensional data: The emergence and influence of hubs,” in *Proc. 26th Int. Conf. on Machine Learning (ICML)*, 2009, pp. 865–872.
10. N. Tomašev, M. Radovanović, D. Mladenić, and M. Ivanović, “The role of hubness in clustering high-dimensional data.” in *PAKDD (1)’11*, 2011, pp. 183–195.
11. E.-H. Han, G. Karypis, and V. Kumar, “Text categorization using weight adjusted k-nearest neighbor classification,” in *Advances in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science, D. Cheung, G. Williams, and Q. Li, Eds. Springer Berlin / Heidelberg, 2001, vol. 2035, pp. 53–65.
12. J. Wang, P. Neskovic, and L. N. Cooper, “Improving nearest neighbor rule with a simple adaptive distance measure,” *Pattern Recogn. Lett.*, vol. 28, pp. 207–213, January 2007.
13. Y. Song, J. Huang, D. Zhou, H. Zha, and C. L. Giles, “Iknn: Informative k-nearest neighbor pattern classification,” in *Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, ser. PKDD 2007. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 248–264.
14. J. E. Keller, M. R. Gray, and J. A. Givens, “A fuzzy k-nearest-neighbor algorithm,” in *IEEE Transactions on Systems, Man and Cybernetics*, 1985, pp. 580–585.
15. R. Jensen and C. Cornelis, “A new approach to fuzzy-rough nearest neighbour classification,” in *Rough Sets and Current Trends in Computing*, ser. Lecture Notes in Computer Science, C.-C. Chan, J. Grzymala-Busse, and W. Ziarko, Eds. Springer Berlin / Heidelberg, 2008, vol. 5306, pp. 310–319.
16. W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and H. Dong, “An adaptive fuzzy knn text classifier,” in *Computational Science — ICCS 2006*, ser. Lecture Notes in Computer Science, V. Alexandrov, G. van Albada, P. Sloot, and J. Dongarra, Eds. Springer Berlin / Heidelberg, 2006, vol. 3993, pp. 216–223.
17. L. Wang, L. Khan, and B. Thuraisingham, “An effective evidence theory based k-nearest neighbor (knn) classification,” in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 797–801. [Online]. Available: http://portal.acm.org/citation.cfm?id=1486927.1487026
18. K. Q. Weinberger, J. Blitzer, and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” in *In NIPS*. MIT Press, 2006.
19. M. R. Min, D. A. Stanley, Z. Yuan, A. J. Bonner, and Z. Zhang, “A deep non-linear feature mapping for large-margin knn classification,” in *ICDM*, 2009, pp. 357–366.
20. J. Aucouturier and F. Pachet, “Improving timbre similarity: How high is the sky?” *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, 2004.
21. J. Aucouturier, “Ten experiments on the modelling of polyphonic timbre,” Docteral dissertation, University of Paris 6, Tech. Rep., 2006.
22. M. Radovanović, A. Nanopoulos, and M. Ivanović, “Hubs in space: Popular nearest neighbors in high-dimensional data,” *Journal of Machine Learning Research*, vol. 11, pp. 2487–2531, 2011.
23. ——, “On the existence of obstinate results in vector space models,” in *Proc. 33rd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2010, pp. 186–193.
24. K. Buza, A. Nanopoulos, and L. Schmidt-Thieme, “Insight: efficient and effective instance selection for time-series classification,” in *Proceedings of the 15th Pacific-Asia conference on Advances in knowledge discovery and data mining - Volume Part II*, ser. PAKDD’11. Springer-Verlag, 2011, pp. 149–160.
25. N. Tomašev and D. Mladenić, “Exploring the hubness-related properties of oceanographic sensor data,” in *Proceedings of the SiKDD conference*, 2011.

26. N. Tomašev, M. Radovanović, D. Mladenić, and M. Ivanović, “Hubness-based fuzzy measures for high dimensional k-nearest neighbor classification,” in *Machine Learning and Data Mining in Pattern Recognition, MLDM conference*, 2011.
27. ——, “A probabilistic approach to nearest neighbor classification: Naive hubness bayesian k-nearest neighbor,” in *Proceeding of the CIKM conference*, 2011.
28. N. Tomašev, R. Brehar, D. Mladenić, and S. Nedevschi, “The influence of hubness on nearest-neighbor methods in object recognition,” in *IEEE Conference on Intelligent Computer Communication and Processing*, 2011.
29. D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. New York, NY, USA: Cambridge University Press, 2002.
30. J. Chen, H. ren Fang, and Y. Saad, “Fast approximate kNN graph construction for high dimensional data via recursive Lanczos bisection,” *Journal of Machine Learning Research*, vol. 10, pp. 1989–2012, 2009.
31. D. R. Hardoon, S. Szedmák, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.
32. H. Hotelling, “The most predictable criterion,” *Journal of Educational Psychology*, vol. 26, pp. 139–142, 1935.

Nenad Tomašev is a PhD student at the Artificial Intelligence Laboratory at J. Stefan Institute. He graduated in 2008 at the Department for Mathematics and Informatics at the University of Novi Sad, Serbia. His research topics include Machine Learning, Data Mining, Artificial Life and Stochastic Optimization. He has been actively participating in teaching at both Petnica Science Center and Višnjan Summer School.

Prof. Dr. Dunja Mladenić works as a researcher and a project manager at J. Stefan Institute, leading Artificial Intelligence Laboratory and teaching at J. Stefan International Postgraduate School. She is an expert on study and development of Machine Learning, Data/Text Mining, Semantic Technology techniques and their application on real-world problems. She is associated with the J. Stefan Institute since 1987. She got her MSc and PhD in Computer Science at University of Ljubljana in 1995 and 1998 respectively. She was a visiting researcher at School of Computer Science, Carnegie Mellon University, USA in 1996-1997 and in 2000-2001. Dunja Mladenić has experience in coordinating EU projects and acting on management board of several European research and development projects. She was a program co-chair of ECML 2007, a general chair of ECMLPKDD 2009. She is a co-inventor on a patent application with Cycorp Inc., USA.

Dunja Mladenić is the Slovenian representative in EC Enwise STRATA ETAN Expert Group ”Promoting women scientists from the Central and Eastern European countries and the Baltic States to produce gender equality in science in the wider Europe”. She serves as project evaluator of project proposals for EC programme on Information and Society Technology (IST). In 2001, she was evaluator of project proposals for National Science Foundation (NSF) initiative on Information Technology Research (ITR), NSF 00-126, USA. She has published papers in refereed journals and conferences, co-editor of several books, served in the program committee of international conferences and organized international events in the area of Text Mining, Link Analysis and Data Mining.

Received: December 11, 2011; Accepted: June 5, 2012.

DOI: 10.2298/CSIS110722008R

AKNOBAS: A Knowledge-based Segmentation Recommender System based on Intelligent Data Mining Techniques

Alejandro Rodríguez-González[1], Javier Torres-Niño[1], Enrique Jimenez-Domingo[1], Juan Miguel Gomez-Berbis[1] and
Giner Alor-Hernandez[2]

[1] Computer Science Department, University Carlos III of Madrid,
28911, Leganes, Madrid, Spain
{alejandro.rodriguez, javier.torres, enrique.jimenez, juanmiguel.gomez}@uc3m.es
[2] Division of Research and Postgraduate Studies, Instituto Tecnológico de Orizaba, 94320, Orizaba, Veracruz, México
galor@itorizaba.edu.mx

Abstract. Recommender Systems have recently undergone an unwavering improvement in terms of efficiency and pervasiveness. They have become a source of competitive advantage in many companies which thrive on them as the technological core of their business model. In recent years, we have made substantial progress in those Recommender Systems outperforming the accuracy and added-value of their predecessors, by using cutting-edge techniques such as Data Mining and Segmentation. In this paper, we present AKNOBAS, a Knowledge-based Segmentation Recommender System, which follows that trend using Intelligent Clustering Techniques for Information Systems. The contribution of this Recommender System has been validated through a business scenario implementation proof-of-concept and provides a clear breakthrough of marshaling information through AI techniques.

Keywords: Data Mining, Clustering, Information Systems, Artificial Intelligence, Use Case

1. Introduction

BeRuby.com, one of the largest cash back companies in Europe, offers Internet users a personalized portal where they are paid for their activity on the Web, as well as for the activity of friends they invite into their BeRuby network. Users can earn money for purchases, registration and even website visits from over 600 advertisers. Users can also create a custom homepage with all of their favorite links on one page. Currently BeRuby.com operates in nine countries (Spain, Italy, France, Germany, Portugal, The Netherlands, The United Kingdom, The United States and Brazil) with more than 800,000 registered users worldwide. BeRuby.com has already paid more than $1,000,000 to its users for their Internet activity. BeRuby.com databases store a huge amount of raw information about these 800K+ users: their navigation habits (sites they visit, or where they register or purchase), user preferences, and profile information. However, the information about users is not capitalized for helping to BeRuby administrators to find groups of users which share things in common. For hence, all the information about the behavior of the users which users the BeRuby platform can be analyzed with the system represented in this paper to find patterns which allows to find groups of users with some interests in common. With these groups we are able to cover the necessity of creating personalized advertisements given the users interests and its relation with other users. Under this scenario, the search of patterns which allows finding groups of users with some interests in common can be considered as a clustering issue. Clustering is one of the most important unsupervised learning problems. A common definition of clustering can be "the process of organizing objects into groups whose members are similar in some way". Another way to say it is ensuring that a cluster is a group of objects which are "similar" between them and are "dissimilar" to the objects that belong to other clusters [1].

The goal of this work is to show the design and implementation of a platform which aims to help BeRuby administrators find groups of users which share things in common. This paper is mainly focused on the design and implementation of clustering techniques to obtain these groups of users, and the analysis of the results obtained by the clustering method used. We have selected a clustering algorithm because satisfies some properties to ensure its performance: scalability, discovering clusters with an arbitrary shape, capacity to deal with noise and uncertainty, high dimensionality, dealing with different kinds of attributes and some requirements for domain knowledge for the determination of input parameters [2]. The algorithm presented here, which is detailed in following sections, fulfills these features and provides interesting results for recommendation systems.

The remainder of the paper is organized as follows. Section 2 outlines related research in the area, Section 3 describes the proposed system, dividing the section into sub sections to explain the different parts of the proposed system. In Section 4, an analysis of the results provided by clustering algorithms is done in order to observe how these results can be interpreted to use them in the recommendation system. In Section 5, a concrete use case of this analysis is presented to see the functioning of the system. Section 6 will show the evaluation of the system using the use case presented in the previous section. This way it is possible to illustrate the accuracy by using as a measurement how useful the classification of the users into profiles is. Finally, Section 7 presents the conclusions and future work.

2. Related Works

Nowadays, clustering is a well-known technique that it is applied to different fields: marketing [3], graphics [4], insurance [5], health [6], biology [7], and classification [8], to mention a few. In the literature, some efforts have applied the clustering algorithm for developing recommendation systems. For instance, in [9], a hotel recommendation system based on clustering and Rankboost algorithm was proposed. This proposal tries to avoid the cold-start and scalability. An improved grid portal recommendation architecture is presented in [10]. The proposed architecture in combination with a clustering algorithm allows solving problems such as over-scale of the grid portal resources management, heavy-load of handing with large-scale querying and processing, low-satisfaction of the users who need access to get the desired the resources. In addition, the paper implements the architecture efficiently through a prototype portal in which both, action layer and render layer, are designed for collaborative filtering. An exploration of how to utilize tagging information to do personalized recommendations is presented in [11]. Based on the distinctive three dimensional relationships among users, tags and items, a new user profiling and similarity measure method is proposed. The experiments suggest that the proposed approach is better than the traditional collaborative filtering recommendation systems using only rating data. In [12], an approach that combines the advantages of memory-based and model-based collaborative filtering by joining the two methods is presented. Firstly, it employs memory-based collaborative filtering to fill the vacant ratings of the user-item matrix. Then, it uses the item-based collaborative filtering as model-based to form the nearest neighbors of every item. At last, it produces prediction of the target user to the target item at real time. In [13], a recommendation algorithm based on the item classification to pre-produce the ratings is proposed. This approach classifies the items to predict the ratings of the vacant values where necessary, and then uses the item-based collaborative filtering to produce the recommendations. An extension of the collaborative filtering approach to design a more effective recommendation system that overcomes those limitations is proposed in [14].

Other works have been proposed centered on the classification of environmental situations such as [15] and [16]. In [15], this work is focused on supervised sea-ice classification in Polar Regions based on fuzzy clustering that allows the system to divide Polar Regions depending on the sea-ice types, while in [16] a segmentation of satellite images and probabilistic methods are used for establishing a cloud classification through self-organizing maps. The value of these algorithms has been widely proven and the results obtained offered enough certainty to enable them to be used in the problem described in this paper. Due to that fact, there are some systems in which clustering techniques are used for recommendation tasks. It is possible to find some examples in [17], where the system suggests products to supermarket shoppers depending on the rest of the products bought using data mining and clustering, [18] recommendation of social tags, which allows noise to be reduced, and identifies trends, and some applications in e-commerce [19] or focuses on the recommendation of web pages based on the interests of the user [20]. In other terms, it is easy to find systems where the main goal is to determine the most suitable advertisement for a concrete user. This is the case of [21], which uses decision-tree induction techniques for offering products in storefronts after a process of marketing rule-extraction, or [22], whose aim is to advise the user in shopping areas based on his profiles and interests by using a neural based planner which identifies the most adequate plan for a given user.

The scope of these kinds of systems has represented important achievements, that even lead to the publication of patents as in [23], where a method for delivering customized electronic advertisements in an interactive communication system is explained, or in [24], a similar system whose aim is also to provide customized advertisings using a repository connected to the World Wide Web and a set of preferences taken from the user. Without a doubt these applications are more and more common every day due to the necessity of adaptation to market pressures and the evolution of society to more customized products in all ambits.

These initiatives suffer from several drawbacks such as: a) require prior knowledge on how to behave under each situation, b) lack of discovering clusters with an arbitrary shape c) lack of different kind of attributes and some requirements for domain knowledge for the determination of input parameters. Our proposal tries improving these aforementioned deficiencies.

3. AKNOBAS

AKNOBAS (Automated Knowledge-Based Segmentation System for Recommendations), which is the system described in this paper, has been developed to build a recommendation system [25] with an advertising aim based on the information stored in the BeRuby database [26]. The objective of the system is to generate several profiles which allow the users of the company to create personalized advertisements for BeRuby's users.

The information stored in the BeRuby database is composed of several data about the users' preferences, visited websites and some other information associated with their profiles. The objective of the system, thus, is to try to process this raw information, and extract groups or users profiles, in such a way that once we have these profiles, we can determine some preferences of a concrete user given that the user belongs to a particular group.

In artificial intelligence there are several tools or techniques that can be used for the aim of the system developed. One technique is data mining [27], the process of extracting patterns from data. Given the nature of the project, the authors consider this process to be the key to achieving the project's goals.

Within artificial intelligence techniques, clustering [28], which is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense, is probably the best approach that can be used.

The following subsections describe the architecture of the generated system and the internal working of their modules as well as the main problems that authors found during its development, and the design decisions made.

3.1. Solution approach

The solution to the use case or problem that we wish to achieve can be reached through the generation of a system which, using clustering as a data mining technique is able to generate groups of users with common commercial features. Commercial features are understood to be those that are representative in the group of data handled.

3.2. Architecture

The solution has a layered design in order to organize its components. This layered-design allows scalability and easy maintenance because its tasks and responsibilities are distributed.

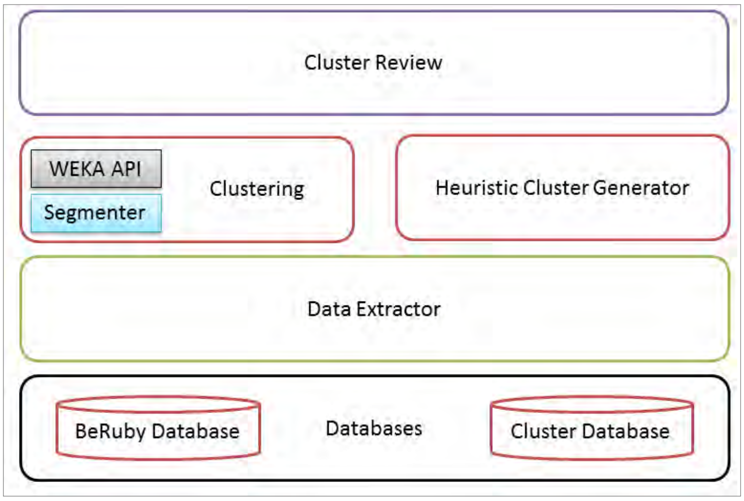


Fig. 1. AKNOBAS architecture

The general architecture is shown in Figure 1. Each component has a function explained as follows:

- **BeRuby DB**: The component BeRuby DB represents the mechanisms of data persistence executing insert, update, delete and query operations in the AKNOBAS's architecture. It contains all the data of the users that are going to be used by AKNOBAS in order to generate the groups (clusters) of users. It also can store the information of the clusters once they have been generated and validated.

- **Data Extractor**: This component obtains all the data which is stored in the BeRuby database through SQL queries with the objective that the following module can build a segmentation of the data what the Heuristic Cluster Generator module can access the data.
- **Heuristic Cluster Generator**: This module generates clusters in a semi-automatic way. The clusters generated by this module are pre-established, and the objective of this module is to generate the groups that the system users (BeRuby employees) have identified through experience. The generation of these groups is based on the application of some heuristics, and thus, these groups always are the same, with the only difference being that the users contained within will change. In this case, the information generated by this module can be directly stored in the database because it is not necessary to supervise the content of the groups as we already know their structure.
- **Segmenter**: This module makes segmentation of all the information that has been extracted by the module Data Extractor. The segments are generated or defined by the BeRuby workers based on their experience in the internal structure of the data that is stored in their databases. They can generate segments of information because they know the possible relation that exists between the tables as experts in the domain. An example of segment is the creation of a segment which contains private information of a user. This information can be his/her age, sex, or similar, with the objective of generating a segment that contains private information about the users. This process provokes the creation of clusters based on such private information.
- **Clustering (WEKA)**: The clustering module generates a cluster using the WEKA framework as clustering API. This module has direct access to the BeRuby database. The reason for this access is because WEKA API implements the possibility of access to a database directly in order to generate the data matrix (individuals) that is used in the clustering process. This module requires extra attention and it is explained with more detail in the following sections.
- **Cluster Review**: The revision module is a item of the architecture, which is composed of two parts:
  1) The first part consists of a manual review of the clusters that are normally generated by a certain segment. The idea is that the expert in the domain, in this case, BeRuby experts, would analyze the content of the cluster that has been generated by the cluster algorithm. With this analysis they are able to interpret the content of the data contained in the cluster, and therefore, be able to generate a certain algorithm that establishes order in the chaos of the results returned by the clustering module.

2) The second part consists of automatic execution of the algorithm that the experts of the domain should design in order to allow the processing of the results given by the clustering algorithm for a certain segment. With the execution of this algorithm, the data is interpreted and it is easier to process the data from the recommender system.

- **Cluster DB**: This module is a database designed to store the information about the clusters that have been generated and processed by the Cluster Review module. The idea is to use this database as a temporary knowledge warehouse.

3.3. Information about BeRuby data

BeRuby stores a lot of data in its databases, around 4GB in binary form for the entire UK version, and 25GB just in the click_account table of the spanish version. The data is distributed in dozens of tables including information for all the widgets (interactive elements), categories, users, clicks of users in the widgets, etc. Only some of this information is relevant to the analysis carried out in this paper.

The data is divided in segments that are a set of fields from one or multiple tables to be taken as a whole for clustering purposes. AKNOBAS automatically creates the required queries to select the data, including joining data from different tables.

The most important segment only has two fields from a single table; the user_id and the widget_id from the click_accounts table. In these fields each row represents a click of one user in a widget.

Some more data is required for the analysis of the output of the clustering system, as a widget id does not give enough information for a meaningful recommendation, the category of each widgets is also needed. This data is located in the categories table, which includes the features of the category, including its parent category in the case of subcategories. Only the name (for displaying purposes) and the parent or children categories are taken from this table.

To relate the widgets with the categories, the table widgets is used, using the fields widget_id and subcategory_id to match widgets from click_accounts with the corresponding category.

Although the dataset contains much more information (including advertisers, ad campaigns, blog and forum posts, to mention a few. Only four concepts are needed: users, clicks, widgets and categories are enough for this analysis, although some of the unused data could be used to improve the results or to provide other kind of recommendations.

3.4. Algorithms: KMeans and XMeans

Once the main parameters of the clustering approach are established, it is necessary to choose the clustering algorithm. There are several options that can be used. However, given that we want to use existing implementations and, if possible, only use a concrete development API, we assume as a restriction the use of the algorithms that WEKA API implements by default.

The main algorithms that WEKA API allows to be used for clustering purposes are EM, KMeans and XMeans.

Given that there are still too many options, we decided to use only two options for implementation based on our own experience with clustering algorithms. The two options selected were KMeans and XMeans; the reason for this selection was based on the fact that KMeans algorithm is widely implemented and used and it has shown itself to be one of the best algorithms for clustering in most scenarios. On the other hand, XMeans is an improvement of KMeans algorithm where the algorithm has a better computational scalability, the number of clusters does not need to be supplied by the user and it offers a partial remedy for the problem of search in KMeans being prone to local minimal [29].

The configuration of the algorithms can be changed by the user of the system depending on the results that the algorithms are returning. KMeans algorithms need the user to provide the number of segments to generate. The actual criterion to determine this number is based on the distribution of the elements in the generated cluster for each segment. The algorithm used to obtain the “optimal” number of clusters to generate is the following:

1. For a concrete segment (which is chosen depending on the type of analysis to be done), KMeans algorithms are executed giving as a parameter the number of clusters to generate that will iterate between 2 (minimal) and a maximum that is established by the expert (normally, the number maximum of clusters is calculated taking into account the number of the attributes of the segment following the formula **1**. This heuristic formula tries to increase the number of clusters as the dataset grows more complex. The heuristic tries to provide at least one cluster per attribute given that we found out that many elements of the datasets could be classified by extreme values in a single attributes, suggesting to use at least a cluster per attribute to be able to classify these. We have used this calculation instead of; for example, add one to the number of attributes to leave a threshold.
2. Once the algorithm has been executed for all the possible values of the segment, the results of each execution are analyzed and a check on how the elements in the clusters are distributed is made, trying to find the distribution that has the most equally distributed elements. For example, if we have these two results: the first result has 4 clusters, and 10,000 elements with the following distribution: c1 = 100, c2 = 3.000, c3 = 6.000, c4 = 900. The second result has

5 clusters, also 10,000 elements, and the following distribution: c1 =2,000, c2 = 2,000, c3 = 2.500, c4 = 2,000, c5 = 1500. As can be seen, the second result has the most equally distributed elements and therefore, the number of clusters = 5 is a better candidate for the analyzed segment. This is shown in Code Listing 1. There are proven techniques for cluster evaluation such as those in [30,31], however we have selected to use this method for its simplicity, as perfect cluster selection is not strictly necessary.

```
foreach num_clusters in [2, num_attr * 3 / 2] {
        mean_instances = instances.size() / num_clusters
        clusters[num_clusters] = kmeans(instances, num_clusters)
        foreach cluster in clusters[num_clusters] {
                diffs[num_clusters] += (cluster.size() - mean_instances)^2
        }
}
return clusters[x] where diffs[x] = min(diffs)
```

Code Listing 1. Cluster selection

Maximum number of clusters formula:

$$Clusters = Segment_{Attributes} + \frac{Segment_{Attributes}}{2} \quad \textbf{(1)}$$

It is necessary to clarify that this criterion can be changed in the future and some other criterions can be chosen.

In the case of XMeans, this algorithm can calculate the number of clusters to generate itself, but it needs to indicate the minimum and maximum thresholds to obtain this number. In this case, the algorithm described before can also be applied to get a concrete and optimal number of clusters. However, we decided to let the algorithm choose the best configuration for each segment in most of the cases. There are some cases that we see (after observing the results that the algorithm returns) that a minimum of clusters needs to be generated in some segments. For this reason, we have the possibility of indicating the system that in a certain segment the thresholds should change to the ones provided instead of using the general configuration of XMeans algorithm.

4. Analysis of the results provided by clustering algorithms

The analysis performed on the results returned by the clustering algorithm was carried out bearing in mind the aim of AKNOBAS to provide recommendations to the users. With this objective, BeRuby experts analyzed the results returned by the clustering algorithm after applying it to the diverse segments that were generated and noted in previous sections. This analysis should be different in each segment because the information stored on each one of them is different and normally this information has no relation with other data stored in the rest of the segments. For this reason, an exhaustive and different analysis of each segment of the system should be applied. The analysis is generally based on some simple statistical factors because the information contained in the segments can be represented as percentages. For example, given a user, we can obtain, from all the possible categories of the system which one is the preferred by this user in percentage form (analyzing the total of categories clicked by the users and obtaining the most popular). If, for example, the category found is "sports", we can assume that the user is interested in sports, and we can offer him, for example, tickets or discounts to sports events. However, this is only one example of the possible analysis that can be done in the different segments.

In the following lines we explain, as a use case, the analysis of one of the most important segments in the BeRuby database.

The segment click_accounts, which is represented by the table "click_accounts" and the column of the table "widget_id" contains very valuable information about the interaction of BeRuby users against the system. In this table all the clicks that the user made on all the widgets of the system are stored. Some analyses made by BeRuby experts before the development of this system detect some "patterns" in the content of the database. For example, BeRuby experts detect a user profile whose behavior consists of clicking on all the available widgets one time, in order to get money for clicking, without exceeding the click limit of the widgets.

The following analysis that is shown in the paper is not unique. Other approaches can be used to analyze the data in order to get other types of information or relations. The procedure designed and implemented for the analysis of this segment is the following:

First of all, we get the output of the clustering algorithm (represented in table 1) and this output is stored in a data structure that, as a result, it generates an easier way to manage the data in the analysis process of the segment. Given that now we are interested in the number of appearances of each user and each widget in all of the generated clusters, we need to create the following elements:

- For each cluster a map is generated which relates the *widget_id* to the number of appearances of the aforementioned widget in the cluster. This structure is repeated for each cluster that has been generated by the clustering algorithm and it is stored on a list (*widgets_clusters*).
- A list with the number of widgets that each cluster contains: *widgets_count*
- For each user, a vector with the number of appearances in each cluster is generated. To facilitate the access through the user ID, it is stored on a map (*users_clusters*) using *user_id* attribute as key.

Figure 2 shows the data structures used in the analysis process.

Table 1. Output of the clustering algorithm

user_id	widget_id	Cluster
1	1	1
1	2	2
1	3	2
2	3	1
2	4	1
3	2	1

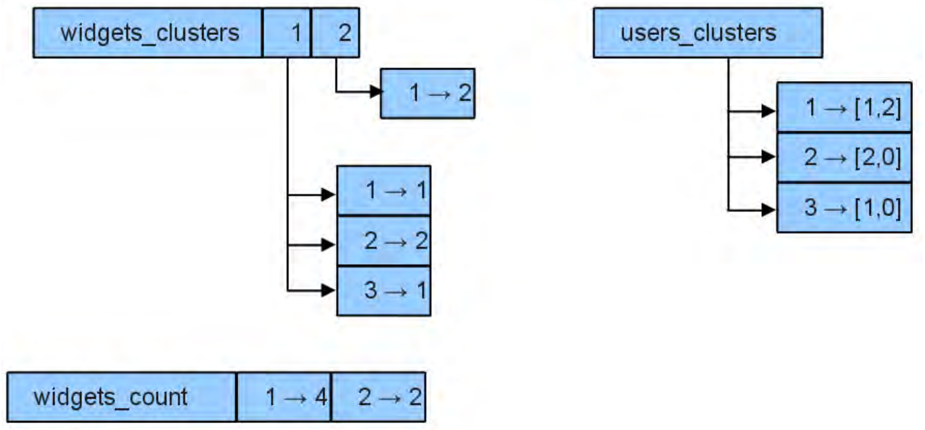


Fig. 2. Data structures used in the analysis process (I)

Hereafter, we proceed to the loading of the categories of the database. Table 2 shows the association between widgets and categories.

Table 2. Association between widgets and categories

widget_id	category_id
1	1
2	2
3	1
4	2
5	1

The categories are stored in two maps with the following associations:

- widget_id → category_id (widget_category)
- category_id → category_name (category_name)

With this data, we have calculated the distribution by categories of each cluster. For doing this, we have inspected all the elements of the list of widgets of each cluster, calculating the percentage that each widget represents of the total (dividing the value of widgets_clusters between the values of the corresponding widgets_count). Once done this, we have added this percentage to the corresponding category-cluster set, which is stored in a map list similar to widgets_clusters: categories_clusters.

The result in our example is shown in Figure 3.

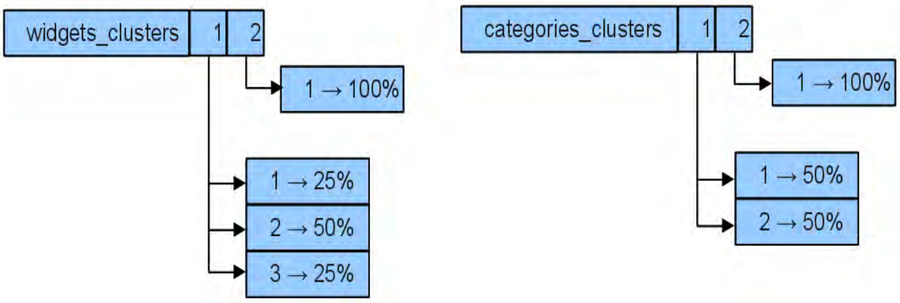


Fig. 3. New data structures generated after applying analysis

Finally, we have calculated the percentage of appearances of a user on a cluster, in other words, the total number of times that an instance with a certain *user_id* has been classified in each cluster, divided by the total of instances in which this user appears. The result of multiplying this value by the percentage of each category in a cluster is a representation of the preference of the user to subjects related to the mentioned category, therefore, the value is high, and it is showed. In our example Table 3 shows the results.

Table 3. Results after applying the analysis

User	Found	Cluster 1		Cluster 2
		Cat 1	Cat 2	Cat 1
1	[1, 2] Total: 3	1/3 * 50% = 16%	1/3 * 50% = 16%	2/3 * 100% = 66%
2	[2, 0] Total: 2	2/2 * 50% = 50%	2/2 * 50% = 50%	0/2 * 100% = 0%
3	[1, 0] Total: 1	1/1 * 50% = 50%	1/1 * 50% = 50%	0/1 * 100% = 0%

In the real case there should be a number of clicks that each user should carry out before considering it for this analysis, because one user with only one click (100% in the same cluster) could be possibly classified as fanatic of a category.

The whole process is summarized in Code Listing 2.

```
foreach inst in clustered_instances {
        widget_clusters[inst.cluster][inst.widget]++
        widgets_count[inst.cluster]++
        users_clusters[inst.user][inst.cluster]++
}
widget_category = sql(“SELECT widget, category FROM categories”)
foreach cluster in clusters {
        foreach (widget, count) in widgets_clusters[cluster] {
                widget_ratio = count / widgets_count[cluster]
                widget_clusters[cluster][widget] = widget_ratio
                cat = widget_category[widget]
                categories_clusters[cluster][cat] += widget_ratio
        }
}
foreach user in users_clusters.keys {
        user_categories = {}
        foreach (cluster, count) in users_clusters[user]  {
                user_ratio = count / sum(users_clusters[user])
                foreach (cat, catperc) in categories_clusters[cluster] {
                        user_categories[cat] += catperc * user_ratio
                }
        }
        foreach (cat, ratio) in user_categories {
                if ratio > threshold {
                        user_favorites[user] = (cat, ratio)
                }
        }
}
```

Code Listing 2. Analysis algorithm

5. Recommender System

The aim of the recommender system is to provide BeRuby administrators with a tool that allows them to do the task of sending personalized advertisements or recommendations to the users that are using the BeRuby platform. It is well known that a user who receives advertisements that are not of interest ignore them in most cases. The explanation of this common behavior is based on the fact that people have a concrete set of hobbies or interests and for this reason all the advertisements related to these interests attract their attention. These interests play an essential role in the AKNOBAS system. As mentioned throughout the paper, the aim of the system is the creation of groups of users (users profiles) which share some kind of features to be able to, for example, recommend to one user some of the preferences of another user that are in the same profile. However, this is only one of the possible points of view that can be used in the design and development of this system. Another kind of recommendation can be made based on the preferences or interests of a certain user, in order to recommend items to this user. For example, if we know that a user is interested in sports, we can recommend sports products to that user.

The exact recommendation that is made to each user depends on the segment analyzed. If, for example, we analyze the *click_accounts* segment, we can make recommendations to each user based on their interests, but we also can make recommendations to users based on the interests of other users (not his own). For this reason the analysis process plays a very important role in this system. It is important to carry out an exhaustive analysis of the data returned by the clustering algorithms to be able to, after analyzing the output, make recommendations.

In the previous sections the use case of the *click_accounts* segment has been studied. In the actual section, a recommendation case based on this segment is presented. However, it is important to note that the following case is only one of the several recommendation cases that could be made with this segment.

The following lines show one of the recommendations that can be made through the clustering and analysis of the *click_accounts* segment. Note that the following example is a real example that comes from the United Kingdom database of the BeRuby system.

Table 4. Results of analysis process

User (ID)	Category	Visited Widgets	Preference
50	Search	38	30.39%
50	e-mail	38	36.9%
79	e-mail	67	34.15%
8705	Blogs	18	41.85%
97	e-mail	14	34.26%
8717	Blogs	64	40.54%
8725	communities	11	30.43%
8725	Blogs	11	30.43%
140	e-mail	65	32.93%
8953	Blogs	52	41.85%
197	e-mail	8733	33.63%
8893	communities	14	31.88%
236	e-mail	28	35.58%
8841	Blogs	12	41.85%
8861	Blogs	88	31.38%
8849	Blogs	14	41.85%
9081	communities	41	35.34%

After executing the clustering and analysis process, we obtain the categories recommended for each user as is shown in Table 4.

Making use of all the knowledge collected, which is:

- Grouping of widgets and users (clustering)
- Each user's preferred categories (analysis)
- Widgets visited by each user (database)
- Popularity (number of clicks) of the widgets (database)

We can provide a recommendation of their possible widgets that can interest the user. For doing this, there are multiple ways of combining the data.

One of the easiest ways is to select the most popular widget of the category that the user prefers but which the user has never visited, as shown in Code Listing 3. So, we can obtain the recommendations as shown in Table 5.

Table 5. Results of recommendation system

User (ID)	Widget Recommended (ID and Name)	Widget Category	Widget Clicks (Popularity)
50	yahoo (3)	search	2547
50	Skype (122)	e-mail	498
79	gmail (8)	e-mail	5589
8705	Livejournal (337)	blogs	130
97	Skype (122)	e-mail	498
8717	Livejournal (337)	blogs	130
8725	Facebook (327)	communities	7872
8725	Livejournal (337)	blogs	130
140	hotmail (10)	e-mail	7210
8953	Livejournal (337)	blogs	130
197	gmail (8)	e-mail	5589
8893	Facebook (327)	communities	7872
236	hotmail (10)	e-mail	7210
8841	Livejournal (337)	blogs	130
8861	Xclusivo (898)	blogs	127
8849	Livejournal (337)	blogs	130
9081	Ideas4All (917)	communities	18890

However, in Table 5 we can notice that the results are very similar for the categories, and that there is a tendency to recommend the same widget to all the users that have the same favorite categories.

To solve this, we can combine these data with the output of the clustering algorithms, in such a way as to provide the most popular widget of the cluster and the category preferred by the user (who has not visited it yet):

For each user, we find the cluster in which most of their instances were classified. In this cluster, we select all the widgets that have a representative presence (in this example, the widget should represent at least 2% of the clicks on the cluster). Finally, the widgets that do not pertain to the user's majority category or that he has already visited is filtered and the most popular of each category is presented. This algorithm is shown in Code Listing 4.

In our example we obtain the results provided in Table 6.

Table 6. Result of recommendation system after cross results with the clusters

User (ID)	Widget Recommended (ID and Name)	Widget Category	Widget Clicks (Popularity)
50	yahoo (3)	search	2547
79	gmail (8)	e-mail	5589
8725	Bebo (969)	communities	3054
140	hotmail (10)	e-mail	7210
197	gmail (8)	e-mail	5589
8893	Twitter (968)	communities	2691
236	hotmail (10)	e-mail	7210

We can appreciate a slightly lower, but also more varied, number of recommendations for the same set of users which can be interpreted as a sign that they are more pertinent than the previous system.

```
foreach (user, (cat, ratio)) in user_favorites  {
        popular = sql(“SELECT * FROM widgets WHERE category = <cat>
                ORDER BY clicks DESC”)
        foreach widget in popular {
                if widget not in user_widgets[user] {
                        recommend(widget, user) and break
                }
        }
}
```

Code Listing 3. Naive recommendation algorithm

```
user_cluster = max(user_clusters[user])
foreach (widget, widget_ratio) in widget_clusters[user_cluster] {
        if widget_categories[widget] = user_favorites[user].cat {
                possible_widgets.add(widget)
        }
}
popular = sql(“SELECT * FROM widgets WHERE id IN <possible_widgets>
                ORDER BY clicks DESC LIMIT 1”)
recommend(popular, user)
```

Code Listing 4. Enhanced recommendation algorithm

Both systems modify their recommendations, given that when a user clicks on one of the recommended widgets, this is deleted from the recommender's list of options, providing more variety.

6. Evaluation

The evaluation of a recommender system can be divided in several types [32], emphasizing concretely two types of evaluations: 1) measuring the accuracy of the recommender system (how good is recommending) and 2) measuring the efficiency (is the system capable of making recommendations in an acceptable time? Is the system using too many resources?). Nowadays, the measurement of the time and resources only need to be done in some cases like when the amount of data used is too vast, or where the architecture doesn't have enough scalability.

In this paper, given that we have presented a use-case about application of data mining techniques to make recommendations, the evaluation was focused in know how good the system is performing these recommendations.

In order to compare the results of the accuracy of the recommender system it is necessary to have something to compare with. Nowadays, these kind of comparison are really hard to perform because it is necessary that both recommender systems are applied in the same domain, and using the same data.

Evaluation process:

The evaluation of AKNOBAS has been done using students from Computer Science studies of two different universities (Universidad Carlos III de Madrid, Spain and Instituto Tecnológico de Orizaba, Mexico). Given that the evaluated system have been developed to be used on real company, it is difficult to have access to the real users of the system. For this reason, we have selected twenty-five students from each university, which will represent the 25 real users of BeRuby platform.

The idea consists in associating a real user to one student of each university. This has been represented in figure 4. With this schema, we have two sets of potential "users", with twenty-five users each set.

Each student received a document with the data associated to the user which represents. This document contains anonymous information (to preserve confidentiality of data) about the clicks that user has done on the BeRuby platform, in order to allow the student to know the main pages in which the user is interested. The idea of this process is to allow the user to "represent" the user, to know how the user works with the system. Note: To simplify, students only have received information about the clicks that the user have made on a certain category (concretely, communities).

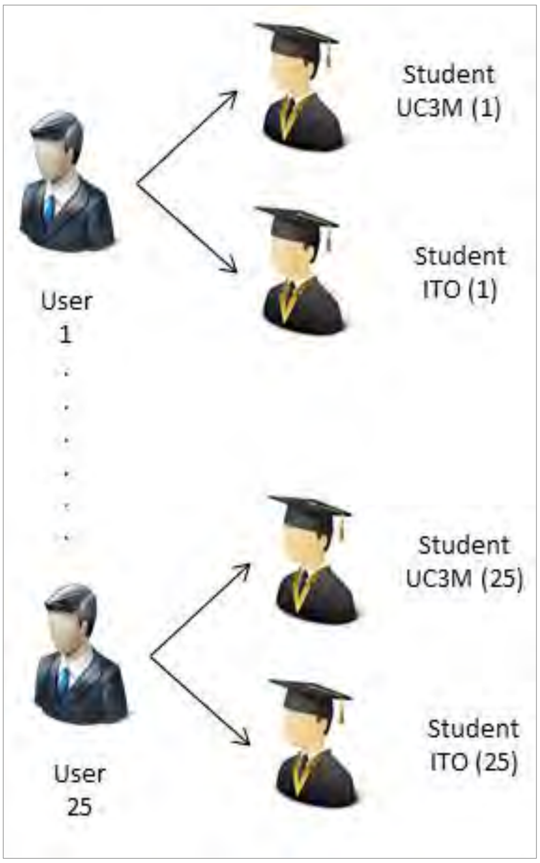


Fig. 4. Association between BeRuby users and students

Once the student has received this information, each student receives a personalized form which contains information about two kinds of recommendations:

- AKNOBAS Recommendation: The student receives three personalized recommendations (related with the category communities) that AKNOBAS system has generated for each user.
- Random Recommendation: We have developed an algorithm which chooses three random elements chosen from the most popular widgets in the “communities” category (excluding the elements which we know that have already been visited).

Then, the student should vote the three recommendations of AKNOBAS and the three recommendations of the Random process, which are presented in random order and without identifying the source of each one, with the following values:

- **Degree**: The user should value how good was the recommendation with a value between one and five: 1 – very bad, 2 – bad, 3 – neutral, 4 – good, 5 – very good.
- **Correct**: The student should indicate if consider that each recommendation of the three proposed, are correct (value 1), or not (value 0) given the knowledge that he have about the user preferences (clicks). If the user (student) considers that the degree of the recommendation is three or lower, the user should mark the recommendation as incorrect (0).

With this information, our goal is to determine how good was the system (in a percentage), and how good the users (students) consider the recommendations done by the system.

Note: The results of the evaluation are publicly available in [33]. However, the questioners send to the students and their concrete responses as well as the data obtained from BeRuby database cannot be published due to the agreement between the BeRuby company and the authors.

Evaluation results:

From the evaluation performed, we obtained two sets of results (one for each institution where the evaluation was done). First, we show and analyze the results of each institution to conclude analyzing the results of both. The file which contains the results of this evaluation could be accessed in [33]. The results provided in this document are focused in the final results, not including information about, for example, measuring each recommendation (R1 to R3).

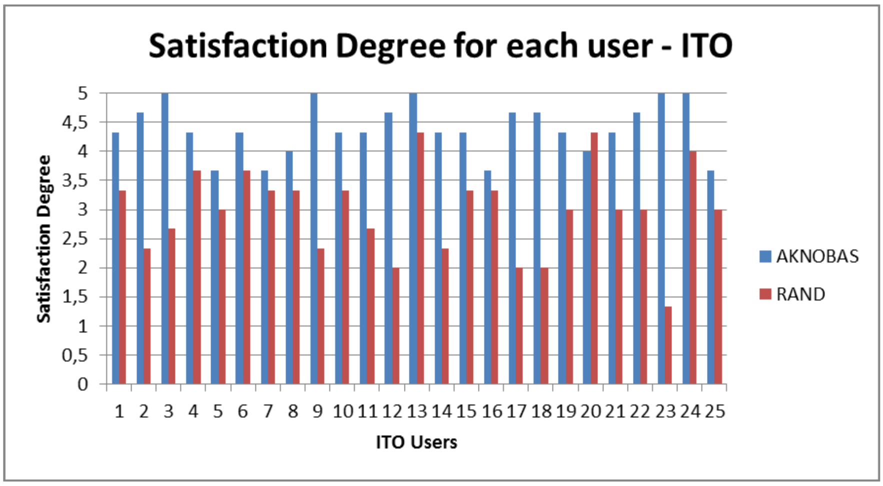


Fig. 5. Satisfaction Degree for each user (ITO)

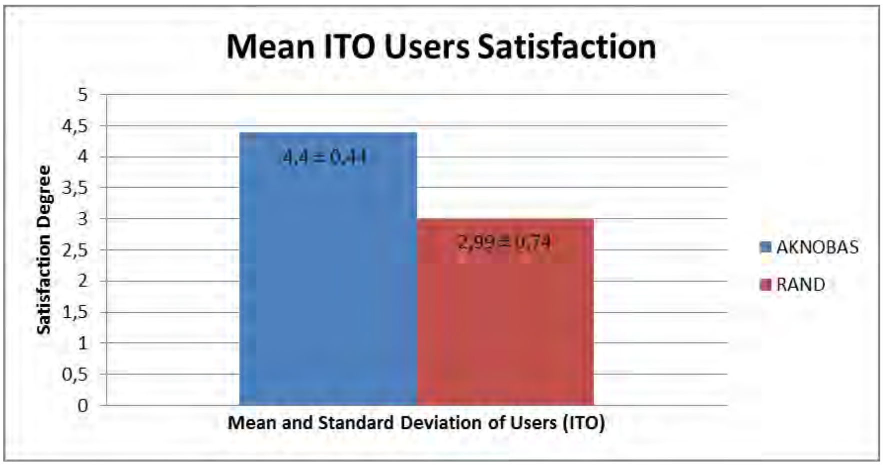


Fig. 6. Mean Satisfaction between ITO users

ITO Results:

As can be seen in file [33], in the section where the results of ITO are shown, we can notice that the satisfaction degree of the results obtained from AKNOBAS is higher than the once obtained with the random recommendation. Figure 5 shows this result. Also, in figure 6 we can notice the mean satisfaction, and the standard deviation referred to the satisfaction of the ITO students.

The accuracy of recommendations made by AKNOBAS or the rand algorithm also was measured. This measurement has been done using basic metrics, calculating the percentage of recommendations which were considered correct with the formula 2:

$$Accuracy = \frac{Correct\ Recommendations}{Total\ Recommendations} \quad \textbf{(2)}$$

Finally, to know if there are significant differences between the results provided by AKNOBAS and the rand algorithm, a T-Student has been applied to the data. T-Student only has been applied to the total data (join of the three available recommendations). The reason to do that, is that we want to know if there are significant differences in the whole process (we have 25 (users) * 3 recommendations = 75 data samples for AKNOBAS and the same for the rand algorithm), not in the individual recommendations (R1 to R3). The appliance of T-Student with a significance of 0.05 (5%) returns (t(148) = -4,935; p<0.05), which means that there are significant differences between AKNOBAS and the rand recommendation.

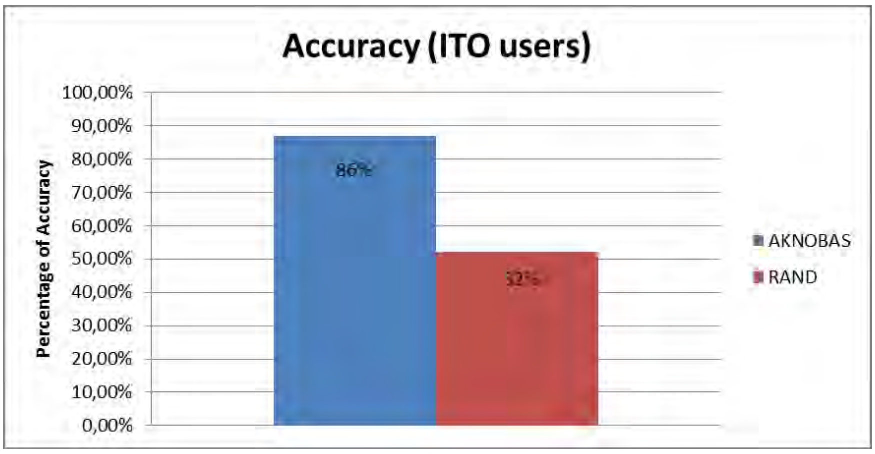


Fig. 7. Mean accuracy (ITO users)

Given that the accuracy of AKNOBAS (with ITO users) is quite higher (34% higher), and from a statistic point of view we can ensure that there are significant differences between both systems, we can demonstrate the utility of AKNOBAS platform in the recommendation process, with the data obtained from students from ITO.

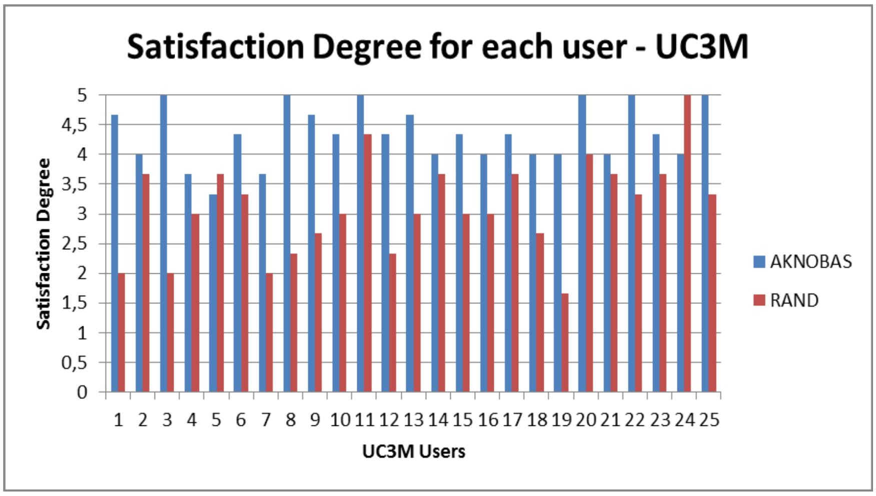


Fig. 8. Satisfaction Degree for each user (UC3M)

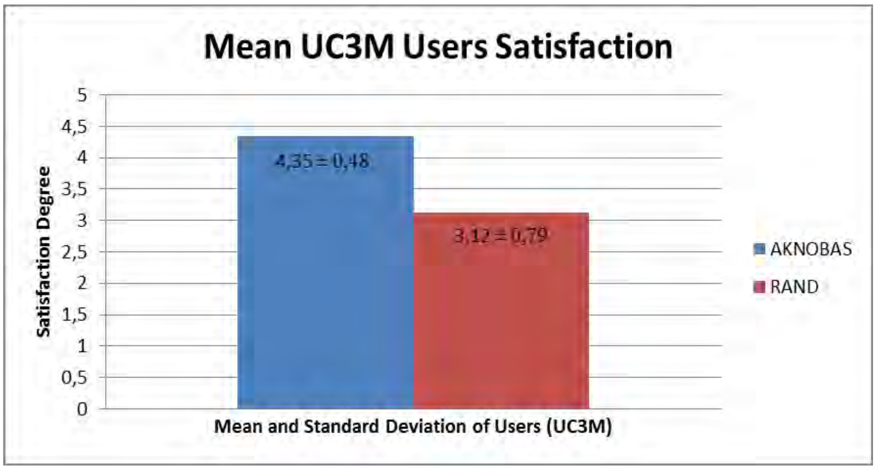


Fig. 9. Mean Satisfaction between UC3M users

UC3M Results:

Again, as can be seen in file [33], in the section where the results of UC3M are shown, we can see that again the satisfaction degree of the results obtained from AKNOBAS is higher than the once obtained with the random recommendation. Figure 8 shows this result for UC3M students. Also, in figure 9 we can notice the mean satisfaction, and the standard deviation referred to the satisfaction of the students from UC3M.

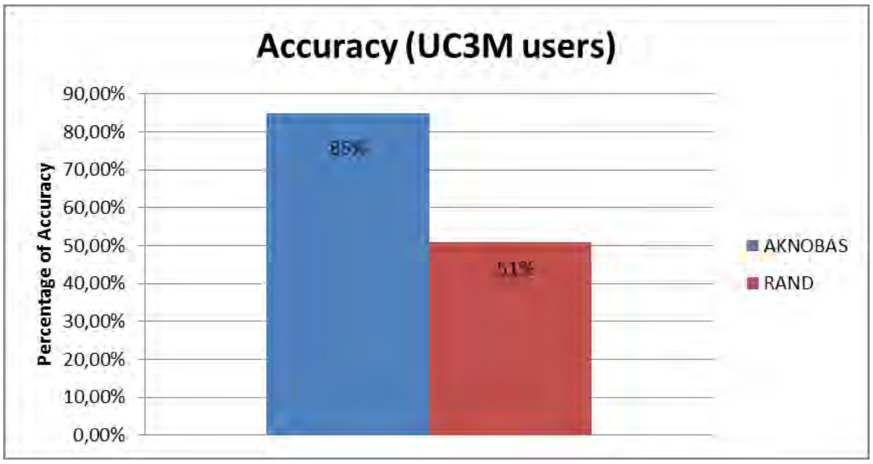


Fig. 10. Mean accuracy (UC3M users)

In Figure 10, the accuracy of AKNOBAS system and the rand algorithm for UC3M students are presented.

Finally, again, we applied T-Student to know if there are significant differences between AKNOBAS and the rand algorithm, The appliance of T-Student with a significance of 0.05 (5%) returns (t(148) = -4,869; p<0.05), which means that there are significant differences between AKNOBAS and the rand recommendation.

Given that the accuracy of AKNOBAS (with UC3M) is, again, quite higher (34% higher), and from a statistic point of view we can ensure that there are significant differences between both systems, we can demonstrate the utility of AKNOBAS platform in the recommendation process, with the data obtained from UC3M students.

UC3M and ITO:

To summarize the results provided by the independent analysis of the two groups of students, we show some figures which represent the final results from the analysis performed.

In this case, we applied again an analysis of the data using a T-Student. We try to know if there are (or not) significant differences between UC3M and ITO users using AKNOBAS, and random algorithm.

In the first one, we analyzed the data to know if there are significant differences between UC3M users and ITO users using AKNOBAS. Results return (t(148) =,234 ; p<0.05), which means that there are not significant differences between UC3M and ITO opinions about the recommendations of AKNOBAS.

In the case of random algorithm, we obtained (t(148) =,162 ; p<0.05), which, again, means that there are not significant differences between UC3M and ITO opinions about the recommendations of random algorithm.

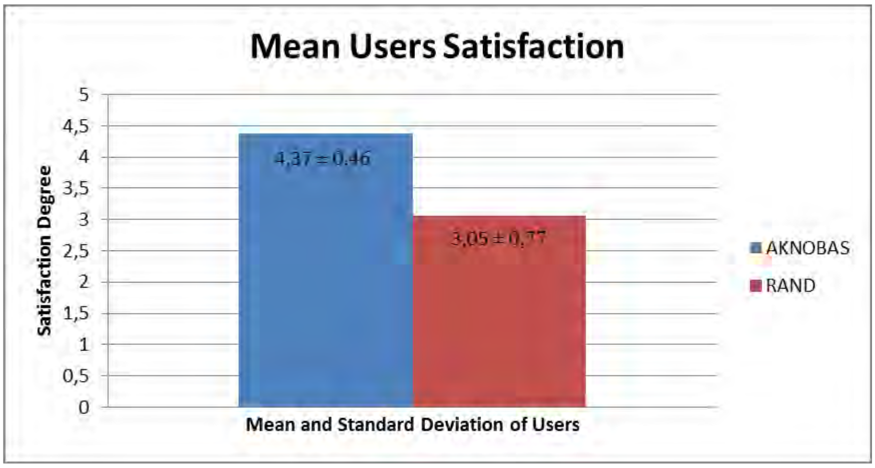


Fig. 11. Mean user satisfaction

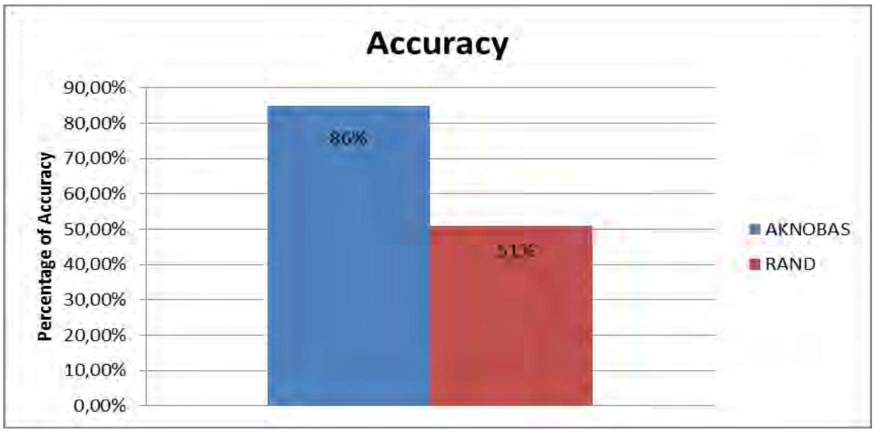


Fig. 12. Mean accuracy

In this case, it is not necessary to perform the T-Student analysis between the join of UC3M and ITO data to know if there are significant differences between AKNOBAS and random algorithm because as we observed in the previous study, in both groups of users we can notice significant differences. For this reason, in the summary of both groups, we found again these significant differences.

Figure 11 shows the mean accuracy of the system, using as input data the results obtained previously with ITO and UC3M users. Figure 12 shows the mean satisfaction degree of the users, using again the results previously obtained with ITO and UC3M users.

Evaluation conclusions:

As we can see in the evaluation conducted, the AKNOBAS system provides quite good results in the recommendation process. The random algorithm, used to allows us to have results to compare AKNOBAS with, has been designed, as was mentioned before, to obtain recommendations from a concrete set of possibilities (the random algorithm only could get recommendation items from a certain list which their elements pertains to the category used in the evaluation: communities). Given this feature of the random algorithm it is possible to notice that this algorithm have a high probability to return correct results given that the recommendation items are obtained from the same category. Even with that, AKNOBAS is more accurate. The reason is because as was explained in the previous sections, the paper tries to obtain these hide relations which in fact exist between different users, and with these relations, it is capable to recommend new items which are preferred by the users.

An important part of this evaluation is the analysis from the statistical point (T-Student) to see that: 1) there are not differences between UC3M users and ITO users (in AKNOBAS and random algorithm results), 2) check that the results provided by AKNOBAS and random algorithm have significant differences.

Also, it is important the measurement about the degree of satisfaction which was done. This degree also allows us to know how good the user considers the recommendation. This is a quite important thing, because this knowledge about "how good" a user finds a certain recommendation, can be reused by the AKNOBAS system in order to perform better recommendations in the future.

7. Conclusions and Future Work

Using data mining algorithms to provide added value to current information systems has been depicted as a potential next generation for advanced information processing based on AI techniques. However, meeting real-world business scenario requirements has been lately leveraged by a new generation of web-based user information, mostly related with user preferences, profiles and, eventually, behavioral analysis. Fundamentally, most Customer Relationship Management (CRM) systems, a particular type of information systems, have already leveraged a number of operational and functional features of most business information systems which can provide an analytical framework to respond and cope with some customer demands or expectations.

However, these systems have not met the challenge in understanding and partially anticipating the users' behavioral patterns. In this paper, we have presented an initiative in terms of using intelligent data mining techniques to a real scenario based on well-known AI algorithms which can provide not only an analytical framework but also a powerful software platform for managing critical mass user demands and desires and turning them into corporate knowledge and substantial desires, changes and expectations awareness.

Our future work in AKNOBAS involves a threefold approach. On the one hand, we will continue evaluating performance thanks to the real-world business oriented requirements from BeRuby, as a leading Web-based user-driven business application and model. On the other hand, we will survey a set of new generation algorithms based on behavioral mechanisms, themselves based on meta-heuristics. Finally, we intend to extend our approach to other different business requirements which could require a higher level of real time reaction, such as those including Web 2.0, Social Networks or Real-Time Web applications (such as Twitter) where the main tenets of our approach could be harnessed.

Acknowledgements. This work was supported by the Spanish Ministry of Industry, Tourism, and Commerce under the projects TRAZAMED (IPT-090000-2010-007), ENERFICIENCY (TSI-020400-2011-56), GECALLIA (TSI-020100-2011-244) and (SEMOSA TSI-020400-2011-51). It was also supported by the Spanish Ministry of Science and Innovation under the project FLORA (TIN2011-27405). Additionally, this work was supported by the General Council of Superior Technological Education of Mexico (DGEST). Additionally, this work was sponsored by the National Council of Science and Technology (CONACYT) and the Public Education Secretary (SEP) through PROMEP.

References

1. Hartigan, J.A.: Clustering Algorithms. John Wiley & Sons, Inc., 1975
2. Gan, G., Ma, C., Wu, J.: Data Clustering: Theory, Algorithms, and Applications. ASASIAM Series on Statistics and Applied Probability. Vol. 20. (2007)
3. Yuan, S., Cheng, C.: Ontology-based personalized couple clustering for heterogeneous product recommendation in mobile marketing. Expert systems with applications. Vol. 26, No. 4, 461–476. (2004)
4. Han, Y., Shi, P.: An improved ant colony algorithm for fuzzy clustering in image segmentation. Neurocomputing. Vol. 70 No.4-6, 665-671. (2007)
5. Yeo, A.C., Smith, K.A., Willis, R.J., Brooks, M.: Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry. Intelligent Systems in Accounting, Finance & Management. Vol. 10, No. 1, 39–50. (2001)
6. Pronk, N.P., Anderson, L.H., Crain, A.L., Martinson, B.C., O'Connor, P.J., Sherwood, N.E., Whitebird, R.R.: Meeting recommendations for multiple healthy lifestyle factors: Prevalence, clustering, and predictors among adolescent, adult, and senior health plan members. American Journal of Preventive Medicine. Vol. 27, No. 2, 25-33. (2004)
7. Kedes, L.H., Birnstiel, M.L.: Reiteration and Clustering of DNA Sequences Complementary to Histone Messenger RNA. Nature new biology. Vol. 230, 165-169. (1971)
8. Xing, H.J., Hu, B.G.: An adaptive fuzzy c-means clustering-based mixtures of experts model for unlabeled data classification. Neurocomputing. Vol. 71, No. 4-6, 1008-1021. (2008)
9. G. Huming, and L. Weili,: A Hotel Recommendation System Based on Collaborative Filtering and Rankboost Algorithm. In Proceedings of the Second International Conference on Multi Media and Information Technology, IEEE Computer Society, Washington, DC, USA, 317-320 (2010).
10. F. Juan, L. Wencan.: An Architecture for Collaborative Filtering in Grid Portal Recommendation System. In Proceedings of the Fourth International Conference on Semantics, Knowledge and Grid, IEEE Computer Society, Washington, DC, USA, 359-363. (2008).
11. H. Liang, Y. Xu, Y. Li, R. Nayak.: Collaborative Filtering Recommender Systems Using Tag Information. In Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society, Washington, DC, USA, 59-62. (2008).
12. S. Gong, H. Ye, H. Tan.: Combining Memory-Based and Model-Based Collaborative Filtering in Recommender System. In Proceedings of the Conference on Circuits, Communications and System, IEEE Computer Society, Washington, DC, USA, 690-693. (2009).
13. H. Tan, H. Ye.: A Collaborative Filtering Recommendation Algorithm Based On Item Classification. In Proceedings of the Conference on Circuits,

Communications and System, IEEE Computer Society, Washington, DC, USA, 694-697. (2009).

14. A. Turati, D. Cerizza, I. Celino, E. Della V.: Analyzing User Actions within a Web 2.0 Portal to Improve a Collaborative Filtering Recommendation System. In Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology – Workshops, IEEE Computer Society, Washington, DC, USA, 65-68. (2009).
15. Eom, K.B.: Fuzzy clustering approach in unsupervised sea-ice classification. Neurocomputing. Vol. 25, No. 1-3, 149-166. (1999)
16. Ambroise, C., Sèze, G., Badran, F., Thiria, S.: Hierarchical clustering of self-organizing maps for cloud classification. Neurocomputing. Vol. 30, No. 1-4, 47-52. (2000)
17. Lawrence, R.D., Almasi, G.S., Kotlyar, V., Viveros, M.S., Duri, S.S.: Personalization of Supermarket Product Recommendations. Data mining and knowledge discovery. Vol. 5, No. 1-2, 11-32. (2001)
18. Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. ACM Conference On Recommender Systems, ACM, New York, NY, USA, 259-266. (2008)
19. Sarwar, B.M., Karypis, G., Konstan, J., Riedl, J.: Recommender Systems for Large-scale E-Commerce: Scalable Neighborhood Formation Using Clustering. In Proceedings of the Fifth International Conference on Computer and Information Technology. ACM, New York, NY, USA, 19-24. (2002)
20. Wu, Y.H., Chen, Y.C., Chen, A.L.P.: Enabling personalized recommendation on the Web based on user interests and behaviors. Research Issues in Data Engineering, Proceedings of Eleventh International Workshop on Research Issues in Data Engineering. Document Management for Data Intensive Business and Scientific Applications. IEEE Computer Society, Washington, DC, USA, 17-24. (2001)
21. Kim, J.W., Lee, B.H., Shaw, M.J., Chang, H.L., Nelson, M.: Application of Decision-Tree Induction Techniques to Personalized Advertisements on Internet Storefronts. International Journal of Electronic Commerce, Vol. 5, No. 5, 45–62. (2001)
22. Corchado, J., Bajom, J. De Paz, Rodríguez, S.: An execution time neural-CBR guidance assistant. Neurocomputing. Vol. 72, No. 13-15, 2743- 2753. (2009)
23. Angles, P., Blattner, D.: System and method for delivering customized advertisements within interactive communication systems. United States patent application publication. (2006)
24. LeMole, S., Howard, S., Thomas, J., Stuntebeck, P.: Method and system for presenting customized advertising to a user on the World Wide Web. United States patent. (1999)
25. Resnick, P., Varian, H.R.: Recommender systems. Communications of the ACM. Vol. 40, No. 3, 56-58. (1997)
26. BeRuby. http://www.beruby.com. Last accessed: July, 21. 2011.
27. Han, J., Kamber, M.: Data Mining. Concept and Techniques. Morgan Kaufman. 2nd Edition. (2006)
28. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Willey. 9th Edition. (2005)
29. Pelleg, D., Moore, A.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters. ICML '00 Proceedings of the Seventeenth International

Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 727-734. (2000)

30. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Cluster validity methods: part I. SIGMOD Rec. 31, 2, 40-45. (2002)
31. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity checking methods: part II. SIGMOD Rec. 31, 3, 19-27. (2002)
32. Stephan Weibelzahl. Framework for the evaluation of adaptive cbr systems. In Proceedings of the 9th German Workshop on Case-Based Reasoning, 254–263, Baden-Baden, Germany (2001).
33. Evaluation results. Available online at: http://nadir.uc3m.es/alejandro/pubs /beruby/ evresults.pdf

Alejandro Rodríguez González, PhD, is a Researcher and Teaching Assistant of Computer Science Department at University Carlos III of Madrid. His main interests include Artificial Intelligence, Medical Informatics, Semantic Web, and Medical Diagnosis Systems.

Enrique Jiménez Domingo is a Researcher, Teaching Assistant and Ph.D. candidate of the Computer Science Department at Universidad Carlos III de Madrid. His main research lines involve Cloud Computing, Artificial Intelligence and Semantic Web.

Javier Torres Niño is a researcher in the Computer Science Department at University Carlos III. His current interests are applied Artificial Intelligence and large scale distributed systems.

Juan Miguel Gómez-Berbís is an Associate Professor at the Computer Science Department of the Universidad Carlos III de Madrid. He holds a Ph.D. in Computer Science from the Digital Enterprise Research Institute (DERI) at the National University of Ireland, Galway.

Giner Alor-Hernandez, PhD, is a full-time researcher of the Division of Research and Postgraduate Studies of the Instituto Tecnologico de Orizaba. His research interests include Web services, e-commerce, Semantic Web, Web 2.0, service-oriented and event-driven architectures and enterprise application integration.

Received: July 22, 2011; Accepted: February 06, 2012.

DOI: 10.2298/CSIS110822011H

Scope of MPI/OpenMP/CUDA Parallelization of Harmonic Coupled Finite Strip Method Applied on Large Displacement Stability Analysis of Prismatic Shell Structures

Miroslav Hajduković[1], Dragan D. Milašinović[2], Miloš Nikolić[1], Predrag Rakić[1], Žarko Živanov[1], and Lazar Stričević[1]

[1] University of Novi Sad, Faculty of Technical Sciences,
Trg D. Obradovića 6, 21000 Novi Sad, Serbia
{hajduk, losmi83, pec, zzarko, lucky}@uns.ac.rs
[2] University of Novi Sad, Faculty of Civil Engineering,
24000 Subotica, Serbia
ddmil@gf.uns.ac.rs

Abstract. This paper presents scope of parallelization in large displacement stability analysis of orthotropic prismatic shells with simply supported boundary conditions along the diaphragm-supported edges. A semi-analytical harmonic coupled finite strip method (HCFSM) is used to solve the large deflection and the post-buckling problems. In the HCFSM formulation the coupling of all series terms highly increases computation time when a large number of series terms is used. Fortunately such computation time increase can be compensated by application of MPI, OpenMP and CUDA approaches to parallelization. This paper shows how these approaches can be applied to HCFSM formulation, and what results can be expected.

Keywords: Prismatic Shells, Stability analysis, HCFSM, MPI, OpenMP, CUDA.

1. Introduction

Typical flat plate structures under consideration here are simply supported by diaphragms and may have arbitrary longitudinal edge conditions. It may be subjected to a transverse loading between its ends (the large-deflection problem) or to an in-plane progressive destabilizing, buckling load (the post-buckling problem) or conceivably to both actions simultaneously. For these structures, the design process should lead to define the optimal morphology of the transversal cross-section, which means its geometry, size, shape and topology.

If a structure undergoes large deformation, the second order terms regarding the strains cannot be ignored. The strain-displacement relations, within the context of a Green-Lagrange strain tensor, represent a sum of linear and non-linear part. Analysis of plates in the post-buckling range is generally performed on the basis of von Karman equations or by employing an energy approach. Only approximate solutions can be obtained, taking into account the in-plane (membrane) and out-of-plane (bending) boundary conditions.

In this work we present the semi-analytical harmonic coupled finite strip method (HCFSM) for geometric non-linear analysis of thin plate structures under multiple loading conditions. This method takes into account the important influence of the interaction between the buckling modes. In contrast, the finite strip method (FSM), in its usual form, ignores this interaction and therefore cannot be used in analysis of these structures [1].

However, already high computational complexity of FSM becomes even higher in the HCFSM formulation that couples all series terms. The result is a very long calculation time of the existing sequential HCFSM software in cases of a large number of series terms [2]. In order to speed up the lengthy calculations various parallelization techniques can be used. Previous research [3] has considered FSM parallelization based on the “networked workstations” and to our knowledge this is the only publicly available result on this subject. Our research treats a new method (HCFSM) and different computer architectures. We show that not only the multicomputer architecture based on message passing scheme (MPI [4]) is suitable for HCFSM parallelization, but also the multiprocessor architecture based on shared memory (OpenMP [5] and CUDA [6]) offers a good potential for HCFSM parallelization [7, 8, 9]. The main intention of this paper is to present scope of HCFSM parallelization by separate application of each of three different available parallelization techniques.

In the second chapter we describe the HCFSM formulation. The third chapter contains a discussion of our hybrid approach to HCFSM parallelization based on application of MPI, OpenMP, and CUDA programming models. This approach results in a noticeable reduction of the execution time and represents a new contribution to the (HCFSM) parallelization. The last two chapters contain acknowledgements and conclusions.

2. Harmonic coupled finite strip method for the stability analysis of thin-walled structures

2.1. Introduction to HCFSM

The contemporary alternatives to perform the stability analysis of thin-walled members are generalized beam theory (GBT) [10], finite element method (FEM) and the FSM [11]. The FSM is derived from the FEM and represents specialization of the FEM to the analysis of engineering structures [2]. The only difference between FEM and FSM consist of the different discretization of the members, as seen in Fig. 1. The FEM uses a mesh that uses discretization of the member transversally and longitudinally, while the FSM needs only transversal discretization, using currently either harmonic or spline functions in the longitudinal direction of the member.

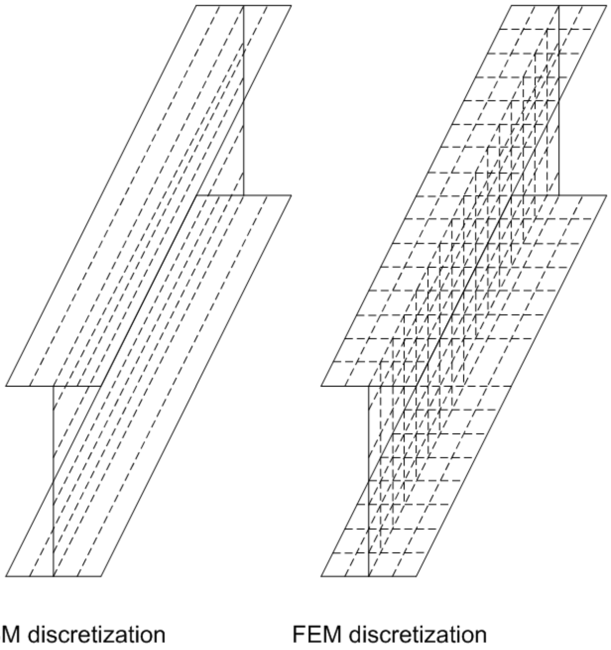


Fig. 1. FSM discretization versus FEM discretization

The FSM was originally developed by Cheung [1] and was widely used by other authors for understanding and predicting the behavior of cold-formed steel members and for bridge decks. There is a reliable harmonic FSM program – CUFSM – for the determination of the critical load parameter of thin-walled prismatic members under a general longitudinal normal stress loading at the extreme cross sections [12].

The well known uncoupled FSM formulation, represents a semi-analytical finite element process. As far as linear analysis is concerned, it takes advantage of the orthogonality properties of harmonic functions in the stiffness matrix formulation. However in the case of the geometric nonlinear analysis, and the geometric stiffness matrix calculation, the integral expressions contain the product of trigonometric functions with higher-order exponents, and here the orthogonality characteristics are no longer valid. All harmonics are coupled, and the stiffness matrix order and bandwidth are proportional to the number of harmonics used. This kind of FSM analysis is named the HCFSM [2, 13].

The nonlinear strain-displacement relations in the finite strip can be predicted by the combination of the plane elasticity and the Kirchhoff plate theory. Using this assumption in the Green-Lagrange strain tensor (1) for in-plane nonlinear strains gives Green-Lagrange HCFSM formulation. Also that, neglecting lower-order terms in a manner consistent with the usual von Karman assumptions gives HCFSM von Karman formulation.

$$\varepsilon_{ij} = 1/2\left(u_{i,j} + u_{j,i} + u_{k,i}u_{k,j}\right). \tag{1}$$

The essential feature of geometric nonlinearity is that equilibrium equations must be written with respect to the deformed geometry – which is not known in advance.

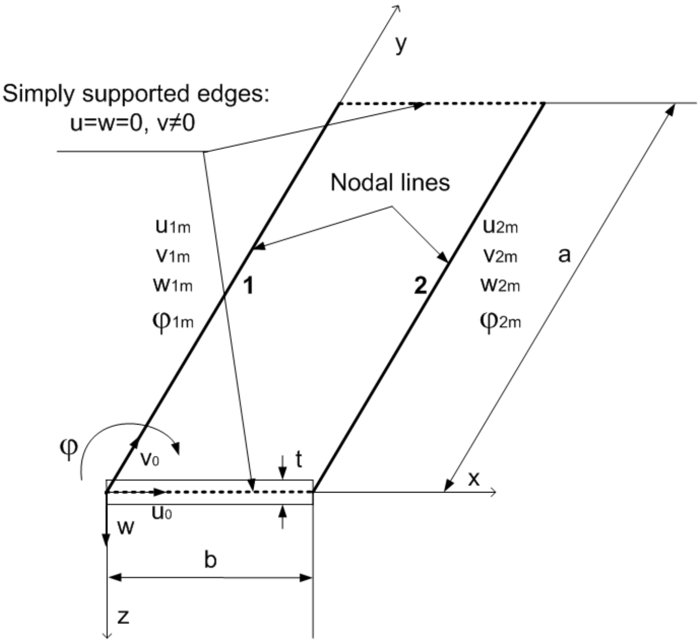


Fig. 2. Simply supported flat shell strip

In the FSM, which combines elements of the classical Ritz method and the FEM, the general form of the displacement function can be written as a product of polynomials and trigonometric functions:

$$f = \sum_{m=1}^{r} Y_m(y) \sum_{k=1}^{c} \mathbf{N}_k(x)\mathbf{q}_{km} \tag{2}$$

where $Y_m(y)$ are functions from the Ritz method and $\mathbf{N}_k(x)$ are interpolation functions from the FEM. We define the local Degrees Of Freedom (DOFs) as the displacements and rotation of a nodal line (DOFs=4), as shown in Fig. 2. The DOFs are also called generalized coordinates.

The following approximate functions are used for the simply supported flat shell strip with edges restrained against in-plane movement:

$$u_0 = \sum_{m=1}^{r} Y_{um}^{u} \mathbf{N}_u^u \mathbf{q}_{um}^u = \sum_{m=1}^{r} Y_{um}^{u} \begin{bmatrix} 1-\frac{x}{b} & \frac{x}{b} \end{bmatrix} \mathbf{q}_{um}^u , \tag{3}$$

$$v_0 = \sum_{m=1}^{r} \frac{a}{m\pi} Y_{um}^{v} \mathbf{N}_u^v \mathbf{q}_{um}^v = \sum_{m=1}^{r} \frac{a}{m\pi} Y_{um}^{v} \begin{bmatrix} 1-\frac{x}{b} & \frac{x}{b} \end{bmatrix} \mathbf{q}_{um}^v ,$$

$$w = \sum_{m=1}^{r} Y_{wm} \mathbf{N}_w \mathbf{q}_{wm} = \sum_{m=1}^{r} Y_{wm} \begin{bmatrix} N_1 & N_2 & N_3 & N_4 \end{bmatrix} \mathbf{q}_{wm} ,$$

$$N_1(x) = 1 - 3\left(\frac{x}{b}\right)^2 + 2\left(\frac{x}{b}\right)^3 , N_2(x) = b\left[\frac{x}{b} - 2\left(\frac{x}{b}\right)^2 + \left(\frac{x}{b}\right)^3\right],$$

$$N_3(x) = 3\left(\frac{x}{b}\right)^2 - 2\left(\frac{x}{b}\right)^3 , N_4(x) = b\left[-\left(\frac{x}{b}\right)^2 + \left(\frac{x}{b}\right)^3\right],$$

$$\mathbf{q}_{wm} = \begin{bmatrix} w_{1m} \\ \varphi_{1m} \\ w_{2m} \\ \varphi_{2m} \end{bmatrix}, \mathbf{q}_{um}^u = \begin{bmatrix} u_{1m} \\ u_{2m} \end{bmatrix}, \mathbf{q}_{um}^v = \begin{bmatrix} v_{1m} \\ v_{2m} \end{bmatrix},$$

$$Y_{um}^{u}(y) = \sin\left(m\pi \frac{y}{a}\right) = Y_{wm} ,$$

$$Y_{um}^{v} = Y_{u,ym}^{u} = \left(\frac{m\pi}{a}\right)\cos\left(\frac{m\pi y}{a}\right), m = 1,2... \; .$$

2.2. HCFSM formulation for stability analysis

Total potential energy

As a preliminary to tracing the equilibrium paths, it is necessary to determine the total potential energy of the structure as a function of the global DOFs. The steps in the computation are detailed discussed in [2].

The total potential energy of a strip is designated *Π* and is expressed with respect to the local DOFs by the HCFSM.

$$
\begin{aligned}
\mathit{\Pi} &= U + W = \left(U_m + U_b\right) + W = \left(1/2 \int_A \mathbf{q}_u^T \mathbf{B}_{u1}^T \mathbf{D}_{11} \mathbf{B}_{u1} \mathbf{q}_u dA + 1/2 \int_A \mathbf{q}_w^T \mathbf{B}_{w3}^T \mathbf{D}_{22} \mathbf{B}_{w3} \mathbf{q}_w \right. \\
&+ \left[\begin{array}{l} 1/8 \int_A \mathbf{q}_w^T \mathbf{B}_{w2}^T \mathbf{W}^T \mathbf{B}_{w1}^T \mathbf{D}_{11} \mathbf{B}_{w1} \mathbf{W} \mathbf{B}_{w2} \mathbf{q}_w dA + 1/4 \int_A \mathbf{q}_w^T \mathbf{B}_{w2}^T \mathbf{W}^T \mathbf{B}_{w1}^T \mathbf{D}_{11} \mathbf{B}_{u1} \mathbf{q}_u dA + \\ + 1/4 \int_A \mathbf{q}_u^T \mathbf{B}_{u1}^T \mathbf{D}_{11} \mathbf{B}_{w1} \mathbf{W} \mathbf{B}_{w2} \mathbf{q}_w dA \end{array} \right] + \\
&+ \left\{ \begin{array}{l} 1/8 \int_A \mathbf{q}_u^{uT} \mathbf{B}_{u2}^{uT} \mathrm{U}^T \mathbf{B}_{u1}^{uT} \mathbf{D}_{11} \mathbf{B}_{u1}^{u} \mathrm{U} \mathbf{B}_{u2}^{u} \mathbf{q}_u^u dA + 1/4 \int_A \mathbf{q}_u^{uT} \mathbf{B}_{u4}^{uT} \mathbf{D}_{11} \mathbf{B}_{u1}^{u} \mathrm{U} \mathbf{B}_{u2}^{u} \mathbf{q}_u^u dA + \\ + 1/4 \int_A \mathbf{q}_u^{uT} \mathbf{B}_{u2}^{uT} \mathrm{U}^T \mathbf{B}_{u1}^{uT} \mathbf{D}_{11} \mathbf{B}_{u4}^{u} \mathbf{q}_u^u dA + \\ + 1/4 \int_A \mathbf{q}_u^{vT} \mathbf{B}_{u5}^{vT} \mathbf{D}_{11} \mathbf{B}_{u1}^{u} \mathrm{U} \mathbf{B}_{u2}^{u} \mathbf{q}_u^u dA + 1/4 \int_A \mathbf{q}_u^{uT} \mathbf{B}_{u2}^{uT} \mathrm{U}^T \mathbf{B}_{u1}^{uT} \mathbf{D}_{11} \mathbf{B}_{u5}^{v} \mathbf{q}_u^v dA + \\ + 1/8 \int_A \mathbf{q}_w^T \mathbf{B}_{w2}^T \mathbf{W}^T \mathbf{B}_{w1}^T \mathbf{D}_{11} \mathbf{B}_{u1}^{u} \mathrm{U} \mathbf{B}_{u2}^{u} \mathbf{q}_u^u dA + 1/8 \int_A \mathbf{q}_u^{uT} \mathbf{B}_{u2}^{uT} \mathrm{U}^T \mathbf{B}_{u1}^{uT} \mathbf{D}_{11} \mathbf{B}_{w1} \mathbf{W} \mathbf{B}_{w2} \mathbf{q}_w \end{array} \right. \\
&- \int_A \mathbf{q}^T \mathbf{A}^T \mathbf{p} dA
\end{aligned} \tag{4}
$$

The multiplication results of the membrane and bending actions in the first bracket of Eq. (4) are uniquely defined and uncoupled, whilst those in second [von Karman assumptions] and third bracket {Green-Lagrange approach} are functions of the displacements *u₀*, *v₀* and *w*. Consequently, the membrane and bending actions are coupled in many ways.

The conventional and the geometric stiffness matrices are, respectively:

$$
\begin{aligned}
&(\hat{\mathbf{K}}_{uu} = \int_A \mathbf{B}_{u1}^T \mathbf{D}_{11} \mathbf{B}_{u1} dA ,\ \hat{\mathbf{K}}_{ww} = \int_A \mathbf{B}_{w3}^T \mathbf{D}_{22} \mathbf{B}_{w3} dA) \\
&[\tilde{\mathbf{K}}_{ww} = \int_A \mathbf{B}_{w2}^T \mathbf{W}^T \mathbf{G}_1 \mathbf{W} \mathbf{B}_{w2} dA ,\ \tilde{\mathbf{K}}_{wu} = \int_A \mathbf{B}_{w2}^T \mathbf{W}^T \mathbf{G}_2 dA ,\ \tilde{\mathbf{K}}_{uw} = \int_A \mathbf{G}_2^T \mathbf{W} \mathbf{B}_{w2} dA] \\
&\{ \tilde{\mathbf{K}}_{uu}^{uu} = \int_A \mathbf{B}_{u2}^{uT} \mathbf{U}^T \mathbf{G}_3 \mathbf{U} \mathbf{B}_{u2}^{u} dA ,\ \tilde{\mathbf{K}}_{uu}^{uu*} = \int_A \mathbf{G}_4 \mathbf{U} \mathbf{B}_{u2}^{u} dA , \\
&\tilde{\mathbf{K}}_{uu}^{uu**} = \int_A \mathbf{B}_{u2}^{uT} \mathbf{U}^T \mathbf{G}_4^T dA ,\ \tilde{\mathbf{K}}_{uu}^{vu} = \int_A \mathbf{G}_5 \mathbf{U} \mathbf{B}_{u2}^{u} dA , \\
&\tilde{\mathbf{K}}_{uu}^{uv} = \int_A \mathbf{B}_{u2}^{uT} \mathbf{U}^T \mathbf{G}_5^T dA ,\ \tilde{\mathbf{K}}_{wu}^{u} = \int_A \mathbf{B}_{w2}^T \mathbf{W}^T \mathbf{G}_6 \mathbf{U} \mathbf{B}_{u2}^{u} dA , \\
&\tilde{\mathbf{K}}_{uw}^{u} = \int_A \mathbf{B}_{u2}^{uT} \mathbf{U}^T \mathbf{G}_6^T \mathbf{W} \mathbf{B}_{w2} dA \} .
\end{aligned} \tag{5}
$$

where

$$[\,\mathbf{G}_1 = \mathbf{B}_{w1}^{T}\mathbf{D}_{11}\mathbf{B}_{w1}\,,\ \mathbf{G}_2 = \mathbf{B}_{w1}^{T}\mathbf{D}_{11}\mathbf{B}_{u1}\,], \tag{6}$$

$$\{\,\mathbf{G}_3 = \mathbf{B}_{u1}^{uT}\mathbf{D}_{11}\mathbf{B}_{u1}^{u}\,,\ \mathbf{G}_4 = \mathbf{B}_{u4}^{uT}\mathbf{D}_{11}\mathbf{B}_{u1}^{u}\,,\ \mathbf{G}_5 = \mathbf{B}_{u5}^{vT}\mathbf{D}_{11}\mathbf{B}_{u1}^{u}\,,\ \mathbf{G}_6 = \mathbf{B}_{w1}^{T}\mathbf{D}_{11}\mathbf{B}_{u1}^{u}\,\}$$

.

The geometric stiffness matrix of structure is built by summing overlapping terms of the component strip matrices; in the same way that conventional stiffness matrix of structure is built by summing terms of the conventional strip matrices using the transformation matrices between the local and global displacements [2].

Stability equations

Stability equations are derived from the virtual work principle and the strain energy methods. In order to obtain the stability equations from the variational relations, the principle of the stationary potential energy will be invoked. Since the principle of the stationary potential energy states that the necessary condition of the equilibrium of any given state is that the variation of the total potential energy of the considered structure is equal to zero, we have the following relation:

$$\delta\Pi = 0\ . \tag{7}$$

We conclude from Eq. (7) that, if the structure is given the small virtual displacements, the equilibrium still persists if an increment of the total potential energy of the structure $\delta\Pi$ is equal to zero. Eq. (7) is the basis to derive the variational equation of equilibrium of a structure, and it is correct for both the pre- and post-critical deformation states. Eq. (7) is satisfied for an arbitrary value of the variations of parameters $\delta\mathbf{q}_m^T$. Thus we have the following conditions, which must be satisfied for any harmonic *m*:

$$\frac{\partial\Pi}{\partial\mathbf{q}_m^T} = \mathbf{0}\ . \tag{8}$$

Next, we calculate derivatives of the total potential energy of a strip and finally, we get a non-homogeneous and nonlinear set of algebraic equations (9), which are the searched stability equations.

$$\left(\hat{\mathbf{K}}_{uu}\mathbf{q}_u + \hat{\mathbf{K}}_{ww}\mathbf{q}_w\right) + \left[1/2\,\tilde{\mathbf{K}}_{ww}\mathbf{q}_w + 1/2\,\tilde{\mathbf{K}}_{wu}\mathbf{q}_u + 1/4\,\tilde{\mathbf{K}}_{uw}\mathbf{q}_w\right] + \tag{9}$$

$$+\left\{\begin{matrix} 1/2\,\tilde{\mathbf{K}}_{uu}^{uu}\mathbf{q}_u^{u} + 3/4\,\tilde{\mathbf{K}}_{uu}^{uu^*}\mathbf{q}_u^{u} + 3/4\,\tilde{\mathbf{K}}_{uu}^{uu^{**}}\mathbf{q}_u^{u} + 1/4\,\tilde{\mathbf{K}}_{uu}^{vu}\mathbf{q}_u^{u} + \\ +1/2\,\tilde{\mathbf{K}}_{uu}^{uv}\mathbf{q}_u^{v} + 1/4\,\tilde{\mathbf{K}}_{wu}^{u}\mathbf{q}_u^{u} + 1/4\,\tilde{\mathbf{K}}_{uw}^{u}\mathbf{q}_w \end{matrix}\right\} - \mathbf{Q} = \mathbf{0}$$

We can visualize the construction of a strip stiffness matrix, which is composed of twelve block matrices. Assembling block matrices into conventional/geometric stiffness matrix of each strip is performed according to the scheme presented in Fig. 3, where: ST1= $\hat{\mathbf{K}}_{uu}$, ST2= $\hat{\mathbf{K}}_{ww}$, ST3= $\tilde{\mathbf{K}}_{ww}$, ST4= $\tilde{\mathbf{K}}_{wu}$, ST5= $\tilde{\mathbf{K}}_{uw}$, ST6= $\tilde{\mathbf{K}}_{uu}^{uu}$, ST7= $\tilde{\mathbf{K}}_{uu}^{uu*}$, ST8= $\tilde{\mathbf{K}}_{uu}^{uu**}$, ST9= $\tilde{\mathbf{K}}_{uu}^{vu}$, ST10= $\tilde{\mathbf{K}}_{uu}^{uv}$, ST11= $\tilde{\mathbf{K}}_{wu}^{u}$ and ST12= $\tilde{\mathbf{K}}_{uw}^{u}$ (ST5=ST4^T, ST8=ST7^T, ST10=ST9^T, ST12=ST11^T).

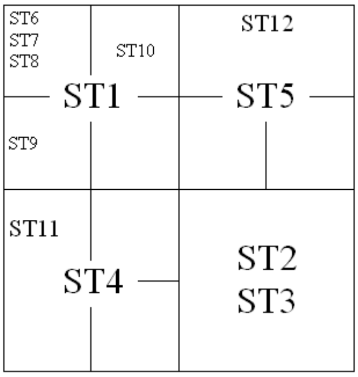


Fig. 3. Strip stiffness matrix assembling

Solution of nonlinear equations

For equilibrium, the principle of stationary potential energy requires that:

$$\mathbf{R} = \partial\Pi/\partial\mathbf{q}^T = \left[\hat{\mathbf{K}} + \tilde{\mathbf{K}}\right]\mathbf{q} - \mathbf{Q} = \mathbf{K}\mathbf{q} - \mathbf{Q} = \mathbf{0} \ . \tag{10}$$

where *Π* is a function of the displacements ***q***, and ***R*** represent the gradient or residual force vector, which is generally nonzero for some approximate displacement vector $\boldsymbol{q_0}$ (the subscript 0 denotes an old value). It is assumed that a better approximation is given by:

$$\mathbf{q}_n = \mathbf{q}_0 + \boldsymbol{\delta}_0 \ . \tag{11}$$

where subscript *n* denotes a new value.

Taylor's expansion of Eq. (10) yields:

$$\mathbf{R}_n = \mathbf{R}(\mathbf{q}_0 + \boldsymbol{\delta}_0) = \mathbf{R}(\mathbf{q}_0) + \bar{\mathbf{K}}_0\boldsymbol{\delta}_0 + ... = \mathbf{R}_0 + \bar{\mathbf{K}}_0\boldsymbol{\delta}_0 + ... \ . \tag{12}$$

where $\bar{\mathbf{K}}_0 = \partial\mathbf{R}/\partial\mathbf{q}$ is the matrix of second partial derivatives of *Π* calculated at $\boldsymbol{q_0}$ (i.e. the tangent stiffness matrix or Hessian matrix). Setting Eq. (12) to zero and considering only linear terms in $\boldsymbol{\delta_0}$ gives the standard expression for Newton-Raphson iteration:

$$\boldsymbol{\delta}_0 = -\bar{\mathbf{K}}_0^{-1}\mathbf{R}_0 \ . \tag{13}$$

Using this approach, a further iteration yields:

$$\boldsymbol{\delta}_n = \bar{\mathbf{K}}_n^{-1}\mathbf{R}_n \tag{14}$$

where $\bar{\mathbf{K}}_n = \partial\mathbf{R}/\partial\mathbf{q}$ at $\boldsymbol{q}_n$.

In addition, the blocks of the conventional and geometric tangent stiffness matrix of each strip are:

$$\bar{\hat{\mathbf{K}}} = \begin{bmatrix} \hat{\mathbf{K}}_{uu} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{K}}_{ww} \end{bmatrix}. \tag{15}$$

$$\begin{aligned} \bar{\tilde{\mathbf{K}}} = {} & \begin{bmatrix} \mathbf{0} & 1/2\,\tilde{\mathbf{K}}_{uw} \\ 1/2\,\tilde{\mathbf{K}}_{wu} & 3/2\,\tilde{\mathbf{K}}_{ww} \end{bmatrix} + \\ & + \begin{bmatrix} 3/2\,\tilde{\mathbf{K}}_{uu}^{uu} + 3/2\,\tilde{\mathbf{K}}_{uu}^{uu*} + 3/2\,\tilde{\mathbf{K}}_{uu}^{uu**} & 1/2\,\tilde{\mathbf{K}}_{uu}^{uv} \\ 1/2\,\tilde{\mathbf{K}}_{uu}^{vu} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & 1/2\,\tilde{\mathbf{K}}_{uw}^{u} \\ 1/2\,\tilde{\mathbf{K}}_{wu}^{u} & \mathbf{0} \end{bmatrix} \\ & . \end{aligned} \tag{16}$$

Comparing these expressions with Eq. (9), it is apparent that the conventional stiffness matrix remains unchanged, while the geometric tangent matrix becomes symmetrical.

Static buckling

The loss of stability of static equilibrium states of structures subjected to conservative loads is in general known as static buckling of the structure. For conservative systems, the principle of minimum of the total potential energy can be used to test the stability of a structure (static equilibriums are extremes of the total potential energy). The Hessian with respect to the local DOFs is denoted as the tangent stiffness matrix, of each strip i.e.:

$$\bar{\mathbf{K}} = \begin{bmatrix} \bar{\hat{\mathbf{K}}} + \bar{\tilde{\mathbf{K}}} \end{bmatrix}. \tag{17}$$

The (local) stability of equilibrium states of conservative systems by HCFSM can be assessed by looking at the eigenvalues of the tangent stiffness matrix of structure $\bar{\mathbf{K}}_{(DOFs\cdot r\cdot(n+1))\cdot(DOFs\cdot r\cdot(n+1))}$ with (*n*+1) nodal lines, which are all real, since tangent stiffness matrix of the strip is a symmetric matrix.

Fig. 4 shows a rectangular plate divided into (*n*) finite strips with (*n*+1) nodal lines, where the plate is assumed to be simply supported on edges *y*=0 and *y*=*a*.

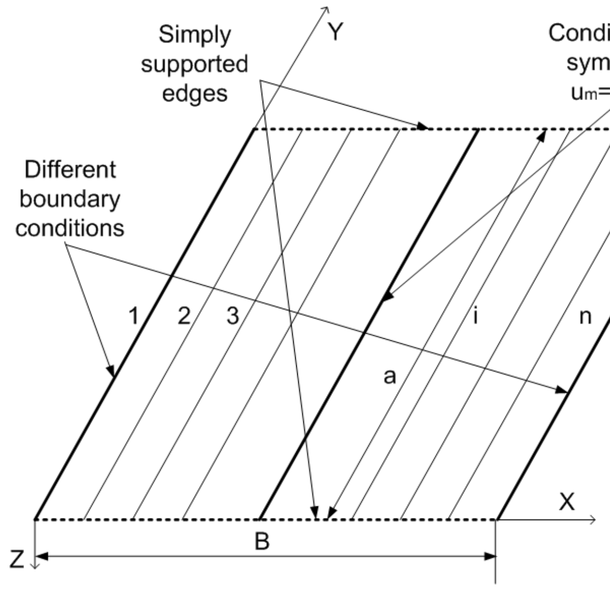


Fig. 4. Discretized plate with a pair of simply supported edges

Let λ_i denote the ith eigenvalue of

$$\bar{\mathbf{K}}_{(DOFs\cdot r\cdot(n+1))\cdot(DOFs\cdot r\cdot(n+1))}\ . \tag{18}$$

Based on theorems of Lagrange-Dirichlet and Lyapunov [14] it can be concluded that an equilibrium state is stable if all λ_i>0, while an equilibrium state is unstable if one or more λ_i<0. If along a load-path (IINCS=1, NINCS), at some equilibrium state one or more λ_i=0, this equilibrium state is denoted as a critical state. Static buckling refers in general to case where, starting from some stable state, a critical state is reached along the load-path.

In general, static buckling is solved by solving Eq. (9) for a varying load *P* with for example some sort of numerical path-following routine [2], while simultaneously tracking the eigenvalues of the tangent stiffness matrix of structure given by Eq. (18). Buckling occurs where the matrix becomes singular.

3. HCFSM Software Parallelization

3.1. HCFSM Software

The HCFSM software enables large displacement stability analysis of orthotropic prismatic shells with simply supported boundary conditions along the diaphragm-supported edges. It consists of four modules as shown in Fig. 5.

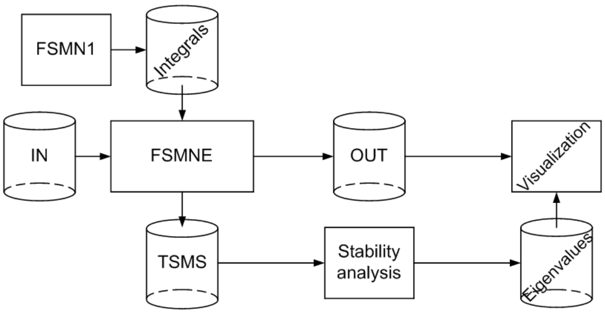


Fig. 5. Overview of HCFSM software

The first module, called FSMN1 program, calculates the values of the integrals using Romberg integration, and stores them in the memory for later computations. The second module, called FSMNE program, calculates displacements and inner forces for each load-part by solving the stability equation (9). The stored integrals are used for the computation of each strip tangent stiffness matrix in FSMNE. The corresponding tangent stiffness matrices of the structure (TSMS) are prepared for the solution of the stability equation and also saved in a file to be analyzed by the third module, called Stability analysis program. The Visualization module provides comparative graphic presentation of the results.

3.2. Scope of FSMN1 program parallelization

The number of integral values is proportional to the power of four of the number of harmonics . For a large number of harmonics computation of the integral values can take a substantial amount of time, even with modern hardware resources. However, calculations of the integral values are independent and can be done in parallel.

To avoid repeating computations the integral values are stored in a file once they are calculated. In this way, every time these integral values are needed for the stiffness matrices calculations, they are not calculated but read from the file.

To avoid numerous readings of the file, the integral values can be read from the file just once and placed in the memory (RAM). This approach significantly reduces the execution time (compared to the previous one) because all integral values are used for every strip stiffness matrix calculation in the every iteration and for every load. The drawback of this approach is that it is not generally applicable as the number of integral values outgrows the capacity of available RAM for higher number of harmonics.

The general solution should provide (reasonably fast) stiffness calculation for any (reasonable) number of harmonics. In our vision this can be achieved by enabling heterogeneous computation where applications utilize the power of both CPU and GPU. These units are treated as separate devices with their own memory spaces allowing for simultaneous computation on both CPU and GPU without contention for memory resources. In particular, using CUDA programming model we are in position to divide independent computations of integral values into different functions (kernels) executed in parallel on general purpose GPUs. Such generated values are used to feed up FSMNE program running on the host processors (CPUs). The globally shared memory located on a GPU device is used to transfer data between the host and device for kernel input and output [6].

For high performance of the system the memory bandwidth between CPU and GPGPU is of crucial importance. To achieve faster communication pipelining we use double buffering technique for the calculation of integral values. Integral values are calculated and placed in two buffers that are used one after the other. At the beginning both buffers are empty. The kernel programs (executing on GPGPU) starts filling the first buffer. When the first buffer is full, FSMNE program (executing on CPU) starts reading integrals from it. At the same time the kernel programs write the next block of integral values to the second buffer. When the first buffer is exhausted, FSMNE program continues to read integrals from the second buffer if it is filled in the mean time. Otherwise, FSMNE program waits for it to be filled. Eventually the kernel programs write the next block of integral values to the second buffer and FSMNE program continues reading them.

The main advantage of using CUDA programming model described above is that it does not require any disk space or memory for storing calculated integral values [8]. Also, it can generate integral values up to 40 times faster than the corresponding sequential approach. However, for efficient implementation of this model we must assure wait-free communication between CPU and the globally shared memory and that represents a major research challenge.

3.3. Scope of FSMNE program parallelization

The essence of FSMNE program can be described as solution of the stability equation (9). The program takes as input parameters the definition of the structure along with the planned load distribution. As a result the program produces the displacement vectors and the vectors of inner forces. The solution for each load-part is obtained using an iterative approach as described in chapter 2.2. Each iteration includes solving the system of the stability equations (SSE) having in mind that TSMS element values depend on the displacements calculated in the previous iteration.

The computational complexity of the problem depend on the number of strips (NELEM) and the number of harmonics (NTERM). Analysis of the program execution reveals that the most time consuming part represents the computation of TSMS, whereas much shorter time is spent on finding the solution of SSE by the method of Gaussian elimination [9]. As the number of strips grows the time necessary for finding the solution of SSE becomes longer. However, the stiffness matrices computation of separate strips stays dominant.

TSMS is composed of the stiffness matrices for each strip. Computation of these stiffness matrices can be based on the von Karman or the Green-Lagrange prediction. The higher complexity of the Green-Lagrange formulation [9] comes from the fact that each stiffness matrix is formed from the 12 block matrices according to the scheme from Fig. 3. The von Karman prediction requires only the first five block matrices. The block matrices are named by the acronyms STi (i=1, ... , 12).

Computations of the stiffness matrices of separate strips, as well as computations of their parts (different STi block matrices), are mutually independent. Therefore, due to its dominant part in the execution time of FSMNE program, they offer natural bases for FSMNE program parallelization. Computation of the separate strip stiffness matrices is suitable for application of multicomputer – MPI approach [7, 9], while computations of the separate STi block matrices are suitable for application of multiprocessor – OpenMP approach [9].

MPI parallelization

The multicomputer based computation of the stiffness matrices of separate strips implies engagement of different node computers to calculate the stiffness matrices of different strips. By default, one of the nodes takes the role of “master” while others become “slaves”. Each slave node does the same. It computes the elements of the stiffness matrices for a particular range of strips. The master node initiates these computations, collects the results from the slave nodes, composes TSMS and performs other non-parallelized tasks (e.g. solving the SSE).

Parallel execution of FSMNE program requires communication between the master and the slave nodes. Therefore, the total parallel FSMNE program execution time (*TP_{total}*) includes the communication time (*TP_c*), the stiffness matrices computation time (*TP_s*) and the rest tasks execution time (*TP_r*), while the total sequential FSMNE program execution time (*TS_{total}*) includes only the stiffness matrices computation time (*TS_s*) and the rest tasks execution time (*TS_r*):

$$\mathrm{TP_{total} = TP_c + TP_s + TP_r}\,. \tag{19}$$

$$\mathrm{TS_{total} = TS_s + TS_r}\,. \tag{20}$$

The effects of FSMNE program parallelization are under influence of various factors: the numbers of strips and harmonics on one side and the number of engaged nodes on the other. It is important to analyze the impact of these factors on the different execution times – TP_{total}, TP_c, TP_s and TP_r. The increase of the strips number is followed by the increase of TP_s and TP_r

due to higher dimensions of the total stiffness matrix and the longer time needed to find the solution of SSE. The same trend connects increase of the harmonics number and increase of TP_s. Increase of the number of nodes causes increase of TP_c and decrease of TP_s. The impact of nodes number increase on TP_{total} depends on the dominant factor: either TP_c increase or TP_s decrease.

From the standpoint of the Amdahl's law [15], the maximum speedup of FSMNE program is limited by the following formulas:

$$\frac{n}{1+f\cdot(n-1)}\ , \tag{21}$$

$$n=\frac{TS_s}{TP_s}\ ,\ f=\frac{TS_r}{TS_{total}}\ .$$

where *f (0 < f < 1)* is the fraction of time spent in the serial part (before the improvement) and *n* represents the speedup of the parallel part of FSMNE program.

From the standpoint of the parallelization effects it is important to distribute load evenly on multicomputer nodes. The maximum degree of parallelization is achieved when the strip number is divisible by the node number. In contrary, when the node load is not even, the (theoretic) speedup depends on the heaviest loaded node due to the effect of barrier synchronization. For example, in the case with ten strips, the speedup for two nodes is 2, the speedup for three nodes is 2.5, the speedup for four nodes is 3.3, the speedup for five nodes is 5, and the speedup for six to nine nodes is again 5, while the speedup for ten nodes is 10.

Effects of MPI parallelization of FSMNE program for the von Karman and the Green-Lagrange prediction are presented on Fig. 6. The presented results are related to the example of prismatic shell in the form of folded plate cross-section [9]. This shell is divided into 20 strips, and the computation is done for 31 harmonic.

From Fig. 6 follows that in the case of the von Karman prediction TP_{total} decreases when the node number increases due to the fact that TP_s represents the significant part of TP_{total}, so the effects of parallelization of stiffness matrices computation are noticeable. These effects are not canceled by the increase of TP_c that follows the node number increase. Therefore MPI parallelization of von Karman prediction is valid in the discussed example.

In the case of the Green-Lagrange prediction TP_{total} significantly decreases when the node number increase, due to the fact that TP_s is the prevalent part of TP_{total}. Thus, we can expect that the effects of parallelization of the stiffness matrices computation are very high. These effects are not under substantial influence of the TP_c increase. Therefore MPI parallelization of the Green-Lagrange prediction is obviously useful in the discussed example.

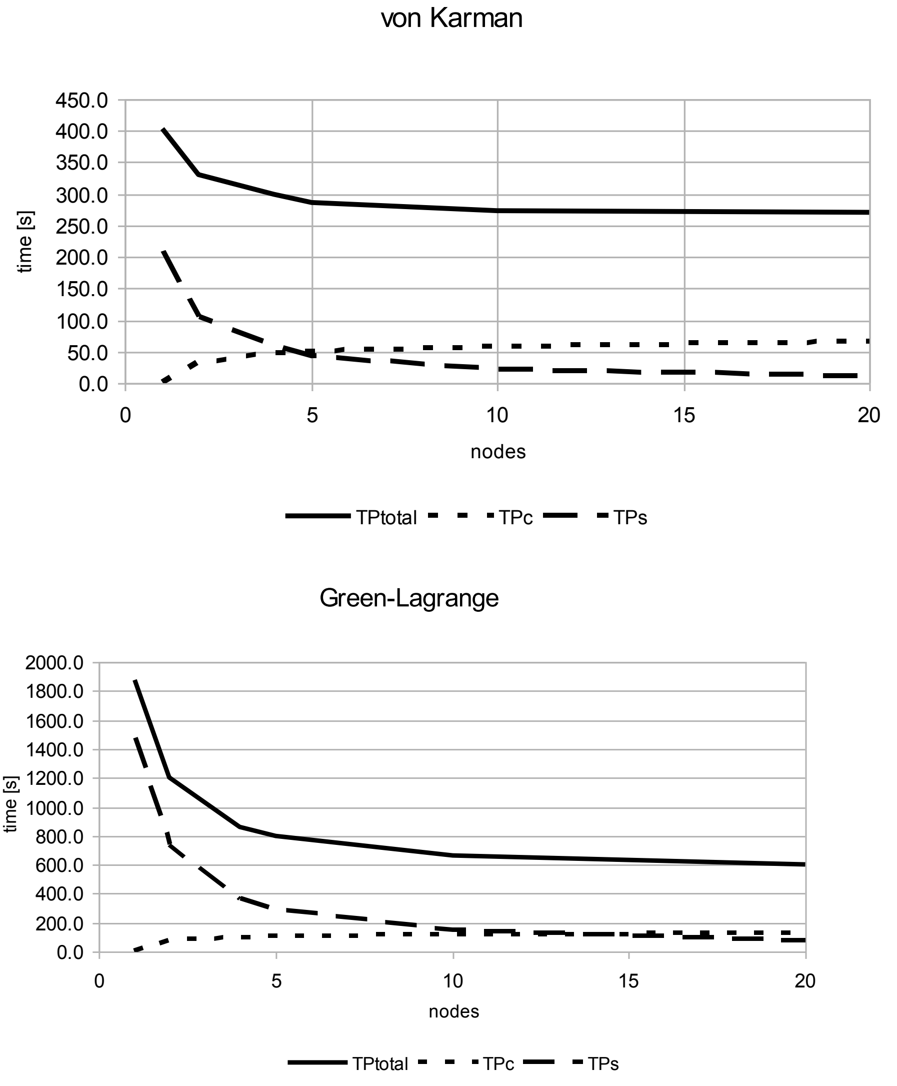


Fig. 6. TP_{total}, TP_c and TP_s for the von Karman and the Green-Lagrange predictions (the presented results are obtained on single CPU core per node multicomputer)

Fig. 6 suggests that in the case of the von Karman, as well as the Green-Lagrange prediction, increase of the node number, when the strip number is high enough, leads to the cancellation of the effect of parallelization of the stiffness matrices computation, due to the TP_c increase. However, Fig. 7 indicates that the harmonic number increase delays such event, due to increase of the stiffness matrices computation complexity.

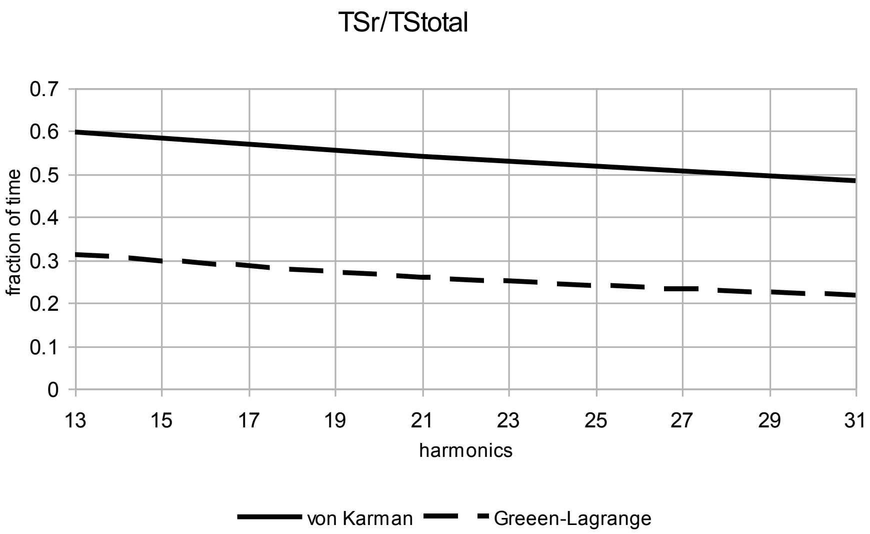
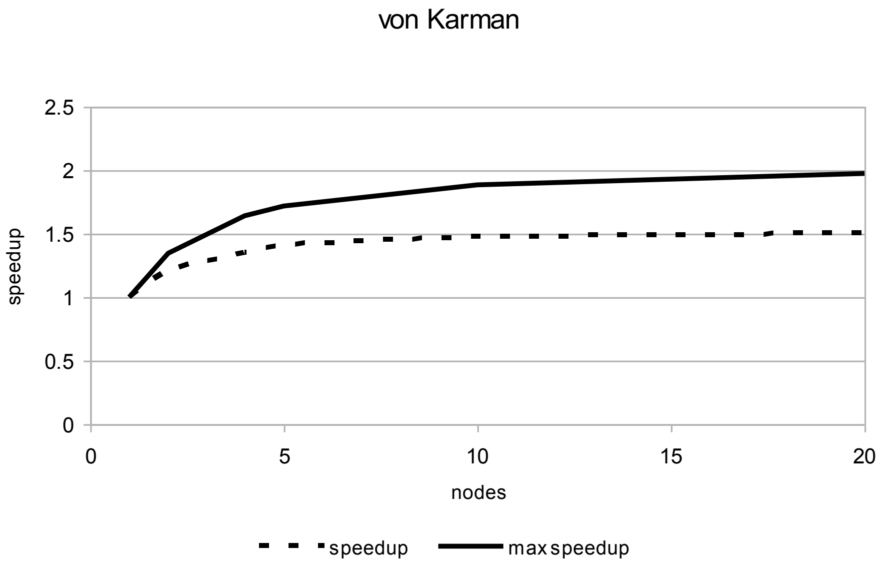


Fig. 7. The impact of the harmonic number increase on the parallelization potential



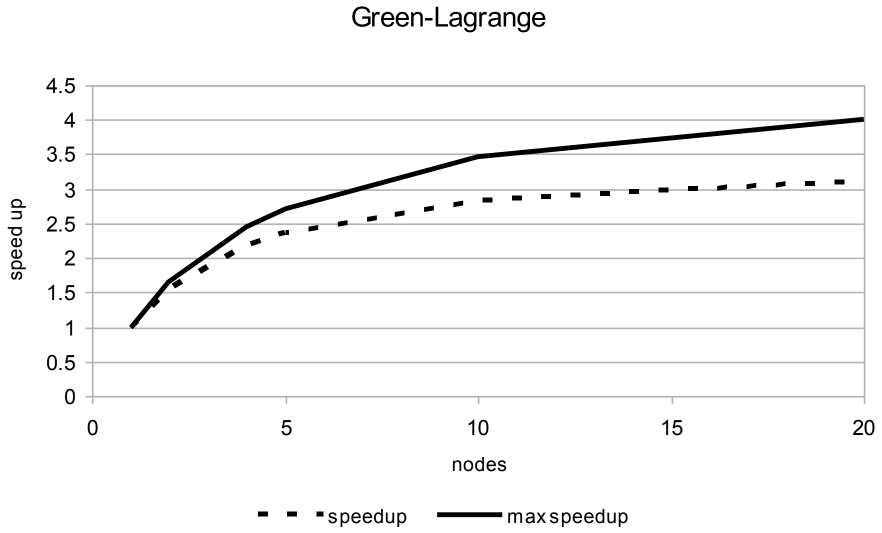


Fig. 8. The speedups of the von Karman and the Green-Lagrange predictions

Fig. 8 suggests that the speedup of the Green-Lagrange prediction is doubled in comparison to the speedup of the von Karman prediction. This is the consequence of the fact that the Green-Lagrange prediction triples the computation time of stiffness matrices in comparison to the von Karman prediction.

OpenMP parallelization

The multiprocessor based computation of STi block matrices in the Green-Lagrange prediction implies distribution of different STi block matrices computation to different cores of a multi-core processor. The complexity of these block matrices varies which reflects on their execution times. Table 1 contains percentages of the total execution time spent in the STi block matrices.

Table 1. Percentage of total execution time spent in the STi block matrices

ST1	ST2	ST3	ST4	ST5	ST6	ST7	ST8	ST9	ST10	ST11	ST12
0.01	0.01	27.38	0.88	0.83	15.15	0.74	0.74	0.74	0.74	24.03	22.56

Table 1 suggests that there is no sense to devote a separate core to every STi block matrices, since the computation times of these block matrices are vary disparate and would keep the majority of cores idle. Table 2 presents a possible distribution of the STi block matrices computations to four evenly loaded cores.

Table 2. A distribution of the STi block matrices computations to different cores

core	STi block matrices	%
1	ST3	27.38
2	ST1, ST2, ST4, ST5, ST6, ST7, ST8	18.36
3	ST9, ST10, ST12	24.04
4	ST11	24.03

Table 2. implies that almost 94% of the stiffness matrices computation time can be parallelized. The remaining 6% of the time is used for preparation of the STi block matrices computations.

The effects of OpenMP parallelization of FSMNE program for the Green-Lagrange prediction are presented on Fig. 9. The presented results are related to the already mentioned example of prismatic shell.

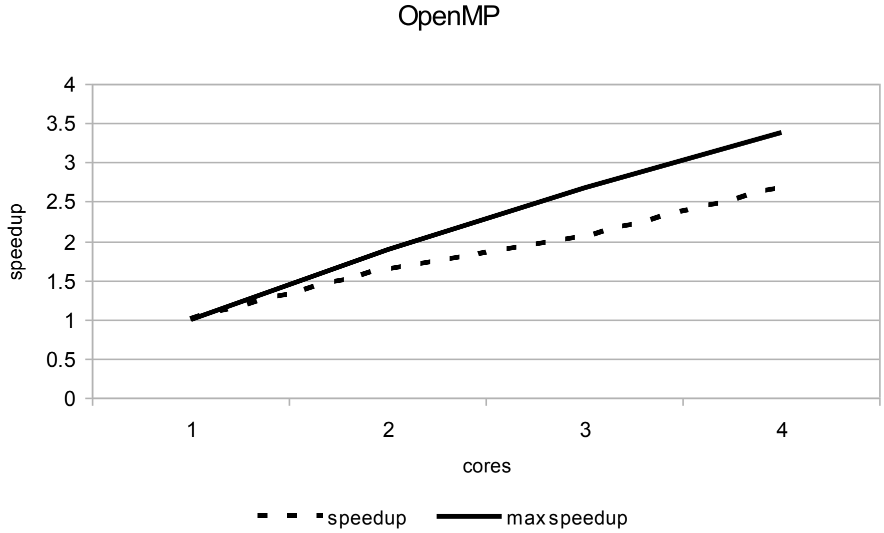


Fig. 9. Results of OpenMP parallelization

The measured speedup follows the Amdahl maximal speedup. Introduction of the third core results in the smaller speedup than expected. The reason is that the data are not evenly distributed among cores anymore, and therefore a good load balancing is not achieved.

It is worth mentioning that as the number of strips grows, the time necessary for finding the solution of SSE grows too. The fact that SSE are solved by master node and our insight to CUDA programming model leads us to the idea of exploring the potential of applying CUDA programming model on the process of solving SSE as the subject of a future work. But for huge number of strips it might be reasonable to explore usage of MPI solver to the same problem.

Acknowledgements. The work reported in this paper is a part of the investigation within the research projects: ON174027 "Computational Mechanics in Structural Engineering" and TR 36017 "Utilization of by-products and recycled waste materials in concrete composites in the scope of sustainable construction development in Serbia: investigation and environmental assessment of possible applications", supported by the Ministry for Science and Technology, Republic of Serbia. This support is gratefully acknowledged.

4. Conclusions

This paper presents a parallelization of finite-strip analysis of geometric nonlinear folded-plate structures applying the von Karman and Green-Lagrange strains, which are both characterized by the coupling of all harmonics. The coupling of all series terms highly increases calculation time in an existing finite-strip sequential program when a large number of series terms are used and justify parallelization efforts.

The HCFSM offers good potential for MPI, OpenMP and CUDA parallelization. Therefore it is not surprise that the MPI/OpenMP/CUDA hybrid approach shows good results in the parallelization that are illustrated on the example folded plate cross-section. The future paper will present cumulative effects of contemporary application of all three discussed parallelization techniques. The further investigation will analyze the negative effects of communication bottlenecks that arise with increase of the number of nodes and will try to answer the question how to postpone such negative effects.

References

1. Cheung, Y.K., Tham, L.G.: Finite Strip Method. Boca Raton: CRC Press. (1998)
2. Milašinović, D.D. The Finite Strip Method in Computational Mechanics. Faculties of Civil Engineering: University of Novi Sad, Technical University of Budapest and University of Belgrade: Subotica, Budapest, Belgrade. (1997)
3. Chen, H.C., Byreddy, V.: Solving plate bending problems using finite strips on networked workstations. Computers & Structures; 62:227-236. (1997)
4. MPI Forum (2009): MPI: A Message-Passing Interface Standard, Version 2.2. http://www.mpi-forum.org/docs/mpi-2.2/mpi22-report.pdf.
5. OpenMP Specifications, http://www.openmp.org/.
6. Garland, M., Grand, S.L., Nickolls, J., Anderson, J., Hardwick, J., Morton, S., Phillips, E., Zhang, Y., Volkov, V.: Parallel Computing Experiences with CUDA. IEEE Micro. IEEE Computer Society, 28:13-27. (2008)
7. Rakić, P., Milašinović, D.D., Živanov, Ž., Hajduković, M.: MPI-CUDA Parallelisation of the Finite Strip Method for Geometrically Nonlinear Analysis. in B.H.V. Topping, P. Iványi, (Editors), "Proceedings of the First International Conference on Parallel, Distributed and Grid Computing for Engineering", Civil-Comp Press, Stirlingshire, UK, Paper 33. doi:10.4203/ccp.90.33. (2009)
8. Rakić, P.S., Milašinović, D.D., Živanov, Ž., Suvajdžin, Z., Nikolić, M. Hajduković, M.: MPI-CUDA parallelization of a finite-strip program for geometric nonlinear

analysis: A hybrid approach. Advances in Engineering Software; 42(5):273-285. (2011)

9. Nikolić, M., Milašinović, D.D., Živanov, Ž., Marić, P., Hajduković M., Borković A., Milaković I.: MPI/OpenMP Parallelisation of the Harmonic Coupled Finite-Strip Method, in P. Iványi, B.H.V. Topping, (Editors), "Proceedings of the Second International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering", Civil-Comp Press, Stirlingshire, UK, Paper 94. doi:10.4203/ccp.95.94. (2011)
10. Davies, J.M.: Generalized Beam Theory (GBT) for Coupled Instability Problems, in: Rondal, J. (Ed.), Coupled Instabilities in Metal Structures, Theoretical and Design Aspects, CISM Courses and Lectures N. 379, p.151-223, Wien – NY: Springer – Verlag. (2000)
11. Rasmussen, K.J.R., Hancock, G.J.: Buckling analysis of thin-walled structures: numerical developments and applications, Progress in Structural Engineering and Materials, V.2, pp. 359-368. (2000)
12. Schafer, B.W.: Progress on the Direct Strength Method, Proceedings of the 16th International Specialty Conference on Cold Formed Steel Structures, Orlando, Fl., U.S.A., pp. 647-662. (2002)
13. Milašinović, D.D.: Geometric non-linear analysis of thin plate structures using the harmonic coupled finite strip method. Thin-Walled Structures; 49(2):280-290. (2011)
14. Bažant, Z.P., Cedolin, L.: Stability of Structures: Elastic, Inelastic, Fracture, and Damage Theories. Oxford University Press, Oxford. (1991)
15. Amdahl, GM.: Validity of the single processor approach to achieving large scale computing capabilities. Proc. AFIPS 1967 Spring Joint Computer Conf. 30 (April), Atlantic City, N.J, 483-485.

Dr Dragan D. Milašinović, born 1954. received his Civil Eng. degree from the Univ. of Sarajevo in 1978 and Ph.D. in 1988. He is currently a faculty member at the Univ. of Novi Sad where he is an Prof. of Plates and Shells Theory and Concepts and Applications of Finite Element Analysis.

Dr Miroslav P. Hajdukovic is full professor at Faculty of Technical Sciences, University of Novi Sad, the Chair for Applied Computer Science. He graduated in 1977 at the Faculty of Electrical Engineering in Sarajevo, finished post-graduate studies in 1980 at the Faculty of Electrical Engineering in Sarajevo, and earned PhD degree in 1984 at the Faculty of Electrical Engineering in Sarajevo. His scope of interest includes computer architecture, operating systems, distributed systems and compilers.

Dr Predrag Rakić is assistant professor at Department of Applied Computer Science at the University of Novi Sad. He received his M.Sc. degree in the operating systems field and Ph.D. in the area wireless sensor networks, both from the University of Novi Sad. His research interests include parallel and distributed computing, generic and template meta programming, data modeling for high performance computing, etc.

Miloš Nikolić is a PhD student/research assistant at École Polytechnique Fédérale de Lausanne (EPFL) since 2010. Prior joining EPFL, he worked as a research assistant at the Chair for Applied Computer Sciences, Faculty of Technical Sciences, University of Novi Sad. His research interest is in the areas of parallel and distributed systems, cloud computing, databases and compilers.

Lazar Stričević has been employed as a teaching assistant at the Faculty of Technical Sciences, University of Novi Sad, the Chair for
Applied Computer Science since 2004. His interests include computer architecture, operating systems, computer networks, distributed systems.

Žarko Živanov is teaching assistant at Faculty of Technical Sciences, University of Novi Sad, the Chair for Applied Computer Science since 2000. His interests include operating systems, computer architecture, compilers and data visualisation.

Received: August 22, 2011; Accepted: May 11, 2012.

DOI: 10.2298/CSIS110906012K

Describing Papers and Reviewers' Competences by Taxonomy of Keywords

Yordan Kalmukov

Department of Computer Systems and Technologies,
University of Ruse,
8 Studentska Str., 7017 Ruse, Bulgaria,
JKalmukov@gmail.com

Abstract. This article focuses on the importance of the precise calculation of similarity factors between papers and reviewers for performing a fair and accurate automatic assignment of reviewers to papers. It suggests that papers and reviewers' competences should be described by taxonomy of keywords so that the implied hierarchical structure allows similarity measures to take into account not only the number of exactly matching keywords, but in case of non-matching ones to calculate how semantically close they are. The paper also suggests a similarity measure derived from the well-known and widely-used Dice's coefficient, but adapted in a way it could be also applied between sets whose elements are semantically related to each other (as concepts in taxonomy are). It allows a non-zero similarity factor to be accurately calculated between a paper and a reviewer even if they do not share any keyword in common.

Keywords: taxonomy of keywords; semantic similarity; objects' description and classification; automatic assignment of reviewers to papers; conference management systems.

1. Introduction

One of the most important and challenging tasks in organizing scientific conferences is the assignment of reviewers to papers. It plays a crucial role in building a good conference image. For highly-ranked conferences, having a low acceptance ratio, it is very important that each paper is evaluated by the most competent, in its subject domain, reviewers. Even a small inaccuracy in the assignment may cause serious misjudgements that may dramatically decrease the conference image and authors' thrust in that event.

Generally the assignment could be done both manually and automatically. Manual assignment is applicable for small conferences having a small number of submitted papers and reviewers well known to the Programme Committee (PC) chairs. It requires the latter to familiarize themselves with all papers and reviewers' competences then assign the most suitable reviewers to each paper while maintaining a load balancing so that all reviewers evaluate roughly the same number of papers. Doing that for a large number of papers and reviewers is not just hard and time consuming, but due to the many constraints (accuracy, load balancing, conflict of interests and etc.) that should be taken into an account the manual assignment gets less and less accurate with increasing the number of papers and reviewers - a motive strong enough to force the developers of all modern commercially available conference management systems to offer an automatic assignment of reviewers to papers.

The non-intersecting sets of papers and reviewers can be represented by a complete weighted bipartite graph (figure 1), where P is the set of all submitted papers and R – the set of all registered reviewers. There is an edge from every paper to every reviewer and every edge should have a weight. In case of a zero weight, the corresponding edge may be omitted turning the graph to a non-complete one. The weight of the edge between paper p_i and reviewer r_j tells us how competent (suitable) is r_j to review p_i. This measure of suitability is called a similarity factor. The weights are calculated or assigned in accordance with the chosen method of describing papers and reviewers' competences [7].

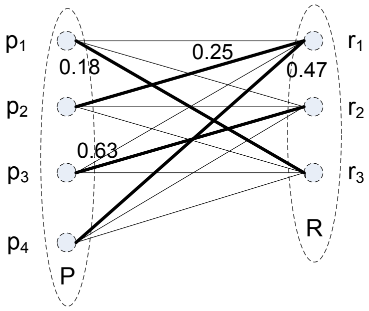


Fig. 1. The sets of papers (P) and reviewers (R) represented as a complete weighted bipartite graph. The edges in bold are the actual assignments suggested by an assignment algorithm. All edges have weights but just those of the assignments are shown for clearness.

Once the weights are calculated the automatic assignment could be easily implemented by using assignment algorithms known from the graph theory – for example the Hungarian algorithm of Kuhn and Munkres [8, 11] or heuristic algorithms like the one proposed in [5, 6]. Some advanced algorithms do really guarantee finding the best possible assignment for a given set of weighted edges, thus the weights' accuracy turns to be a key point in performing an accurate automatic assignment.

This article proposes a method of describing papers and reviewers' competences, based on hierarchically-structured set (taxonomy) of keywords. It significantly increases the weights' accuracy in comparison to other methods used in the existing conference management systems. The suggested similarity measure (equations 3 to 7) is derived from the well-known and widely-used Dice's coefficient, but it could be also applied over sets whose elements are semantically related to each other. It allows a non-zero similarity factor to be accurately calculated between a paper and a reviewer even if they do not share any keyword in common.

From now on the terms "conference topics" and "keywords" are used interchangeably.

2. Related Work

2.1. Conference and Journal Management

In respect to paper submission and review, conference management and journal management have pretty much in common. The key difference however is in the assignment of reviewers to papers. Conferences have a paper submission deadline. Once the deadline has been reached there are N number of submitted papers and M number of registered reviewers. The assignment is then handled as an assignment problem [8, 11, 5, 6] in bipartite graphs (as shown on figure 1). In contrast, journals usually do not have a submission deadline and papers are processed individually. The lack of clusters of papers makes the automatic assignment almost needless. Instead the journal management system should be able just to identify the most suitable reviewers (who are both competent and not overloaded) for a specified article without handling the assignment as a global optimization problem as the conference management system does. Despite that calculating similarity factors between a specific paper and all of the reviewers as accurate as possible is still very important for finding the most suitable reviewers to evaluate the paper.

One of the well known and probably the most used journal management system is Editorial Manager (EM) [27]. It is fully featured and highly configurable software that provides literally all editors need for flexible management. EM relies on hierarchical classification of terms to describe articles and reviewers' competences and to find the most competent reviewers. In contrast to the proposed solution however EM just takes into account the number of exactly matching classification nodes and omits the semantic relationships between them. Thus if the manuscript is described by a classification node and a reviewer has selected its child node (for example), then according to EM the article and the reviewer will have nothing in common as the number of matching nodes is 0. However as the node selected by the reviewer is a child node of the one describing the manuscript then the reviewer should be suitable to evaluate it – a statement based on the assumption that if the reviewer has an expertise in a specific field then he/she should have sufficient knowledge in the more general one as well. In this case EM suggestions seem not to be quite correct.

2.2. Describing Papers and Reviewers' Competences in Conference Management Systems

Most of the conference management systems (CMSs) are actually online, often cloud-based, conference management services. End users do not get any software but a completely hosted service. Thus analyzing existing solutions is mostly based on their official documentation and/or user experience. Some systems have very detailed documentation while others do not reveal any technical issues or algorithms at all.

A detailed review and comparative analysis of the existing methods of describing papers and reviewers' competences could be found in [7]. The phrase "method of describing papers and competences" includes general concepts, algorithms, data structures, mathematical formulas, proper organization of the user interface and etc. so that all these allow authors not just to submit a single file, but to outline what the paper is about and reviewers to state the areas of science they feel competent to review papers in [7].

Generally the methods of describing could be divided into two main groups:

- Explicit methods – require users to explicitly outline their papers and/or competences, i.e. to provide some descriptive metadata;
- Implicit methods – intelligent methods that automatically extract the required descriptive data from the papers and from reviewers' previous publications available on the Internet.

2.2.1 Implicit Methods of Describing Papers and Competences

The implicit methods of describing papers and competences use intelligent techniques to fetch the required metadata from the papers themselves or from the Internet. They usually perform a text analysis of papers' content and/or online digital libraries, indexes and others resources - DBLP [25], ACM Digital Library [20], CiteSeer [33], Google Scholar [28], Ceur WS [21] and etc.

Andreas Pesenhofer et al. [12] suggest that *the interest of a reviewer can be identified based on his/her previous publications available on the Internet*. The proposal suggests that the reviewer's name is used as a search query sent to CiteSeer and Google Scholar to obtain his/her publications. Then Euclidian distance (used as a measure of similarity) between the titles of every submitted paper and every reviewer's publication is calculated based on the full-text indexed feature vector [12].

Stefano Ferilli et al. [3] suggest a similar solution - *paper topics are extracted from its title and abstract, and expertise of the reviewers from the titles of their publications available on the Internet*. The proposed method assumes there is a predefined set of conference topics and *it tells which topics exactly apply to which papers / reviewers*. It is done by applying a Latent Semantic Indexing (LSI) [30] over the submitted papers' title and abstract and the titles of reviewers' publications fetched from DBLP [25].

Evaluated by the IEA / AIE 2005 conference organizers, the result showed a *79% accuracy on average*. *As to the reviewers the resulting accuracy was 65%* [3].

Marko Rodriguez and Johan Bollen [16] suggest that *a manuscript's subject domain can be represented by the authors of its references*. In comparison to the previous two methods this one does not exploit the titles and the abstracts, but the names of the authors who appear in the reference section. The approach uses a co-authorship network, initially built from the authors' names in the reference section of the paper; then their co-authors, the co-authors of the co-authors and etc. are fetched from DBLP, and a relative-rank particle-swarm algorithm is run for finding the most appropriate experts to review the paper.

Implicit methods are convenient and time saving but *they rely on external data sources on the Internet that are more or less inertial and contain sparse information.* New papers and articles are indexed with months delay and not by one and the same bibliographic index. That results in incomplete papers' and/or reviewers' profile. Rodriguez and Bollen report that for JCDL 2005 *89% of the PC members and only 83% of the articles with bid data were found in DBLP* [16]. Pesenhofer states, for ECDL 2005, *for 10 out of 87 PC members no publications have been retrieved from CiteSeer and Google Scholar* [12]. Obviously, for those not found within the bibliographic indexes, assignment was at random.

All commercially available conference management systems rely on explicit methods of describing papers and competences. Implicit methods however could be very useful for automatic detection of conflicts of interest and a couple of systems employ them for exactly that purpose.

2.2.2 Explicit Methods of Describing Papers and Competences

Explicit methods require authors and reviewers to provide additional metadata in order to describe their papers respectively competences. Existing CMSs use the following three different ways of providing such metadata:

- Bidding / rating papers
  (reviewers state their interest / preference to each paper individually);
- Choosing keywords from a predefined unordered set of conference topics;
- Combined - topics + bidding.

Bidding requires reviewers to browse the list of all papers and to indicate if they would like to review or not each one of them. Reviewers usually browse just titles and abstracts and state their willing to review by selecting an option from a drop down menu (figure 2).

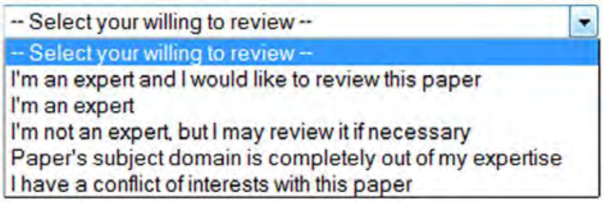

-- Select your willing to review --
-- Select your willing to review --
I'm an expert and I would like to review this paper
I'm an expert
I'm not an expert, but I may review it if necessary
Paper's subject domain is completely out of my expertise
I have a conflict of interests with this paper

Fig. 2. A drop down menu allowing reviewers to indicate how willing they are of reviewing a specified paper

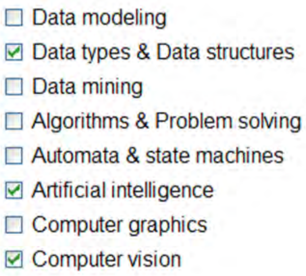
A similarity factor directly corresponds to each one of the options, where the highest factor is assigned to the highest level of willing to review. This method is considered to be the most accurate way of determining similarity factors as nobody knows better than the reviewer himself if he is competent to review the specified paper or not. It however does not really describe papers nor reviewers, but just the explicitly-stated relationship between them. If a paper has not been rated by enough reviewers then they will be assigned to it at random as there is no way the conference management system to determine the paper's subject domain or reviewers' competences.

Philippe Rigaux proposes a clever improvement of the bidding process which is trying to overcome the random assignments. The proposal is called an Iterative Rating Method (IRM) [15]. Its basic idea is to predict the missing ratings by iteratively applying a collaborative filtering algorithm to the ratings that has been explicitly provided by reviewers or the ratings that have been previously predicted (within previous iterations). The iterative rating method is fully implemented in the very popular The MyReview System [34].

Describing papers and reviewers' competences by a set of conference topics assumes that PC chairs set up a predefined list (unordered set) of keywords that best describe the conference coverage area. During paper submission authors are required to select those topics that best outline their papers. Reviewers do the same - during registration they are required to select the keywords corresponding to the areas of science they are competent in. Topics could be user-weighted or not. Non-weighted topics are usually selected from HTML checkboxes (figure 3). The user-weighted keywords allow users not only to state that a topic applies to their papers or area of expertise, but to indicate how much exactly it applies. They are usually selected from drop down menus.

Most of the conference management systems do not reveal how exactly topics are used in calculating similarity factors. Those that do, do not exploit them in the best possible way.

EasyChair [26] does not use any complex similarity measure but just counts the number of keywords in common. If a paper has more than one common topic with the PC member, it will be regarded as if he chooses "I want to review this paper". If a paper has exactly one common topic with the PC member, it will be regarded as "I can review it" [26]. This allows just 3 different levels of similarity – "entirely dissimilar" (0 topics in common); "slightly similar" (1 topic in common); "very similar" (2+ topics in common).



- □ Data modeling
- ☑ Data types & Data structures
- □ Data mining
- □ Algorithms & Problem solving
- □ Automata & state machines
- ☑ Artificial intelligence
- □ Computer graphics
- ☑ Computer vision

Fig. 3. Example of a list of non-weighted keywords

In contrast to bidding, describing objects by list of keywords provides an independent and stand-alone description of every single paper and reviewer that in most cases reduces the “random assignments” problem. There is no theoretical guarantee however that calculated similarity factors are 100% accurate as they are not directly assigned in accordance with subjective human judgments, but are more or less distorted by mathematics trying to represent the subjective human perception.

Combination of both topics and bidding aims to bring the advantages of the two previously discussed methods together and to reduce the influence of their disadvantages. The best sequence of actions is: 1) Based on the selected topics the conference management system suggests to each reviewer a small subset of let say 20 papers which should be the most interesting to him / her; 2) Reviewers rate the suggested papers; and 3) If a paper has been rated by a reviewer then his/her rating completely determines the similarity factor between them. Otherwise, if the paper has not been rated by the specified reviewer, the similarity factor is calculated based on the selected topics.

Although most of the CMSs support both topics selection and bidding not many of them follow this sequence of actions.

Based on topics matching the MyReview System [34] selects a subset of papers to be explicitly rated by each reviewer. Then missing bids are either predicted by the IRM or implicitly determined by topics matching (papers are described just by a single topic while competences by multiple topics). If the paper's topic matches one of the reviewer's topics the bid is set to “Interested”. If there is no topic in common the bid is set to neutral (“Why not”) [15, 34]. The latter seems not to be a very relevant decision as the “why not” bid still indicates minor competences in the paper's subject domain that could not be justified as there is no explicit bid and no common topic.

OpenConf [32] facilitates automatic assignment based on topics matching or bidding (pro version only) but not based on both simultaneously.

ConfTool [23], Confious [22], CyberChair [24, 17] and Microsoft CMT [31] state they are using both topics and bidding but there is no information how exactly they calculate the similarity factors between papers and reviewers.

EasyChair, as previously mentioned, do not use any complex similarity measure, but just counts the number of topics in common allowing only three levels of similarity.

CyberChair allows reviewers not only to state which topics they are competent in, but also to indicate how much exactly they are competent in. This “topics weighting” feature is supposed to increase the assignment accuracy.

GRAPE [2] system’s fundamental assumption is to prefer topics matching approach over the reviewers’ bidding one, based on the idea that they give assignments more reliability [2]. It uses reviewers’ preferences just to tune the assignments.

2.2.3 A Brief Analysis of the Explicit Methods of Describing Used by the Existing Conference Management Systems

The existing explicit methods of describing papers and competences are most effective when they are used together. However if used separately or not in accordance with the already suggested sequence of actions, the accuracy of the assignment could not be guaranteed because of the following major reasons:

1. Bidding is the most reliable way of determining if a reviewer is competent to evaluate a specific paper. However in case of a large number of submitted papers, reviewers never bid on all of them. If a paper has not been explicitly rated by enough reviewers then *they are assigned to it at random*.
2. The non-structured (unordered) set of conference topics provides an independent stand-alone description of all papers and reviewers, however its size should be limited to a reasonable number of semantically non-overlapping elements. If there are overlapping topics then the author and the reviewer may select different ones although they have the same subject domain in mind. If that happens the similarity factor between them will be calculated as zero, i.e. *reviewers will be again assigned to the paper at random*.
3. The limit in the list’s size results in less detailed keywords or in incomplete coverage of the conference topics. If keywords are too general they can not describe objects in details. If keywords are specific (detailed) enough their number may not be sufficient to cover the entire conference’s area of science leading again to *assigning reviewers at random*.
4. Simple matching based similarity measures or the three-level similarity used in EasyChair do not allow similarity factors to be calculated precisely due to the *low number of distinguished levels*.

Solving random assignment requires solving the following:

1. Similarity factors should be calculated by complex similarity measures that allow factors to change smoothly within the range of [0, 1] on many distinguished levels.
2. A new method should be suggested that allows the conference to be described with many more *semantically-related* keywords. The larger number of keywords will provide more detailed and precise description of papers and competences, and will significantly decrease the probability of authors or reviewers not being able to find relevant keywords for them. Semantic relationship between separate keywords will allow accurate non-zero similarity factors to be calculated between a paper and a reviewer even if they do not share any keyword in common.
3. Existing similarity measures should be adapted or new ones should be proposed that not just count the number of matching (common) keywords, but calculates semantic similarity of the non-matching keywords as well.
4. Evaluation criteria should be proposed for assessing the accuracy of the suggested method of describing papers and reviewers' competences.

Solving task 1 is easy. As the list of conference topics is an unordered set the most reasonable way of calculating similarity factors is by using Jaccard's index (1) or Dice's similarity measure (2).

$$SF(p_i, r_j) = w(e_{p_i r_j}) = \frac{|KWp_i \cap KWr_j|}{|KWp_i \cup KWr_j|} \quad (1)$$

where:

$SF(p_i,r_j)$ - similarity factor between *i*-th paper and *j*-th reviewer

KWp_i - set of keywords, describing the *i*-th paper

KWr_j - set of keywords chosen by the *j*-th reviewer

$$SF(p_i, r_j) = w(e_{p_i r_j}) = \frac{2 \times |KWp_i \cap KWr_j|}{|KWp_i| + |KWr_j|} \quad (2)$$

Tasks 2 and 3 are solved by the suggested, in section 3, method of describing conferences, papers and reviewers' competences by taxonomy of keywords.

3. Describing Papers and Reviewers' Competences by Taxonomy of Keywords

The advantage of the proposed method comes from the hierarchical structure itself. It provides an important additional information – the semantic relationship between the separate topics (keywords) that allows similarity measures to take into account not only the number of keywords in common between a paper and a reviewer, but *in case of non-matching keywords to calculate how semantically close they are*. Thus if a paper and a reviewer do not share even a single topic a non-zero similarity factor could still be accurately calculated.

Before the paper submission and reviewer registration the conference chairs define taxonomy (figure 4) of topics that best describe the conference coverage area. The particular taxonomy on figure 5 is derived from the ACM classification scheme [19] and can be used to describe a software-related conference.

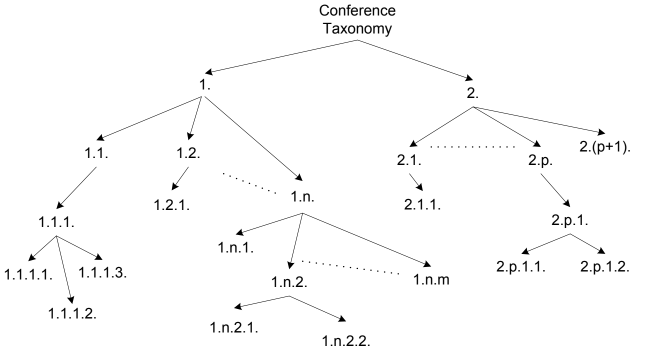


Fig. 4. General taxonomy structure

During paper submission authors are required to select all keywords that apply to their papers. During registration, reviewers select all topics that describe their areas of expertise.

Keywords describing the *i*-th paper and the *j*-th reviewer are stored into unordered sets noted as KWp_i and KWr_j respectively.

KWp_i = {Relational databases, Content analysis, Web-based services, Architectures}

KWr_j = {Relational databases, Distributed databases, Spatial DB & GIS, Information storage and retrieval, Content analysis, Data sharing, Software Engineering, Programming languages, C++}

Although keywords, chosen by authors and reviewers, are stored into unordered sets the semantic relationship between their elements is still available within the conference taxonomy. The most precise way of computing similarity factors is by using a measure of similarity between two sets of concepts in a common taxonomy.

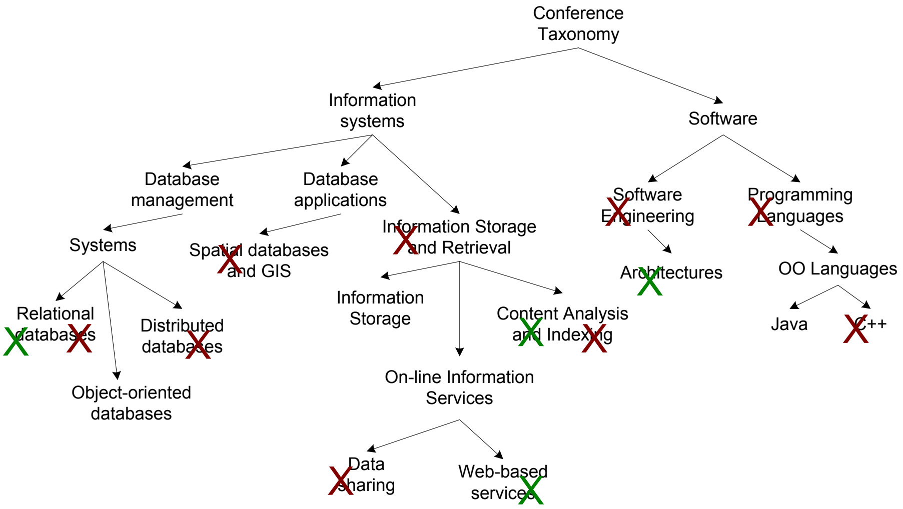


Fig. 5. An example for conference taxonomy with nodes chosen by the author of the i-th paper (coloured in green) and nodes chosen by the j-th reviewer (coloured in light brown)

There are many ways to compute similarity between two unordered sets (Dice, Jaccard and etc.), and many measures calculating the similarity between two individual concepts in a common taxonomy (Wu & Palmer [18], Lin [9], Resnik [14] and etc.). However there are just a few proposals for measuring similarity between sets of concepts in a common taxonomy and they are all asymmetric (Maedche & Staab [10], Haase et al [4]) or take into account all combinations between the elements of the two sets (Rada et al [13], Bouquet et al [1]) which is not applicable to the current subject domain.

The rest of this section proposes a way to compute similarity between two sets of concepts by combining similarity measures between sets with those measuring similarity between individual taxonomy concepts.

The original Dice's similarity measure is applicable to unordered sets only. Applying it to taxonomies will immediately convert the taxonomies to unordered sets as it measures commonality just by the number of exactly matching elements and ignores the semantic relationship between them. If a paper and a reviewer have no topics in common the similarity factor between them calculated by using (1) or (2) will be 0, although the reviewer could be competent to evaluate the paper. Consider for example a paper is described by "Object-oriented programming languages" and the reviewer has chosen "Java". In this case it is obvious the reviewer is competent to review the paper but as the topics do not exactly match the calculated similarity is zero.

To overcome that we can substitute the intersect in the numerator of (2) with the right-hand side expression in (3) that does not just count the number of exactly matching keywords, but calculates the *semantic commonality* between the keywords describing the specified paper and reviewer, so that if they do not share even a single common topic the numerator will not be 0 in case the reviewer is competent to evaluate the paper.

$$| KWp_i \cap KWr_j | \Rightarrow \sum_{k_m \in KWp_i} \max_{k_n \in KWr_j} (Sim(k_m, k_n)) \tag{3}$$

where:

k_m – the *m*-th keyword chosen by the author of the *i*-th paper.

k_n – the *n*-th keyword chosen by the *j*-th reviewer.

KWp_i – set of keywords, describing the *i*-th paper.

KWr_j – set of keywords chosen by the *j*-th reviewer.

$Sim(k_m, k_n)$ – semantic similarity between the *m*-th keyword chosen by the author of the *i*-th paper and *n*-th keyword chosen by the *j*-th reviewer.

$\max\limits_{n \in KWr_j} (Sim(k_m, k_n))$ – semantic similarity between the *m*-th keyword chosen by the author and its closest neighbour (in the conference taxonomy) chosen by the reviewer.

$Sim(k_m, k_n)$ is calculated in a way that $Sim(k_m, k_n) \in [0,1]$. Thus the semantic similarity between the *m*-th keyword chosen by the author and its semantically closest neighbour chosen by the reviewer will lie within the same interval of $[0,1]$.

If both the author of the *i*-th paper and the *j*-th reviewer have selected the same keyword, i.e. $m \equiv n$ then $Sim(k_m, k_n) = 1$ and $\max\limits_{n \in KWr_j} (Sim(k_m, k_n)) = 1$

Thus

$$\sum_{k_m \in KWp_i} \max_{k_n \in KWr_j} (Sim(k_m, k_n)) = | KWp_i \cap KWr_j | \tag{4}$$

in case $Sim(k_m, k_n) = 1$ for $\forall m \in KWp_i$ and $\forall n \in KWr_j$, i.e. if the author and the reviewer have chosen the same keywords.

Furthermore if the taxonomy is converted to an unordered set (by ignoring the semantic relationship between its elements) then (4) is always true – unconditionally, because the similarity between k_m and k_n could be just 1 or 0. This proves substitution (3) is relevant and does not change the spirit of the Dice's measure (2).

The intersection between two unordered sets is a commutative operation however the semantic commonality between two sets of concepts within a common taxonomy is not, thus:

$$\sum_{k_m \in KWp_i} \max_{k_n \in KWr_j} (Sim(k_m, k_n)) \neq \sum_{k_n \in KWr_j} \max_{k_m \in KWp_i} (Sim(k_n, k_m)) \tag{5}$$

The reason (5) to be true is that the different number of elements within the two sets matters as each element has a non-zero similarity with an element from the other set.

In case of taxonomy, applying the suggested substitution (3) to the Dice's measure (2) results in (6). Note that the numerator is not twice the left-hand side of (5) but both sums are taken into account.

$$SF_{(p_i, r_j)} = \frac{\sum\limits_{k_m \in KWp_i} \max\limits_{k_n \in KWr_j} (Sim(k_m, k_n)) + \sum\limits_{k_n \in KWr_j} \max\limits_{k_m \in KWp_i} (Sim(k_n, k_m))}{| KWp_i | + | KWr_j |} \quad (6)$$

where all notations are previously explained.

As proven earlier if the taxonomy is converted into an unordered set, by ignoring the semantic relationship between its elements, then (6) produces the very same result as the Dice's measure (2).

Equation (6), in contrast to the original Dice's measure, could be also applied between sets whose elements are semantically related to each other. Its asymmetric version is (7).

$$SF_{(p_i, r_j)} = \frac{\sum\limits_{k_m \in KWp_i} \max\limits_{k_n \in KWr_j} (Sim(k_m, k_n))}{| KWp_i |} \quad (7)$$

where all notations are previously explained.

The symmetric similarity measure (6) takes into account:

a) how competent a reviewer is to evaluate a specified paper;

b) how focused the reviewer is on the paper's subject domain;

c) how much the paper's topics cover the reviewer's interests.

b) and c) show how worthy is to assign (spend) the reviewer for that paper exactly or it is better to keep him for other papers.

In contrast the asymmetric similarity measure (7) takes only into account how competent the reviewer is to evaluate the specified paper. So, which one to use – (6) or (7)? They both have their significance and which one applies better depends mainly on the assignment algorithm.

If reviewers are assigned to papers by a greedy algorithm then it is mandatory to adopt a symmetric measure. Greedy algorithms process papers in tern, assigning the most competent available reviewers to the current paper, but ignoring the possibility that another paper may appear later whose only suitable reviewers are already busy, i.e. overloaded and no more papers could be assigned to them. As a result the latter paper may remain without reviewers although there are competent ones to evaluate it but they have been previously and unreasonably used. In this case papers being processed first usually get the most competent, in their areas, reviewers however many reviewers may be assigned at random to the papers processed at last. The symmetric similarity measure (6) takes into account not just how competent the reviewer is, but also how worthy is to spend him for that paper exactly. In this way it significantly reduces the amount of random assignments to last processed papers.

If the assignment is handled as a global optimization problem then an asymmetric similarity measure could be used as well since the algorithm will find an optimal solution where there are competent reviewers for all papers.

To calculate the similarity factor between any paper and any reviewer by using (6) or (7), the software needs to know the semantic similarity between any two nodes in the taxonomy – the *m*-th keyword chosen by the author of the *i*-th paper and the *n*-th keyword chosen by the *j*-th reviewer, i.e. $Sim(k_m, k_n)$. There are two general ways of calculating semantic similarity between taxonomy nodes:

- by using the structural characteristics of the taxonomy – distance, depth, density and etc.
- based on the information content of the nodes.

A commonly accepted similarity measure based on structural characteristics is the one formulated by Zhibiao Wu and Martha Palmer [18] (8).

$$Sim_{Wu\ \&\ Palmer}(k_m, k_n) = \frac{2 \times N_0}{2 \times N_0 + N_1 + N_2} \tag{8}$$

where:

N_0 - the distance (in number of edges) between the root and the closest common ancestor *C₀* of the two nodes (C_m) and (C_n) whose similarity is being calculated (Figure 6).

N_1 - the distance from C_0 to one of the nodes, say C_m. C_m represents the *m*-th keyword describing the paper.

N_2 - the distance from C_0 to the other node – C_n. C_n is the node representing the *n*-th keyword chosen by the reviewer.

As Wu and Palmer's measure is a symmetrical one it does not matter if C_m is the node chosen by the author and C_n the one chosen by the reviewer or the opposite. Closer look at (8) reveals it is in fact the Dice's coefficient applied over the definition of the term "path", i.e. a set of edges – twice the number of common edges divided by the number of all edges (including duplicates).

The calculated semantic similarity between any two nodes is only useful if it complies with the real human perception of similarity between these nodes. Dekang Lin [9] evaluated the Wu & Palmer's similarity measure by correlating the calculated similarity factors on 28 pairs of concepts in WordNet [35] with the assessments made by real humans during the Miller and Charles survey

[14]. The experiment shows very good correlation (0.803) with the Miller and Charles results.

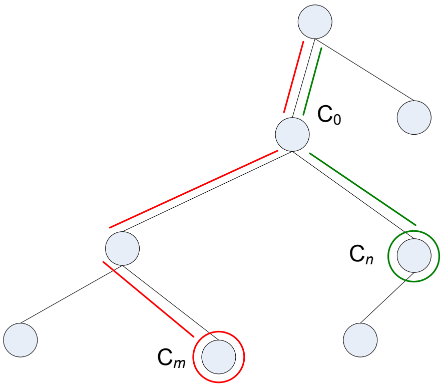


Fig. 6. Visual representation of Wu and Palmer's similarity measure.

Dekang Lin himself suggested another similarity measure based on the information content of the taxonomy nodes [9]. The taxonomy is augmented with a function *p:C* -> [0, 1], such that for any node (concept) *c* ∈ *C*, *p(c)* is the probability of encountering the concept *c* or an instance of it. So if c_1 IS-A c_2, then $p(c_1) <= p(c_2)$. The probability of the root node (if any) is 1. Lower probability results in higher information content. Obviously and expectedly nodes deeper in the hierarchy are more informative than the ones on the top.

Lin's similarity measure [9] (9) looks similar to the one of Wu and Palmer, but takes into account the nodes' information content instead of distances.

$$Sim_{Lin}(k_m, k_n) = \frac{2 \times \log P(C_0)}{\log P(C_l) + \log P(C_m)} \tag{9}$$

where:

$P(C_0)$ - the probability of encountering (in the conference taxonomy) the closest common ancestor C_0 (of the two nodes whose similarity is being calculated) or an instance of it;

$P(C_m)$ - the probability of encountering the node representing the *m*-th keyword chosen by the author or an instance of it;

$P(C_n)$ - the probability of encountering the node representing the *n*-th keyword chosen by the reviewer or an instance of it.

Lin's measure was part of the same experiment and evaluation that Lin performed on the Wu and Palmer's measure as well. It proved Lin's measure has a bit better correlation (0.834) with the real human assessments taken from the Miller and Charles data.

In case of user-weighted keywords the semantic similarity calculated by using equations (8) or (9) could be additionally modified in the form of (10) to take the levels of competence and applicability into an account as well. These levels are usually explicitly provided by the users – authors and reviewers indicate how much exactly each one of the chosen keywords applies to their papers, respectively competences. The main idea behind this modification is to lower the semantic similarity in case the reviewer is less competent in respect to a specified keyword than the level it applies to the paper.

$$\left| \begin{array}{l} Sim_{weighted}(k_m, k_n) = Sim(k_m, k_n) \times C_{levels} \\ \text{If } w_{rj}(k_n) >= w_{pi}(k_m), C_{levels} = 1 \\ \text{else } C_{levels} = 1 - (w_{pi}(k_m) - w_{rj}(k_n)) \end{array} \right. \qquad (10)$$

where:

$w_{rj}(k_n)$ - the expertise of reviewer r_j on the topic k_n. $w_{rj}(k_n) \in$ [0, 1].

$w_{pi}(k_m)$ - the level that k_m applies to paper p_i. $w_{pi}(k_m) \in$ [0, 1].

If both the author and the reviewer have selected one and the same keyword, i.e. $m \equiv n$, then:

- If the level of expertise (on that particular keyword) of the reviewer is higher than or equal to the level chosen by the author then the reviewer is considered to be highly competent to evaluate the specified paper *in respect to that particular keyword only*.
- If the level of expertise of the reviewer is less than the level chosen by the author, then the reviewer is less competent to evaluate the paper in respect to that keyword.

Logic is the same even if the author and the reviewer have selected different, but semantically close topics, i.e. $m \neq n$. If the modified semantic similarity $Sim_{weighted}(k_m, k_n)$ (10) is used in (6) or (7) instead of just $Sim(k_m, k_n)$ then it is possible for a semantically closest neighbour of the *m*-th keyword to be chosen not the nearest node selected by the *j*-th reviewer but another one for which the *j*-th reviewer has specified a higher level of expertise.

The system (10) takes into account the *relative level of competence*, i.e. whether the reviewer is more competent in *n* than *m* applies to the paper and if so the semantic similarity between *m* and *n* remains the same, otherwise it is lowered proportionally to the difference between the levels of *m* and *n*. Alternatively the *absolute level of competence* of the reviewer could be used instead. In that case not the difference between the two levels matters but just the level the reviewer is competent in *n*. Thus (10) is transformed to (11).

$$Sim_{weighted}(k_m, k_n) = Sim(k_m, k_n) \times w_{rj}(k_n) \qquad (11)$$

where all notations are previously explained.

4. Experimental Evaluation

A similarity factor should be calculated between any paper and any reviewer. The number of reviewers is usually proportional to the number of submitted papers so that reviewers’ workload is kept in a reasonable range. Thus in respect to the number of submitted papers and registered reviewers the proposed method of describing the conference has a quadratic computational complexity, $O(n^2)$, where n – the number of submitted papers. The method of describing by an unordered set of conference topics has the same complexity as well, but significantly lower constant factor. The one suggested here reasonably runs slower due to the necessity of calculating semantic similarity between every keyword, *m*, describing the *i*-th paper and every one, *n*, chosen by the *j*-th reviewer. Fortunately the number of calculations could be significantly reduced by using optimization techniques like dynamic programming.

Evaluating accuracy of the methods of describing papers and reviewers’ competences is a challenging task as they involve many subjective aspects like self-evaluation, self-classification, consistency of the taxonomy structure and etc.

The most reasonable way of assessing a method’s accuracy is by comparing the calculated similarity factors to the similarity evaluations provided by real PC members over the same dataset. A reliable way of collecting reviewers’ preferences is the method of bidding [7]. If the conference utilizes both the suggested method and bidding the similarity factors, calculated by the above equations could be compared to the bids provided by the reviewers. Five bid options are commonly available to reviewers to indicate their preferences and willing to review specific papers or not. These are:

1. I’m an expert in the paper’s subject domain and I want to review it;
2. I’m an expert in the paper’s subject domain;
3. I’m not an expert in the paper’s subject domain, but I can review it;
4. The paper’s subject domain is completely out of my expertise;
5. Conflict of interests.

Alternatively the calculated similarity factors could be compared to the reviewers’ self-evaluation of expertise in respect to the papers they evaluate. The self evaluation of expertise is provided by each reviewer during the review submission, i.e. after reading the entire paper. It allows the conference management system to detect also errors of the following types “Selection of too general nodes”, “Incomplete / inappropriate description of papers”, “Ambiguous taxonomy structure” and others.

The proposed method of describing papers and reviewers’ competences has been used by the international broad-area, multidisciplinary, ICT-related conference *CompSysTech* [29] in 2010 and 2011. During review submission every reviewer was required to specify his/her level of expertise in the paper’s subject domain by selecting one of the following: ***High***, ***Medium*** or ***Low*** level of expertise. Then the calculated similarity factors are compared to the reviewers' self-evaluations.

The method is considered to be accurate if most of the similarity factors are accurately calculated, i.e. comply with one of the following rules:

A similarity factor associated with ***low*** level of competences is correctly calculated if, for the paper it applies, it ***is less than*** all correctly calculated factors related to ***medium*** and ***high*** level of competences.

A similarity factor associated with ***medium*** level of competences is correctly calculated if it is ***non-zero*** and, for the paper it applies, it is ***higher than*** all correctly calculated similarities corresponding to ***low*** level of competences and ***less than*** all correctly calculated similarities corresponding to ***high*** level of competences.

A similarity factor associated with ***high*** level of competences is correctly calculated if it is ***non-zero*** and, for the paper it applies, it is ***higher than*** all correctly calculated similarities related to ***low*** and ***medium*** levels of competences.

In other words, in respect to a particular paper*,* the reviewer who stated to be an expert should have higher similarity factor than the one who is not an expert but still capable of reviewing the paper, and the latter should get higher similarity factor than the reviewer who is completely out of the paper's subject domain.

The relationship between the similarity factors on a local scale (per paper) is what really matters, not the specific values and not even the relationship on a global scale (for all papers simultaneously). This statement is true because the most competent reviewer to evaluate a particular paper is the one who has the highest similarity factor with the paper and no other papers can influence that similarity. Furthermore the actual value of the similarity factors depends on the number of chosen keywords while there is no rule stating that each paper or reviewer should be described by precisely the same number of topics.

The assignment algorithm used by the CompSysTech's conference management system is a heuristic but not greedy. It tries to find a global solution (although it does not guarantee finding the best possible one) and it does not assign reviewers to any paper until it finds suitable reviewers for all papers. That allows both symmetric and asymmetric similarity measures to be used. Since the reviewers' self-evaluation of expertise takes into account just the level of competence (which is an asymmetric in nature) it is expected that the relative amount of correctly calculated similarity factors will be similar for both the symmetric and the asymmetric measures. Results for CompSysTech'11 are summarized in table 1.

Table 1. Accuracy evaluation of the similarity factors of CompSysTech 2011

	CST'11 (asymmetric measure)	**CST'11** (symmetric measure)
Number of assignments	525	525
Correctly calculated similarity factors	**83.24 %**	**83.05 %**

Although the relationship between similarity factors on a global scale is not suitable for evaluation of the method's accuracy it could be very useful for identifying the reasons for having inaccurately calculated similarities. Figure 7 shows the distribution of the similarity factors of the three groups (Low, Medium and High) within the interval of [0, 1]. Line chart is used instead of bar chart as lines better illustrate trends. Furthermore in case of multiple data series on the same chart lines look much clearer than thick bars.

Despite the high extent of overlapping, similarity factors of the different groups tend to cluster where "High" have their maximum within the subinterval [0.8, 0.9), followed by "Medium" in [0.7, 0.8) and then "Low" in [0.6, 0.7).

Two "anomalies" are easily noticeable from the graphic:

- Relatively high amount of similarity factors associated with *High* level of competences within the lower subinterval of [0.3, 0.5];
- Relatively high mode value, 0.6-0.7, of the similarity factors associated with *Low* level of competences.

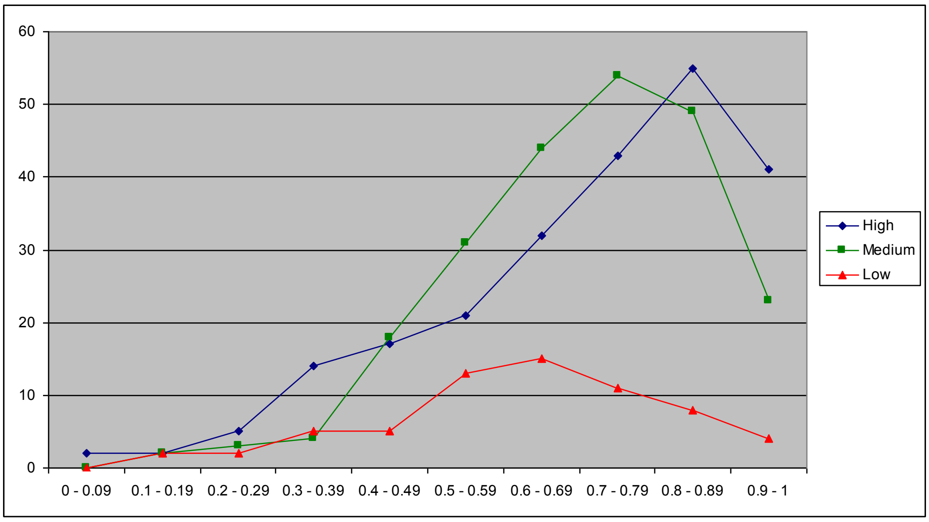


Fig. 7. Distribution of similarity factors (calculated by an asymmetric measure) of the three groups (Low, Medium and High) within the interval of [0, 1] for CompSysTech'11.

A detailed review of the inaccurately calculated similarity factors reveals the following major reasons for having inaccuracies:

- ***Incomplete description of competences*** - reviewers choose keywords just within a narrow area of science, but during the budding phase they bid as experts on papers outside that area.
- ***Incomplete description of papers*** – multi-disciplinary papers described by keywords from just one area. For example: e-learning on hardware subjects, but described by e-learning keywords only.
- ***Selection of two general, low-informative nodes***. For example: Software or Computer Applications. The lack of details does not allow papers and competences to be described precisely.
- ***Multi-disciplinary papers with higher contribution in a non-primary area*** (non-computing area in our context) which the reviewer could not evaluate without having a sufficient knowledge in that area. For example, with computing as a primary area, papers related to yoghurt processing, internal combustion engines, medicine, finance and etc.
- ***Ambiguous taxonomy structure.*** For example duplicate keywords located within different nodes in the tree.

Incorrectly calculated similarity factors not due to the specified reasons are marked as “Unclassified”. Their nature is more or less random and could be a result of other, unforeseen, subjective factors or inappropriate taxonomy structure. It is preferable taxonomy to be narrow and deep rather than broad and shallow. Figures 8 and 9 show the contribution of the individual reasons to the presence of *High* competences associated with low similarity factors (fig. 8) and *Low* competences associated with high similarity factors (fig. 9).

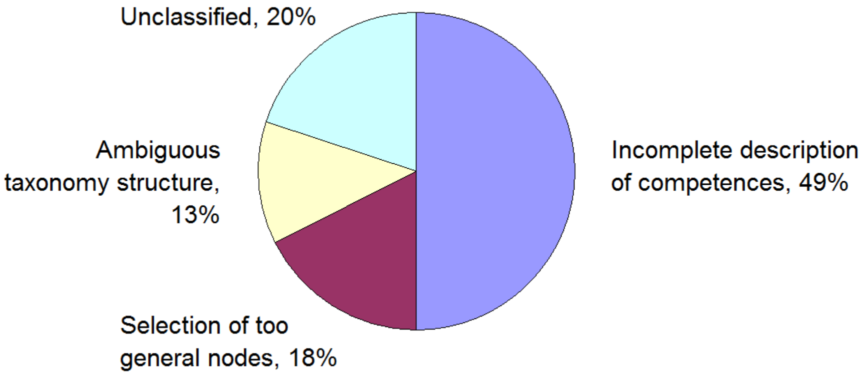


Fig. 8. Contribution of the individual reasons to the presence of low similarity factors (< 0.5) associated with *High* level of competences.

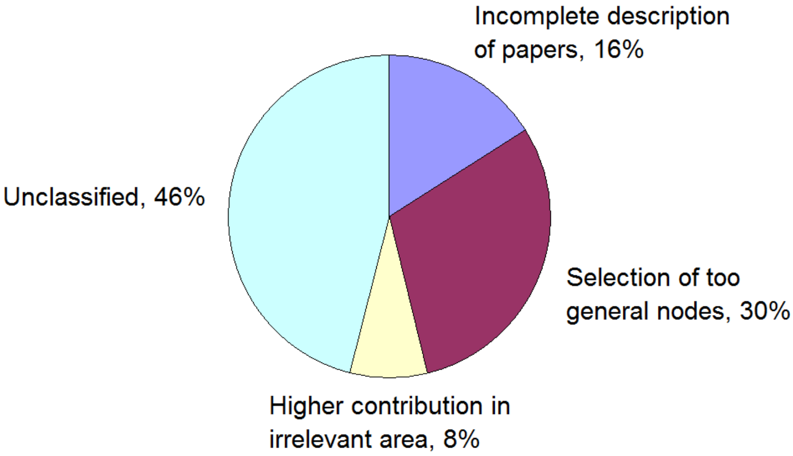


Fig. 9. Contribution of the individual reasons to the presence of high similarity factors (> 0.5) associated with Low level of competences.

Since the major causes of inaccuracy are identified they may be partially or completely avoided and/or compensated. Refer to section “Conclusions and Future Improvements” for more information.

Comparing the suggested method to the one describing papers and competences by an unordered set (often referred as list) of keywords

Comparing the accuracy of two or more explicit methods is a challenging task due to the subjective aspects of self-classification. *The fairest comparison* requires that different methods are implemented in a single conference management system so *they are used in parallel, in one and the same event, and by the same authors and PC members – a situation that is more or less unrealistic*.

Fortunately alternative ways of indirect comparison are also possible. For example CompSysTech 2009 was described by an unordered set of keywords but the assignment was done by the same assignment algorithm, configured in the very same way and the Programme Committee in 2009 was the same as in 2010. In 2011 PC was larger but contained all members from 2009 as well. Then besides the papers the only other difference is the method of describing papers and competences. Everything else (the assignment algorithm and the PC members) is the same. That allows us to indirectly compare the two methods.

Table 2 shows the relative amount of correctly calculated similarity factors (according to the accuracy rules stated earlier) for CompSysTech 2009. Expectedly the asymmetric measure shows a little bit higher result due to the asymmetric natures of the reviewers' self-evaluation of expertise. The asymmetric measure is calculated as the number of common keywords divided by the number of keywords describing the paper, while the symmetric measure is the Jaccard's index (1).

Table 2. Accuracy evaluation of the similarity factors of CompSysTech 2009

	CST'09 (asymmetric measure)	**CST'09** (symmetric measure)
Number of assignments	356	356
Number of random assignments (zero calculated similarities)	64	64
Number of zero-calculated similarity factors associated with High and Medium level of competences	57	57
Correctly calculated similarity factors	**67.70 %**	**67.13 %**

As discussed in section 2 the method of describing papers and competences by unordered set of keywords is characterized by a large number of random assignments, i.e. zero-calculated similarity factors. As seen on the table most of them are actually related to medium and high level of competences. Although the zero-calculated similarities, reviewers are assigned to the corresponding papers just because they (reviewers) have directly chosen the papers they would like to evaluate. If reviewers were unable to do that, papers should have been assigned to them at random, significantly decreasing the overall number of assignments associated with high and medium level of expertise.

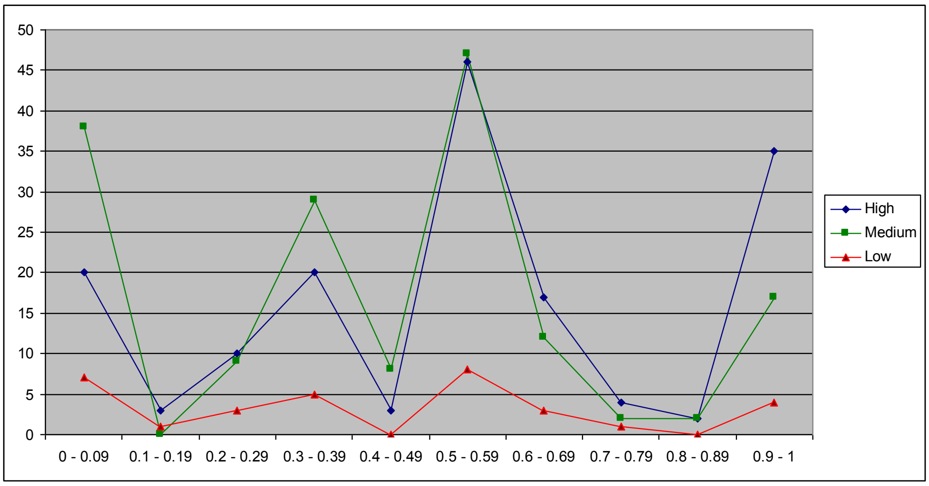


Fig. 10. Distribution of similarity factors (calculated by an asymmetric measure) of the three groups (Low, Medium and High) for CompSysTech'09.

The global scale distribution of similarity factors, calculated by the asymmetric similarity measure, for CompSysTech 2009 is shown on figure 10.

In contrast to the bell-shaped like distribution produced by the suggested method, the asymmetric similarity measure of the one using unordered set of keywords provides reduced set of values together with more chaotic distribution. As seen on the figure all the three groups (Low, Medium and High) have their maximums in the same subinterval of [0.5, 0.6). However after 0.6 similarity factors associated with *High* level of competences are more than the ones corresponding to *Medium*, i.e. the unordered set of keywords could be also used to describe papers and competences although it provides less precision and accuracy, in comparison to the taxonomy, especially for broad area and/or multi-disciplinary conferences.

One may say the accuracy levels in 2011 and 2009 are just lucky - accidentally achieved. Table 3 contradicts that statement by showing accuracy levels for all years from 2008 to 2011. For both years when the conference management system used taxonomy for describing papers and competences the accuracy levels are higher than the two years when the system relied on an unordered set of keywords. In 2008 papers were evaluated by just two reviewers that improperly make the overall accuracy higher (the fewer number of similarity factors per paper decreases the possibility that they are wrong since one of the similarities is taken as a reference value, i.e. considered to be correct) but it is still under 70%.

Table 3. Accuracy evaluation of the similarity factors of CompSysTech 2008 to 2011 (asymmetric similarity measures)

	CST'11	**CST'10**	**CST'09**	**CST'08**
Method of description	taxonomy	taxonomy	unordered set	unordered set
Reviewers per paper	3	3	3	2
Number of assignments	525	385	356	196
Number of random assignments (zero similarities)	0	0	64	35
Correctly calculated similarity factors	**83.24 %**	**80.26 %**	**67.70 %**	**68.88 %**

5. Conclusions and Future Improvements

The suggested method relies on taxonomy of keywords (conference topics) used as descriptive metadata. The implied hierarchical structure allows similarity measures to calculate similarity between a paper and a reviewer even if they do not share any common keyword at all. That avoids random assignments and increases the weights' accuracy in general, resulting in more precise automatic assignment.

The method however does not fully eliminate the inaccurately calculated similarity factors. Since most of the reasons for having inaccuracies are identified during the analysis they could be completely or partially avoided and/or compensated as follows:

1. To avoid selection of too general, low informative nodes (ex. Computer Applications; Software and etc.) it is enough to technically disallow users to select these nodes (or all nodes within the first one or two levels after the root).

2. Partial description of competences could be significantly compensated by using *collaborative filtering* techniques. If a reviewer bids as an expert on papers described by keywords he/she has not chosen, then they could be automatically added to his/her set of keywords. Additional constraints may be applied here to avoid improper propagation of keywords due to incorrect selection by authors - for example keywords could be added to the reviewer's set only if: they are chosen by other reviewers who bid as experts on the same paper (proving the author's selection of keywords is correct); or if the keywords also describe other papers the reviewer has bid on as an expert (proving the reviewer has competences within the specified area but missed to mark the relevant keywords).

3. In some cases reviewers tend to generalize their competences and instead of selecting all nodes within a specific narrow area of science reviewers select just their (the nodes) common ancestor stating high level of competence on it. For example: If the reviewer feels highly competent on all aspects of information systems then he/she is more likely to choose the general node "Information Systems" rather than its sub-nodes. Achieving higher accuracy however requires users to select nodes deeper in the hierarchy. The following rule helps increasing the accuracy while providing reviewers' comfort by allowing them to generalize their competences by selecting just a single node.

If

1. A reviewer has stated *high level of competence* for a node n_i
and 2. n_i *is between* […] and […] level in the hierarchy
and 3. n_i *has children*
and 4. the reviewer ***has not*** *selected any* of n_i's children.

then:

Add n_i's direct successors to the set of keywords chosen by the reviewer (assuming the reviewer is generalizing his knowledge and instead of selecting all children he has selected just their common ancestor for convenience)

Although the proposed method achieves higher accuracy in comparison to other methods used in the existing conference management systems its accuracy could be further increased by implementing the three proposals described above. Additionally the method should be verified with specialized narrow-area conferences where its accuracy is expected to be even higher.

A comparable solution based on paper-to-paper similarity could be used to group papers in working sessions.

The suggested, in this article, method and all accompanying formulas have been exclusively designed to precisely calculate similarity between papers and reviewers however they may be used to describe any other objects and subsequently to compute similarities between them – for example men and women on dating sites; or job vacancies and candidates and etc.

Acknowledgement. This paper is supported by project: Creative Development Support of Doctoral Students, Post-Doctoral and Young Researches in the Field of Computer Science, BG 051PO001-3.3.04/13, European Social Fund 2007–2013, Operational Programme "Human Resources Development"

References

1. Bouquet, P., G. Kuper, M. Scoz, S. Zanobini. Asking and answering semantic queries. Proc. of Meaning Coordination and Negotiation Workshop (MCNW-04) in conjunction with International Semantic Web Conference (ISWC-04), 2004, pp. 25-36.
2. Di Mauro N., T. M.A. Basile, S. Ferilli. GRAPE: An Expert Review Assignment Component for Scientific Conference Management Systems. Proceedings of the 18th international conference on Innovations in Applied Artificial Intelligence. Italy, 2005, pp. 789 – 798.
3. Ferilli S., N. Di Mauro, T.M.A. Basile, F. Esposito, M. Biba. Automatic Topics Identification for Reviewer Assignment. 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2006. Springer LNCS, 2006, pp. 721-730.
4. Haase, P., R. Siebes, F. van Harmelen. Peer Selection in Peer-to-Peer Networks with Semantic Topologies. Semantics of a Networked World, Lecture Notes in Computer Science, vol. 3226, Springer, 2004, pp. 108-125.
5. Kalmukov, Y. Analysis and experimental study of an algorithm for automatic assignment of reviewers to papers. ACM ICPS Vol. 285, ACM Press N.Y., 2007, ISBN: 978-954-9641-50-9.
6. Kalmukov, Y. An algorithm for automatic assignment of reviewers to papers. Proceedings of the international conference on computer systems and technologies CompSysTech'06. Avangard Print, 2006.
7. Kalmukov, Y., B. Rachev. Comparative Analysis of Existing Methods and Algorithms for Automatic Assignment of Reviewers to Papers. *Journal of Information Technologies and Control* 2010:(2), ISSN 1312-2622, pp. 20-31.
8. Kuhn, Harold W. "The Hungarian Method for the assignment problem". *Naval Research Logistics Quarterly*. 1955:(2), pp. 83–97.
9. Lin, D. An Information-Theoretic Definition of Similarity. Proceedings of the Fifteenth International Conference on Machine Learning ICML '98, ISBN:1-55860-556-8, 1998.
10. Maedche, A., S. Staab. Measuring Similarity between Ontologies. Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, Lecture Notes in Computer Science, vol. 2473, Springer, 2002, pp. 15-21.

11. Munkres, J. "Algorithms for the Assignment and Transportation Problems". *Journal of the Society for Industrial and Applied Mathematics*. 5:1, 1957, pp. 32–38.
12. Pesenhofer A., R. Mayer, A. Rauber. Improving Scientific Conferences by enhancing Conference Management System with information mining capabilities. Proceedings IEEE International Conference on Digital Information Management (ICDIM 2006), ISBN: 1-4244-0682-x; S. 359 - 366.
13. Rada, R., H. Mili, E. Bicknell, M. Blettner. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, Jan/Feb 1989, pp. 17 - 30, ISSN: 0018-9472.
14. Resnik, Ph., Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, Vol. 11 (1999), pp. 95-130.
15. Rigaux Ph. An Iterative Rating Method: Application to Web-based Conference Management. Proceedings of the 2004 ACM Symposium on Applied Computing (SAC'04), ACM Press N.Y., pp. 1682 – 1687,
ISBN 1-58113-812-1.
16. Rodriguez M., J. Bollen. An Algorithm to Determine Peer-Reviewers. Conference on Information and Knowledge Management (CIKM 2008), ACM Press, pp. 319-328.
17. van de Stadt, R., CyberChair: A Web-Based Groupware Application to Facilitate the Paper Reviewing Process. 2001, Available at www.cyberchair.org.
18. Wu, Z., M. Palmer. Verb semantics and lexical selection. 32nd. Annual Meeting of the Association for Computational Linguistics (1994), pp. 133 -138.
19. ACM Computing Classification System - http://www.acm.org/about/class/
20. ACM Digital Library - http://portal.acm.org/dl.cfm
21. Ceur Workshop Proceedings “Ceur WS” – http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/
22. Confious Conference Management System - http://confioussite.ics.forth.gr/
23. ConfTool Conference Management Tool - http://conftool.net/
24. CyberChair Conference Management System - http://www.borbala.com/cyberchair/
25. DBLP Computer Science Bibliography - http://www.informatik.uni-trier.de/~ley/db/
26. EasyChair Conference Management System - http://www.easychair.org/
27. Editorial Manager – Online Manuscript Submission and Peer Review. http://www.editorialmanager.com/homepage/home.htm
28. Google Scholar - http://scholar.google.com/
29. International Conference on Computer Systems and Technologies CompSysTech, http://www.compsystech.org/
30. Latent Semantic Indexing - http://en.wikipedia.org/wiki/Latent_semantic_indexing
31. Microsoft Conference Management Toolkit - http://cmt.research.microsoft.com/cmt/
32. OpenConf Conference Management System - http://www.openconf.com/
33. Scientific Literature Digital Library and Search Engine CiteSeer - http://citeseerx.ist.psu.edu
34. The MyReview System - http://myreview.lri.fr/
35. WordNet – A Lexical Database for English, http://wordnet.princeton.edu/

Yordan Kalmukov received his BSc and MSc degrees in Computer Technologies at the University of Ruse in 2005 and 2006 respectively. Currently he is a PhD student at the Department of Computer Systems and Technologies, University of Ruse and teaches BSc students in web programming. His research interests include information retrieval, automated document management and classification systems, web-based information systems and databases.

Received: September 06, 2011; Accepted: March 30, 2012.

DOI: 10.2298/CSIS110801009I

Journal evaluation based on bibliometric indicators and the CERIF data model

Dragan Ivanović[1], Dušan Surla[2], and Miloš Racković[2]

[1] University of Novi Sad, Faculty of Technical Sciences, Trg D. Obradovića 6,
21000 Novi Sad, Serbia
chenejac@uns.ac.rs
[2] University of Novi Sad, Faculty of Sciences, Trg D. Obradovića 4,
21000 Novi Sad, Serbia
surla@uns.ac.rs, rackovic@dmi.uns.ac.rs

Abstract. In this paper we propose an application of extended CERIF data model for storing journal impact factors and journal scientific fields and also propose a journal evaluation approach based on these data. The approach includes an algorithm for journal evaluation based on one metric for journals ranking that is also stored using the CERIF data model and that is in accordance with the rule book for evaluation of scientific-research results which is prescribed by the Republic of Serbia. The algorithm does not unambiguously evaluate journal, i.e. the algorithm suggests possibly journal categories according to the values of the metric, but final decision is made by commission. The proposed evaluation approach is implemented within CRIS UNS and verified on scientific-research results of researchers employed at Department of Mathematics and Informatics, University of Novi Sad. The complete evaluation approach proposed in this paper is based on the CERIF standard that allows an easy application of this evaluation approach in any CERIF-compatible CRIS system.

Keywords: CERIF; evaluation of scientific-research results; impact factor; CRIS UNS.

1. Introduction

Main output of scientific-research activity is scientific-research result: journal paper, monograph, etc. Evaluation of scientific-research results is important for research policy of state, region or institution. There are three approaches for evaluation of scientific-research results: creation of an experts group (a commission) that evaluates scientific-research results according to some rules, usage of bibliometric indicators (impact factor, h-Index, citation, etc.) and combination of previous two approaches, i.e., creation of commission which takes into account some bibliometric indicators for results evaluation.

Many research management systems have been developed in recent years. The central part of those systems is input of metadata about scientific-research results. Those metadata can be used for evaluation of scientific-research results of researchers in accordance with national, regional and international rule books. In order to do that, it is necessary to classify the results into appropriate categories defined by some rule book. Seljak and Bošnjak [1] present usage of SICRIS system for evaluation of researcher's results. Yi and Kang [2] present development of the evaluation system of the Electronics and Telecommunications Research Institute in Korea.

Standardization of data model used in those systems is necessary in order to enable researchers to find the desired data in various research management systems as well as to enable efficient data exchange between those systems. CERIF (*Common European Research Information Format* – http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1) is a standard that defines data model of research management systems. A non-profit organisation euroCRIS (http://www.eurocris.org/) has been responsible for the development of CERIF since 2002. The European Union encourages the development of national research management systems in accordance with CERIF standard because the European Union wants to achieve maximum competitiveness of Europe at all levels of research activity. A research management system compatible with the CERIF data model is called CRIS (*Current Research Information System*).

The process of evaluation of scientific research is prescribed by law in the Republic of Serbia, i.e., there is the rule book which was issued by the Ministry of Education and Science of the Republic of Serbia. The Ministry uses the previously mentioned rule book for funding scientific projects and faculties and scientific institutions use the rule book for selection of teaching and scientific position. For those purposes, researchers have to evaluate their own scientific results according to the rule book and to submit their evaluation reports to competent institutions. After that, those evaluation reports are controlled by institutions' boards. The previously described evaluation approach, regardless to great effort made for controlling evaluation reports, still does not provide a consistent application of the rule book. There are some cases when same publication has been differently evaluated in a few final evaluation reports. In order to overcome those shortcomings a software system for evaluation of scientific results called CRIS UNS has been developed at University of Novi Sad since 2008. The first phase of CRIS UNS development covered implementation of a module for entering metadata about scientific-research results. The second phase of that system development is related to evaluation of those results. The main idea of using CRIS UNS for results evaluation is that researchers enter data about their scientific-research results by themselves and that some commission (group of experts) evaluates those results. The system is going to generate evaluation reports using metadata about scientific-research results entered by researchers and results evaluations entered by some commission. This approach does not require that researchers have to know the rule book for the evaluation and to apply it to their results, i.e., CRIS UNS improves procedure of evaluation of scientific-research results. The current rule book (which is published in 2008) prescribes that evaluation of a paper published in a journal depends on journal scientific field and journal value (category of journal), which means it is sufficient to evaluate journals (it is not necessary individual evaluation of all papers published in journals). The current rule book introduces three categories of journals based on value of the journal impact factor and scientific field (see Table 2). Therefore, it is necessary to enable storage of journals impact factors and scientific fields in the CRIS UNS data model. In that way, the CRIS UNS system could help commission members to evaluate journals. Due to the needs for increasing availability of scientific-research results, the CRIS UNS system is built on the CERIF data model that enables data exchange with other CERIF-compatible research management systems. Therefore, storage of the journals impact factors, scientific fields and the rule book in the CERIF compatible data model is necessary.

In this paper we propose an application of the CERIF model extension for storing previously mentioned data. This proposal includes algorithm for storing journals impact factors and scientific fields that does not require addition of new types of entities to existing data model, i.e., only instantiations of existing types of CERIF entities are required. This paper also proposes a journal evaluation method based on the journal impact factor (JIF) and journal scientific fields stored in the data model in previously described way. That method includes algorithm for journal evaluation based on one metric for journals ranking which takes into account the JIF and journal scientific fields. The metric for journals ranking is in accordance with the rule book prescribed by the Republic of Serbia and it is also stored in the CERIF data model. The proposed algorithm for evaluation does not unambiguously evaluate journal, i.e. the algorithm suggests possibly journal categories according to the values of the metric, but final decision is made by commission. The proposed evaluation approach is implemented within CRIS UNS and verified on scientific-research results of researchers employed at Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad. Furthermore, the proposed approach allows simultaneously using different metrics for journals ranking that can be based on various bibliometric indicators. The complete evaluation approach proposed in this paper is based on the CERIF standard that allows an easy application of this evaluation approach in any CERIF-compatible CRIS system.

2. Background and related work

This section shows fundamental concepts and related work for the paper research area. Those concepts are basis for investigation presented in this paper. The section is divided into three subsections: Current Research Information Systems, Journal impact factor and Journal Citation Report and Rule books for evaluation of scientific-research results.

There are systems that evaluate scientific-research results. Seljak and Bošnjak [1] present evaluation of researcher's results stored in SICRIS system that is based on CERIF-compatible model, but this evaluation does not use bibliometrics indicators for evaluation unlike the evaluation approach presented in this paper that is based on bibliometrics indicators stored in the CERIF *cfMetrics* entity. Yi and Kang [2] present development of the evaluation system of the Electronics and Telecommunications Research Institute in Korea, but this system is not CERIF-compatible. The evaluation approach presented in this paper evaluates results stored in current research information systems that are interoperable with all CERIF-compatible systems. This interoperability increases accessibility of scientific-research results, i.e., enhances the further development of science.

2.1. Current Research Information Systems

Standardization of data model used in research management systems is necessary in order to enable data exchange between those systems. CERIF is a leading standard for data exchange between those systems and research management systems compatible with the CERIF data model are called CRIS (*Current Research Information Systems*).

CERIF (*the Common European Research Information Format*) defines a data model that contains information about people, projects, organizations, publications, patents, equipments, etc. The first version of CERIF was published in 1991 and the current version is the CERIF2008 v1.2 published in 2010. The CERIF data model has six groups of entities:

- Core entities,
- Result entities,
- 2nd level entities,
- Link entities,
- Semantic layer entities and
- Multilingual entities.

Entities *Project*, *Person* and *OrganizationUnit* are the CERIF *Core entities*. Entities *ResultPublication*, *ResultPatent* and *ResultProduct* belong to the CERIF *Result entities* and contain metadata about scientific-research results. Furthermore, the group *2nd level entities* contains entities dedicated to hold relevant data for scientific-research activity that do not belong to groups *Result entitites* and *Core entities*. The group contains the following entities: *cfCite, cfCountry, cfCurrency, cfCV, cfEAddr, cfEquip, cfEvent, cfExpSkills, cfFacil, cfFundProg, cfLang, cfMetrics, cfPAddr, cfPrize, cfQual, cfSrv*. Those entities are linked with the CERIF *Core entities* and *Result entities* through the CERIF *Link entities*. The *Link entities* hold references to the two related entities, time period in which relation between entities applied (attributes *startDate* and *endDate*) as well as the classification of relationship. The CERIF data model has the CERIF *Semantic layer entities* that enable classification of entities and relations between entities in accordance with some classification scheme. For example, classification of relations between the *Person* entity and the *ResultPublication* entity is: author of publication, editor of publication, lecturer. Examples of classification of the *ResultPublication* entity are: thesis, dissertation, paper published in journal, monograph, etc. In addition to the list of entities that belong to the *Semantic Layer entities*, CERIF prescribes values that those entities can hold (provides a list of possible classifications). The CERIF data model provides storage for certain data, such as title, abstract, keywords, research area in multiple languages. Those data are stored using the CERIF *Multilingual entities*.

Due to specific local or national requirements CRIS systems are built on different modifications of the CERIF data model. *Asserson, Jeffery and Lopatenko* [3] describe CERIF data model extensions that were created in order to satisfy requirements of a CRIS that was developed at the University of Bergen and a CRIS that was developed by “Science and Technology Facilities Council”. Moreover, a CERIF data model extension that was created in order to satisfy requirements of the IST World portal is described in papers [4;5;6;7]. Furthermore, papers [8;9] describe a CERIF data model extension that uses the formalized Dublin Core for description of scientific-research results. *Ivanović, Surla and Konjović* [10] propose a CERIF-compatible data model based on the MARC 21 format (http://www.loc.gov/marc/). In that model, part of the CERIF data model related to scientific-research results is replaced with data model of the MARC 21 format. MARC 21 is a standard that prescribes a format for bibliographic data storing.

By 2011, there are many CRIS systems: IST World (http://www.ist-world.org/), HunCRIS (http://nkr.info.omikk.bme.hu/HunCRIS_eng.htm), SICRIS (http://sicris.izum.si/default.aspx?lang=eng), CRIStin (http://www.cristin.no/), Pure (http://www.atira.dk/en/pure/), CRIS UNS (http://cris.uns.ac.rs/), etc. IST World (*Information Society Technology World*) is a portal that provides access to scientific-research results from several countries. This portal was developed within a FP6 (*Sixth Framework Programme*) project. The data model created for the purpose of that system is a CERIF data model extension. HunCRIS and SICRIS are Hungarian and Slovenian national CRIS systems based on the CERIF data model. CRUStin is an information system used by scientific institutions in Norway. Pure is commercial software that can be installed and customized for the needs of scientific institutions. That software system is used by many universities such as the University of Helsinki and the University of Copenhagen.

CRIS UNS is a CERIF compatible research management system that has been being developed since 2008 at the University of Novi Sad in the Republic of Serbia. Experience gained from developing of the BISIS library information system (http://www.bisis.uns.ac.rs/) is used for developing CRIS UNS. The BISIS system has been being developed since 1993 at the University of Novi Sad. The current version 4 is based on XML technology. Within the version, an XML editor for cataloguing in the UNIMARC and MARC21 format [11;12] is developed. The first phase of CRIS UNS development covered implementation of a system for entering metadata about scientific-research results in the following forms: papers published in journals, papers published in conferences proceedings, monographs, papers published in monographs. The next phase of the system development is related to evaluation of those results. The system is built on the CERIF-compatible data model based on the MARC21 format presented in the paper [9]. The system implementation is described in the papers [13;14], and a module for automatic extraction of metadata from scientific papers in PDF format for CRIS UNS is described in the paper [15]. Scientific-research results from the system are available to anonymous user via Internet. The system is in accordance with CERIF and meets requirements prescribed by Ministry of Education and Science of the Republic of Serbia in the field of scientific-research results evaluation [16]. Moreover, the system is implemented as web application that enables authors to input metadata about their own research results without the knowledge of CERIF and MARC21. The system data model and architecture enable easy integration of the system with library information systems.

2.2. Journal impact factor and Journal Citation Report

One of bibliometric indicators is the journal impact factor (JIF). The JIF measures the importance of a journal. The JIF is a measure reflecting the average number of citations to articles published in science journals. *Glänzel and Moed* [17] discuss about advantages and disadvantages of using that indicator. Papers [18;19] explain the history and the proper uses of the JIF. *Bensman* [20] and *Bensman, Smolinsky and Pudovkin* [21] investigate the journals impact factors distribution. *Bordons, Fernández and Gómez* [22] discuss problems of using the JIF as measure in non-English speaking countries. *Van Leeuwen and Moed* [23] also discuss some shortcomings of the JIF and propose a new measure called *Journal to Field Impact Score* (JFIS). *Buela-Casal* [24] proposes to take into account the reputation of the publication from which the citation emanates. Moreover, *Bollen, Rodriguez and van de Sompel* [25] propose the measure of citations based on popularity (number of citations) and prestige (expert appreciation). *Sombatsompop, Markpin and Premkamolnetr* [26] propose the modification to the JIF from the point of citation period (half-life of the journal). *Frandsen and Rousseau* [27] generalize the definition of the JIF to allow different publications and citation periods. *van Leeuwen and Moed* [28] study the relationship between the JIF and other indicators. *Egghe* [29] compares the evaluation based on h-Index with evaluation based on the JIF. There are also other metrics for journals' scientific prestige such as Eigenfactor [30] and SCImago Journal Rank [31].,It is obvious that the JIF has weaknesses that researchers have been tried to eliminate by proposing some modifications of the JIF. Despite its weaknesses and existence of other metrics for journal ranking, journal impact factor is widely spread bibliometric indicator for evaluation of journals. Also, the JIF can be start point for measuring other aspects relevant for scientific-research activity. *Chinchilla-Rodriguez and colleagues* [32] propose an approach based on the JIF that visualize international scientific collaboration.

In a given year, the JIF is the average number of particular year citations received per paper published in that journal during some preceding period. If period of previous two years is used then it is two-year JIF, and if period of previous five years is used then it is five-year IF. Hereinafter, term JIF refers to two-year JIF (it is a common practice).

For instance, 131 papers in 2007 and 133 papers in 2008 were published in the journal *Scientometrics*. In other words, 264 papers were published in that journal in two-year period 2007-2008. Those papers were cited 572 times in all the papers published in 2009. The JIF for 2009 of *Scientometrics* is quotient of numbers 572 and 264, i.e., the JIF for 2009 is 2.167.

Journals impact factors are published in Journal Citation Report (JCR) every June for the previous year. The calculations are performed based on the situation in all three citations databases (SCIe(xpanded), SSCI, AHCI) on the first day of March.

JCR has two partly overlapped sections: *JCR Science Edition* and *JCR Social Science Edition*. Each journal can belong into several of 220 scientific fields and can have one annual impact factor. A journal position by the value of impact factor can be determined within each scientific field to which a journal belongs. **Table 1** presents the two-year impact factors of *Scientometrics* for the period 2001-2009. In the third and fourth row are positions of the journal within scientific fields. For example, within the scientific field *Computer Science, Interdisciplinary Applications* were 76 journals in 2001 and *Scientometrics* was on 25th place in a sorted list of journals. The list is sorted in decreasing order by value of JIF.

Table 1. Impact factors

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009
Two-year JIF	0.676	0.855	1.251	1.120	1.738	1.363	1.472	2.328	2.167
Scientific field: Computer Science, Interdisciplinary Applications	25/76	19/80	16/83	24/83	18/83	24/87	24/92	13/94	18/95
Scientific field: Information Science & Library Science	16/55	16/55	9/55	14/54	5/55	12/53	12/56	8/61	10/65

2.3. Rule books for evaluation of scientific-research results

Definition of different rule books for evaluation and quantitative expression of scientific-research results is possible. Those rule books contain types of results that can be evaluated (the elements of evaluation) and quantitative measures assigned to them. Types of results can be decomposed to several hierarchical levels. For example, the first level can be journal of international importance, journal of national importance, etc. The second level can be the type of publication in the journal: research article, review, etc. In general, it is possible to continue decomposition to arbitrary number of hierarchical levels (three, four, etc.). Different quantitative measures can be joined to same type results within different groups of sciences.

The paper [16] proposes an extension of the CERIF data model for storage of rule books using the following entities:

- The *RuleBook* entity contains name and description of a rule book, time period in which a rule book applies.
- The *ResultType* entity presents element of evaluation. Set of instances of the *ResultType* entity are connected to an instance of the *RuleBook* entity and a recursive relation defined for the *ResultType* entity enables that types of results can be decomposed to arbitrary number of hierarchical levels.
- The *SciencesGroup* entity represents model of group of sciences. This entity can be replaced with the CERIF entity *cfClass*.
- The *ResultTypeMeasure* entity represents model of quantitative measure of result type. This entity defines how many points researcher gets for result which belongs to some result type (relation between the *ResultTypeMeasure* entity and the *ResultType* entity) and belongs to some sciences group (relation between the *ResultTypeMeasure* entity and the *SciencesGroup* entity).

The extension proposed in the paper [16] is included in the CRIS UNS data model and that extension is used for description of the rule book for evaluation of scientific-research results prescribed by the Republic of Serbia. The rule book evaluates result depending on result type and group of sciences to which result belongs. Types of results are the followings: papers published in journals, papers published in conferences proceedings, monographs, papers published in monographs, etc. Three categories of journals depending on the value of JIF are introduced (**Table 2**).

Table 2. Journal categories

Journal category	Definition
Leading journal of international importance	Ranked among the top 30% in the JCR list of journals.
Outstanding journal of international importance	Ranked between the top 30% and 50% in the JCR list of journals.
journal of international importance	Exists on the JCR list of journals, but is not ranked among the top 50%.

Types and quantifications of results related to journals of international importance are presented in **Table 3**. There are three sciences groups: (1) *Mathematics and Natural Sciences*, (2) *Technical and Technological Sciences*, (3) *Social Sciences*. Within those sciences groups, types of results are quantified with numerical value, e.g., type of result *Paper published in journal of international importance* within sciences groups (1) and (2) is quantified with 3 points, and within sciences group (3) is quantified with 4 points.

Table 3. Types and quantifications of results related to journals

Result type	Code	Sciences group		
		(1)	(2)	(3)
Paper published in leading journal of international importance	M21	8	8	8
Paper published in outstanding journal of international importance	M22	5	5	5
Paper published in journal of international importance	M23	3	3	4
Paper published in journal of international importance verified by special decision	M24	3	3	4
Book review published in outstanding journal of international importance	M25	1,5	1,5	1,5
Book review published in journal of international importance	M26	1	1	1
Editor of outstanding journal of international importance	M27	3	3	3
Editor of journal of international importance	M28	2	2	2

Evaluation of scientific-research results using CRIS UNS requires:

- Input of results metadata,
- Input of a rule book for evaluation,
- Input of a commission categorization of journals, monographs and conferences and
- Generation of evaluation reports.

Evaluation of scientific-research paper published in a journal depending on group of sciences to which a researcher belongs: *Mathematics and Natural Sciences*, *Technical and Technological Sciences*, *Social Sciences*. The paper is evaluated using journal categories suggested by the CRIS UNS system. Calculation of suggested journal categories is based on positions by value of impact factor within each scientific field to which a journal belongs. For example:

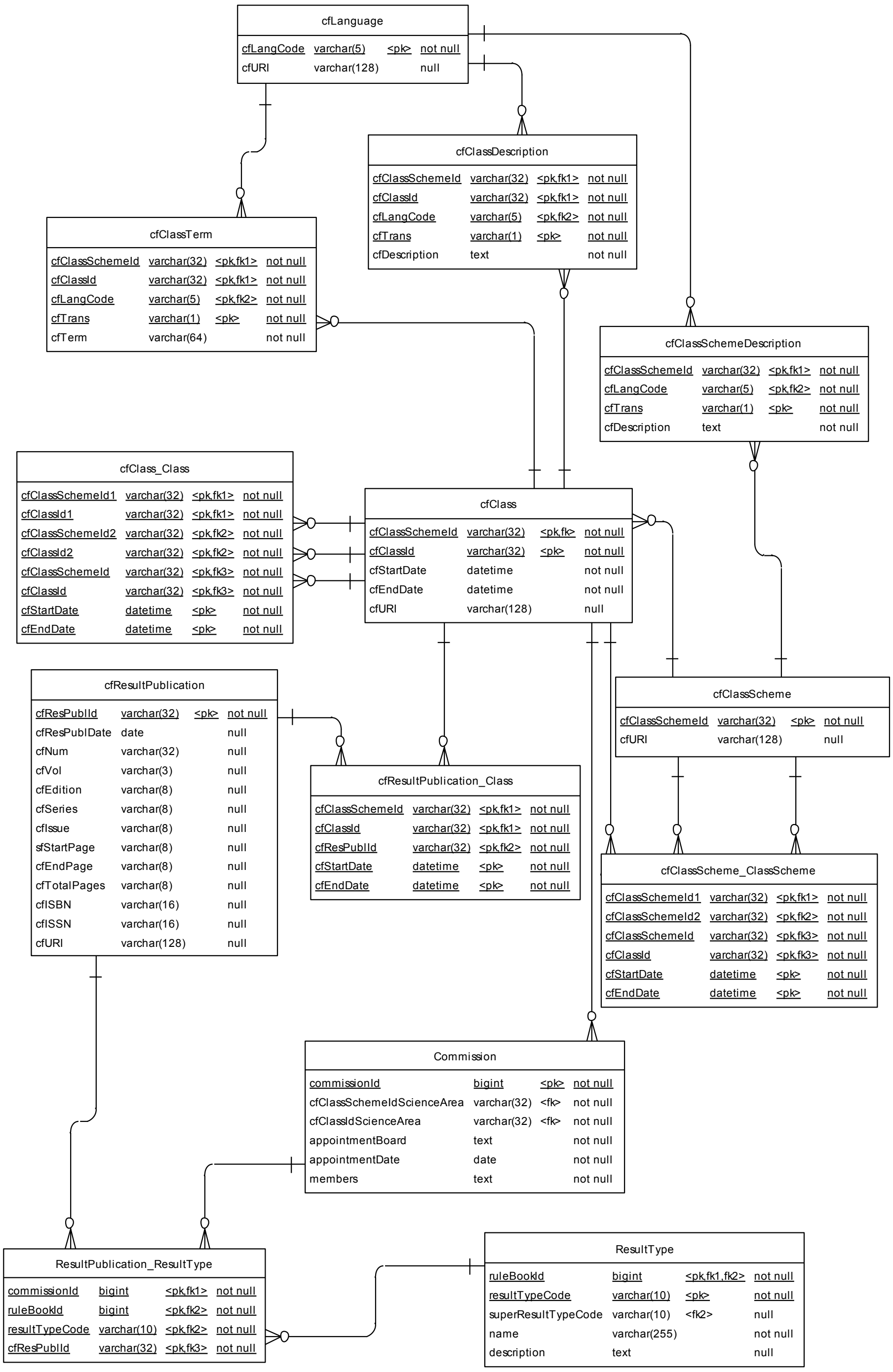
- The journal belongs to the scientific fields Computer Science, Interdisciplinary Applications and Information Science & Library Science.
- Both scientific fields belong to the same sciences group: *Technical and Technological Sciences*.
- The researcher also belongs to the same sciences group. The system suggests journal categories using journal positions by value of impact factor in both scientific fields.
- The commission categorizes the journal (usually adopts one of the suggested categories).

Storage of impact factors and scientific fields of journals in the CRIS UNS data model enables CRIS UNS can suggest categorization of a journal that has an impact factor. A commission adopts or rejects suggested categorization. Also, the commission on the basis of their scientific competence evaluates papers published in scientific journals that do not have impact factors.

3. Findings

Evaluation and quantitative expression of scientific-research results using the rule book prescribed by the Republic of Serbia depends on results scientific fields. Result published in a journal can be evaluated on the basis of journal position by the value of impact factor within scientific field to which the journal belongs. Data about a journal IF, scientific fields and positions within scientific fields are necessary for previously described evaluation. On the basis of those data, some metric can be assigned to the journal. Those data storage in the CERIF data model is proposed in this section. The storage does not require addition of new types of entities as well as attributes to existing CERIF data model entities, i.e., only instantiations of existing types of entities are required.

Journals scientific fields can be stored using the CERIF group *Semantic layer entities* (**Figure 1**). That group contains the entities *cfClass* and *cfClassSheme* that describe classes and classification schemes, respectively. Relation between a class and a classification scheme can be established using the *cfClassSchemeId* attribute of the *cfClass* entity. A classification scheme is additionally described using the *cfClassSchemeDescription* entity that belongs to the CERIF group *Multilingual entities.* Furthermore, a class is additionally described using the entities *cfClassTerm* and *cfClassDescription* that also belong to the group *Multilingual entities*. A relation between two classes can be established using the *cfClass_Class* entity that belongs to the group *Link entities*. The attributes *cfClassSchemeId1* and *cfClassId2* hold foreign key values to the first, and the attributes *cfClassSchemeId2* and *cfClassId2* hold foreign key values to the second linked class. The attributes *cfClassSchemeId* and *cfClassId* classify the relation, i.e., those attributes hold foreign key values to a class that presents type of the relation. A relation between two classification schemes can be established using the *cfClassScheme_ClassScheme* entity that belongs to the group *Link entities*. Other entities of the CERIF data model are linked with semantic layer through the *cfClass* entity, e.g., the *cfResultPublication_Class* entity (**Figure 1**) links the entities *cfResultPublication* and *cfClass*. The *cfResultPublication_Class* entity defines results scientific fields (publication can present some scientific-research result) and the *cfResultPublication_ResultType* entity defines types of results by some rule book. The *cfCommission* entity is dedicated to hold basic information about commissions.



cfLanguage			
cfLangCode	varchar(5)	<pk>	not null
cfURI	varchar(128)		null

cfClassDescription			
cfClassSchemeId	varchar(32)	<pk,fk1>	not null
cfClassId	varchar(32)	<pk,fk1>	not null
cfLangCode	varchar(5)	<pk,fk2>	not null
cfTrans	varchar(1)	<pk>	not null
cfDescription	text		not null

cfClassTerm			
cfClassSchemeId	varchar(32)	<pk,fk1>	not null
cfClassId	varchar(32)	<pk,fk1>	not null
cfLangCode	varchar(5)	<pk,fk2>	not null
cfTrans	varchar(1)	<pk>	not null
cfTerm	varchar(64)		not null

cfClassSchemeDescription			
cfClassSchemeId	varchar(32)	<pk,fk1>	not null
cfLangCode	varchar(5)	<pk,fk2>	not null
cfTrans	varchar(1)	<pk>	not null
cfDescription	text		not null

cfClass_Class			
cfClassSchemeId1	varchar(32)	<pk,fk1>	not null
cfClassId1	varchar(32)	<pk,fk1>	not null
cfClassSchemeId2	varchar(32)	<pk,fk2>	not null
cfClassId2	varchar(32)	<pk,fk2>	not null
cfClassSchemeId	varchar(32)	<pk,fk3>	not null
cfClassId	varchar(32)	<pk,fk3>	not null
cfStartDate	datetime	<pk>	not null
cfEndDate	datetime	<pk>	not null

cfClass			
cfClassSchemeId	varchar(32)	<pk,fk>	not null
cfClassId	varchar(32)	<pk>	not null
cfStartDate	datetime		not null
cfEndDate	datetime		not null
cfURI	varchar(128)		null

cfResultPublication			
cfResPublId	varchar(32)	<pk>	not null
cfResPublDate	date		null
cfNum	varchar(32)		null
cfVol	varchar(3)		null
cfEdition	varchar(8)		null
cfSeries	varchar(8)		null
cfIssue	varchar(8)		null
sfStartPage	varchar(8)		null
cfEndPage	varchar(8)		null
cfTotalPages	varchar(8)		null
cfISBN	varchar(16)		null
cfISSN	varchar(16)		null
cfURI	varchar(128)		null

cfClassScheme			
cfClassSchemeId	varchar(32)	<pk>	not null
cfURI	varchar(128)		null

cfResultPublication_Class			
cfClassSchemeId	varchar(32)	<pk,fk1>	not null
cfClassId	varchar(32)	<pk,fk1>	not null
cfResPublId	varchar(32)	<pk,fk2>	not null
cfStartDate	datetime	<pk>	not null
cfEndDate	datetime	<pk>	not null

cfClassScheme_ClassScheme			
cfClassSchemeId1	varchar(32)	<pk,fk1>	not null
cfClassSchemeId2	varchar(32)	<pk,fk2>	not null
cfClassSchemeId	varchar(32)	<pk,fk3>	not null
cfClassId	varchar(32)	<pk,fk3>	not null
cfStartDate	datetime	<pk>	not null
cfEndDate	datetime	<pk>	not null

Commission			
commissionId	bigint	<pk>	not null
cfClassSchemeIdScienceArea	varchar(32)	<fk>	not null
cfClassIdScienceArea	varchar(32)	<fk>	not null
appointmentBoard	text		not null
appointmentDate	date		not null
members	text		not null

ResultPublication_ResultType			
commissionId	bigint	<pk,fk1>	not null
ruleBookId	bigint	<pk,fk2>	not null
resultTypeCode	varchar(10)	<pk,fk2>	not null
cfResPublId	varchar(32)	<pk,fk3>	not null

ResultType			
ruleBookId	bigint	<pk,fk1,fk2>	not null
resultTypeCode	varchar(10)	<pk>	not null
superResultTypeCode	varchar(10)	<fk2>	null
name	varchar(255)		not null
description	text		null

Figure 1. Journals and Semantic layer entities

In order to define scientific fields prescribed by JCR, the followings are necessary:

- Creation of an instance of the *cfClassScheme* entity with mnemonic name *scientific field*; and creation of 220 instances of the *cfClass* entity that are linked with the classification scheme *scientific field* and present scientific fields prescribed by JCR. In other words, the attribute *cfClassSchemeId* of those 220 instances holds the primary key value of the *cfClassScheme* entity instance with mnemonic name *scientific field*.
- Creation of an instance of the *cfClassScheme* entity with mnemonic name *section*; and creation of 2 instances of the *cfClass* entity that are linked with the classification scheme *section* and present the sections prescribed by JCR: *JCR Science Edition* and *JCR Social Science Edition*.
- Creation of an instance of the *cfClassScheme* entity with mnemonic name *scientific field – section - relation* as; and creation of an instance of the *cfClass* entity that is linked with the classification scheme *scientific field – section - relation* and has mnemonic name *scientific field belongs to section*.
- Creation of instances of the *cfClass_Class* entity for linking instances of the *cfClass* entity that represent scientific field (the attributes *cfClassSchemeId1* and *cfClassId1* of the *cfClass_Class* entity) and section (the attributes *cfClassSchemeId2* and *cfClassId2* of the *cfClass_Class* entity). Those instances of the *cfClass_Class* entity are classified using instance of the *cfClass* entity that has mnemonic name *scientific field belongs to section* (the attributes *cfClassSchemeId* and *cfClassId* of the *cfClass_Class* entity).

In order to link scientific fields and sciences groups prescribed by some rule book for evaluation, the followings are necessary:

- Creation of an instance of the *cfClassScheme* entity with mnemonic name *sciences group*; and creation of three instances of the *cfClass* entity that are linked with the classification scheme *sciences group* and have the following mnemonic names *Mathematics and Natural Sciences*, *Technical and Technological Sciences*, *Social Sciences*. Those three instances of the *cfClass* entity represent sciences group defined by the rule book for scientific-research results evaluation of the Republic of Serbia. As already mentioned, the CERIF extension for evaluation that is adopted for CRIS UNS replaces the *SciencesGroup* entity with the CERIF entity *cfClass*. Some rule book can have more groups of sciences. Within different groups of sciences same type results can be differently quantified.
- Creation of an instance of the *cfClassScheme* entity with mnemonic name *scientific field – sciences group - relation*; and creation of an instance of the *cfClass* entity that is linked with the

classification scheme *scientific field – sciences group - relation* and has mnemonic name *scientific field belongs to sciences group*.

- Creation of instances of the *cfClass_Class* entity for linking instances of the *cfClass* entity that represent scientific field and section, respectively. Those instances of the *cfClass_Class* entity are classified using instance of the *cfClass* entity that has mnemonic name *scientific field belongs to sciences group.*

In order to link a journal with result types and scientific fields, the followings are necessary:

- Creation of instances of the *cfResultPublication_ResultType* entity for linking an instance of the *cfResultPublication* entity that represents a journal and an instance of the *cfResultType* entity that represents a result type to which the journal belongs by some commission (the *commissionId* attribute of the *cfResultPublication_ResultType* entity).
- Creation of instances of the *cfResultPublication_Class* entity for linking an instance of the *cfResultPublication* entity that represents a journal and an instance of the *cfClass* entity that represents a scientific field to which the journal belongs. The attributes *cfStartDate* and *cfEndDate* define time period in which the journal belongs to some scientific field (journal can change scientific field over time).

The JIF is bibliometric indicator. The CERIF data model has the *cfMetrics* entity that is intended for storage of various metrics and this entity can be used for storage of the JIF. The *cfMetrics* entity and connected entities are shown in **Figure 2**. A metric name and description is stored using the multilingual entities *cfMetricsName* and *cfMetricsDescription*. A journal is stored using the *cfResultPublication* entity. That entity can be also used for storage of other publications, such as monographs, conference proceedings, etc. Those publications could be also linked with appropriate metrics. Multilingual entities linked with the *cfResultPublication* entity are not shown in **Figure 2** (*cfResultPublicationTitle*, *cfResultPublicationKeywords*, etc.). A value of metric in some year for some publication can be defined using the attributes *cfCount*, *cfFraction* and *cfYear* of the *cfResultPublication_Metrics* entity.

In order to store journals impact factors, the followings are necessary:

- Creation of instances of the *cfMetrics* entity with mnemonic names *two-year impact factor* and *five-year impact factor.*
- Creation of an instance of the *cfClassScheme* entity with mnemonic name *journal – impact factor - relation*; and creation of an instance of the *cfClass* entity that is linked with the classification scheme *journal – impact factor - relation* and has mnemonic name *value of impact factor.*
- Creation of instances of the *cfResultPublication_Metrics* entity for linking instances of the *cfResultPublication* entity that represent

journal and instances of the *cfMetrics* entity with mnemonic name *two-year impact factor*. Those instances of the *cfResultPublication_Metrics* entity are classified using the instances of the *cfClass* entity with the mnemonic name *value of impact factor.* There is an instance for each year. Year is stored using the *cfYear* attribute, and value of JIF is stored using the *cfCount* attribute of the *cfResultPublication_Metrics* entity.

- Creation of instances of the *cfResultPublication_Metrics* entity for linking instances of the *cfResultPublication* entity that represent journal and instances of the *cfMetrics* entity with mnemonic name *five-year impact factor*. Those instances of the *cfResultPublication_Metrics* entity are classified by the analogy to the classification described in the previous item.

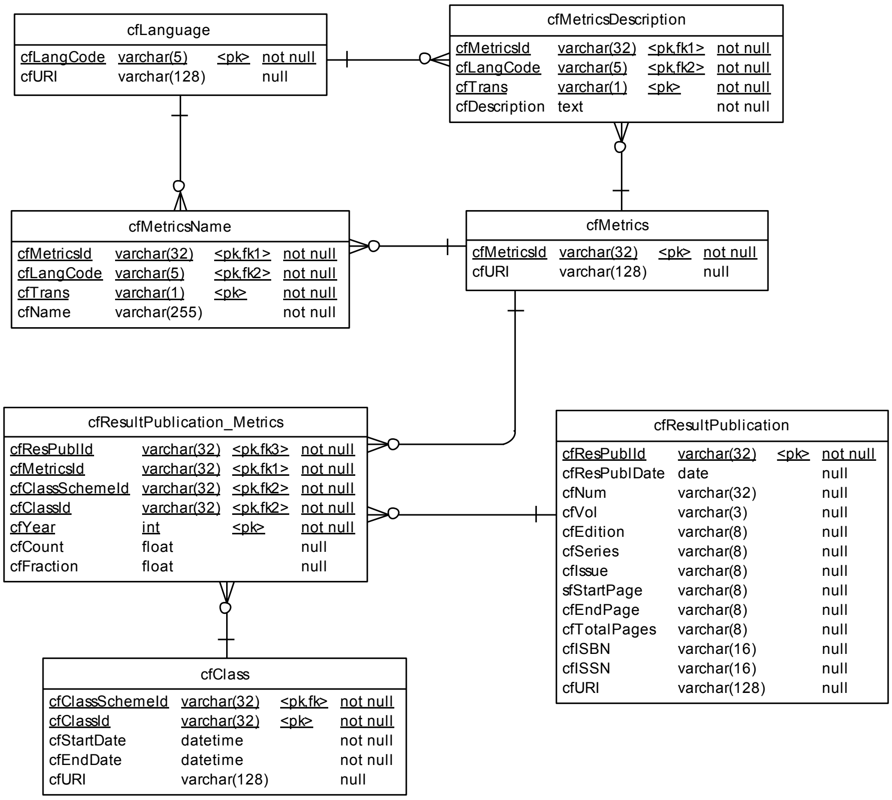

cfLanguage
cfLangCode varchar(5) <pk> not null
cfURI varchar(128) null
cfMetricsDescription
cfMetricsId varchar(32) <pk,fk1> not null
cfLangCode varchar(5) <pk,fk2> not null
cfTrans varchar(1) <pk> not null
cfDescription text not null
cfMetricsName
cfMetricsId varchar(32) <pk,fk1> not null
cfLangCode varchar(5) <pk,fk2> not null
cfTrans varchar(1) <pk> not null
cfName varchar(255) not null
cfMetrics
cfMetricsId varchar(32) <pk> not null
cfURI varchar(128) null
cfResultPublication_Metrics
cfResPublId varchar(32) <pk,fk3> not null
cfMetricsId varchar(32) <pk,fk1> not null
cfClassSchemeId varchar(32) <pk,fk2> not null
cfClassId varchar(32) <pk,fk2> not null
cfYear int <pk> not null
cfCount float null
cfFraction float null
cfResultPublication
cfResPublId varchar(32) <pk> not null
cfResPublDate date null
cfNum varchar(32) null
cfVol varchar(3) null
cfEdition varchar(8) null
cfSeries varchar(8) null
cfIssue varchar(8) null
sfStartPage varchar(8) null
cfEndPage varchar(8) null
cfTotalPages varchar(8) null
cfISBN varchar(16) null
cfISSN varchar(16) null
cfURI varchar(128) null
cfClass
cfClassSchemeId varchar(32) <pk,fk> not null
cfClassId varchar(32) <pk> not null
cfStartDate datetime not null
cfEndDate datetime not null
cfURI varchar(128) null

Figure 2. Metrics

By the rule book for scientific-research results evaluation of the Republic of Serbia, a paper published in a journal is evaluated on the basis of scientific fields and the journal category, which means categorization of the journal is sufficient, i.e., evaluation of each paper is not necessary. Category of journal referred in JCR within each scientific field is determined using a metric for journals ranking which is defined as quotient of *journal position in sorted list in decreasing order by value of JIF* and *total number of journals belonging to scientific field*. In order to store the metric for journals ranking, the following is necessary:

- Creation of instances of the *cfResultPublication_Metrics* entity for linking instances of the *cfResultPublication* entity that represent journal and instances of the *cfMetrics* entity with mnemonic name *two-year impact factor*. Those instances of the *cfResultPublication_Metrics* entity are classified using the instances of the *cfClass* entity that represent scientific fields to which a journal belongs. There is an instance for each year and each scientific field. Year is stored using the *cfYear* attribute, and position in sorted list of journals (in decreasing order by value of JIF) is stored using the *cfCount* attribute of the *cfResultPublication_Metrics* entity. The *cfFraction* attribute of the *cfResultPublication_Metrics* entity holds the metric for journals ranking. The metric for journals ranking holds a value from the interval (0, 1] and the value of metric establishes a category of journal according to the rule book. For example: A journal is a leading journal of international importance (**Table 2**) if value of its metric for ranking belongs to the interval (0, 0.3].

A scientific-research paper authorized by a researcher is evaluated in accordance with sciences group to which the researcher belongs: *Mathematics and Natural Sciences, Technical and Technological Sciences, Social Sciences*. The paper is evaluated on the basis of journal categorization (**Table 2**) within journal scientific fields which belong to the same sciences group like researcher. That journal categorization is done by a commission for evaluation. A CRIS can send queries to a database and provide values of the previously described metric to a commission. In that way, the CRIS provides help to the commission in journals categorization.

An example of such a query is shown in **Listing 1**. The query returns the value of the metric for *„Scientometrics“* for 2001 within the scientific fields that belong to the group of sciences „*Technical and Technological Sciences*“. The query is described using a meta-language that uses mnemonic names instead of the primary keys of entities. **Figures 1** and **2** shown in the previous section are necessary and sufficient for this query understanding. Those figures shows entities used in this query. Instances of the *cfClass_Class* entity link instances of the *cfClass* entity that represent scientific fields (the attributes *cfClassSchemeId1* and *cfClassId1* of the *cfClass_Class* entity) and instances of the *cfClass* entity that represent sciences groups (the attributes *cfClassSchemeId2* and *cfClassId2* of the *cfClass_Class* entity). Those instances of the *cfClass_Class* entity are classified using the instance of the *cfClass* entity with mnemonic name *scientific field belongs to sciences group* (the attributes *cfClassSchemeId* and *cfClassId* of the *cfClass_Class* entity).

```
select   journal_metric.cfFraction
         from     cfResultPublication journal, cfMetrics metric,
                  cfResultPublication_Metrics journal_metric
         join     journal, metric, journal_metric
         where    journal is „Scientometrics“ and
                  metric is „two-year impact factor“ and
                  journal_metric.cfYear is „2001“ and
                  (journal_metric.cfClassSchemeId, journal_metric.cfClassId)  in
                          (select cfClassSchemeId1, cfClassId1
                               from cfClass_Class
                               where   mnemonicName(cfClassSchemeId,   cfClassId)   is
                                               „scientific field belongs to sciences group“ and
                                   mnemonicName(cfClassSchemeId2,   cfClassId2)   is
                                               „Technical and Technological Sciences“)
```

Listing 1. Query

As already shown in **Table 2,** the journal *„Scientometrics“* in 2001 belonged to the scientific fields *„Computer Science, Interdisciplinary Applications“* and *„Information Science & Library Science“*. Both of those scientific fields belong to the group of sciences „*Technical and Technological Sciences*“. Within the first scientific field the metric for journals ranking has value 25/76 = 0.329, and within the second 16/55 = 0.291. Those two values are result of execution of query shown in **Listing 1.** Based on those values the journal can be categorized (**Table 2**) as *Outstanding journal of international importance* (value of the metric lies in interval (0.3, 0.5]) or as *Leading journal of international importance* (value of the metric lies in interval (0, 0.3]). A commission for evaluation has to make decision and categorize the journal.

4. Discussions

Previous section proves that journals impact factors and scientific fields can be stored in the CERIF data model. The CERIF semantic layer and the *cfMetrics* entity are used for that storage. The *cfMetrics* entity can be also used for storing other bibliometric indicators. On the basis of those data, evaluation of journals scientific-research results in accordance with some rule book is possible. A rule book can be stored using the extension of the CERIF data model presented in the paper [16].

The proposed evaluation approach is implemented within CRIS UNS and verified on scientific-research results of researchers employed at Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad. Journals listed in JCR and their impact factors (see **Table 1**) are stored in the system database. The database contains only data related to journals that contain papers authorized by some researcher affiliated at the University of Novi Sad, i.e., all journals referred in JCR are not present in the system database. If the database contained all journals referred in JCR, the data supplied in the last two rows in **Table 1** could be derived. The department founded a commission for scientific-research results evaluation. The commission had to categorize all scientific journals in which at least one researcher employed at the department published some scientific-research result. In order to do that, the journal evaluation approach presented in this paper is used by the commission. As already mentioned, the approach uses a metric for journals ranking based on journals impact factors and journals scientific fields. The first encountered problem was the proposed metrics for different journal scientific fields can give different journal categorizations. The second encountered problem was missing JCR for some years. One possible approach to overcome these problems is evaluation of each journal for each year by the commission. The second approach that is applied in the CRIS system is based on rules for overcoming these problems created by the commission:

1. If a journal for a certain year has at least one of the scientific fields that belong to a sciences group in which the evaluation is done, then the best journal category within scientific fields belonging to the sciences group is adopted, otherwise the worst journal category within different scientific fields is adopted. Precondition for this rule is the journal has the impact factor for certain year.
2. If a journal for a certain year does not have impact factor, then the best journal categorization within a period of previous four years and next two years is adopted.
3. If a journal after applying the rule 1 and the rule 2 still does not have categorization for a certain year, then the journal is categorized as M53 (scientific journal) for the certain year.

Introducing of other rules defined by some other commission by which CRIS UNS automatically categorizes journals is also possible. However, a commission has to categorize journals that do not have JIF for any year. Journals categorizations are stored in data model as it is described in the paper [16].

Using of others metrics for journals categorizations is also possible:

- Value of JIF lies in some interval of values,
- Quartile score (http://researchassessment.fbk.eu/quartile_score),
- Eigenfactor (http://www.eigenfactor.org/),
- SCImago Journal Rank (http://www.scimagojr.com/),
- Position of a journal in a sorted list of journals by decreasing value of JIF within a scientific field (regardless to the total number of journals in the scientific field),
- etc.

Using the other metrics can be easy achieved in a way analogous to the proposed algorithm by creating new instances of the *cfMetrics* entity and new instances of *cfResultPublication_Metrics* entity that hold appropriate values.

The CERIF date model also enables storage of number of publication citations using the *cfCite* entity. This aspect of CERIF has not been considered in this paper because the rule book of the University of Novi Sad does not evaluate papers published in journals based on citations.

5. Conclusions

A proposal for evaluation of journal based on the JIF and journal scientific fields stored in a CERIF model extension is presented in this paper. The *cfMetrics* entity of the CERIF data model is used for storage of impact factors, and the CERIF semantic layer is used for storage of scientific fields. The storage does not require addition of new types of entities as well as attributes to existing CERIF data model entities, i.e., only instantiations of existing types of entities are required. Algorithm for journal evaluation presented in this paper is based on a metric for journals ranking that is in accordance with the rule book for evaluation of scientific-research results which is prescribed by the Republic of Serbia. The algorithm for evaluation does not unambiguously evaluate journal, i.e. the system suggests possibly journal categories which calculation is based on the metric, but final decision is made by commission.

On the basis of the categorized journals evaluation of all the scientific results published in journals can be done. A category assigned to a journal for a certain year can be used for evaluation of all results published in the journal in the certain year, or results can be evaluated using the best journal category for some period of time. The rule book for evaluation prescribed by the Republic of Serbia uses a period of three years (the result publication year and the previous two years).

CRIS UNS is CERIF-compatible research management system of the University of Novi Sad that stores journals impact factors and scientific fields as it is described in this paper and uses those data for evaluation of papers published in journals in accordance with the results evaluation rule book of the Republic of Serbia. The complete evaluation approach proposed in this paper is based on the standard CERIF that allows an easy application of this evaluation approach in any CERIF-compatible CRIS system.

This paper proves it is possible to create a research management system which:

- is interoperable with other systems based on the CERIF data model and
- can store bibliometric indicators about journals that can be used for evaluation of scientific-research results published in journals.

Acknowledgment. This paper is part of the research project "*Infrastructure for Technology Enhanced Learning in Serbia*" supported by the Ministry of Education and Science of the Republic of Serbia [Project No. 47003].

References

1. Seljak, T. and Bošnjak, A.: Researchers' bibliographies in COBISS.SI. Information Services and Use, Vol. 26, No. 4, 303-308. (2006)
2. Yi, C.-G. and Kang, K.-B.: Developments of the evaluation system of government-supported research institutes in Korean science and technology. Research Evaluation, Vol. 9, No. 3, 158-170. (2000)
3. Asserson, A., Jeffery, K. and Lopatenko, A.: CERIF: Past, Present and Future: An Overview. Proceedings of the 6th International Conference on Current Research Information Systems. University of Kassel, August 29 - 31, 2002, 33-40. (2002)
4. Ferlež, J.: Public IST World Deliverable 1.3 – Data Model for Representation of Expertise, 12 p., available at: http://ist-world.dfki.de/downloads/deliverables/ISTWorld_D1.3_DataModelForRepresentationOfExpertise.pdf (2005)
5. Kiryakov, A., Grabczewski, E., Ferlež, J., Uszkoreit, H. and Jörg, B.: Public IST World Deliverable 1.1 – Definition of the Central Data Structure, 23 p., available at: http://ist-world.dfki.de/downloads/deliverables/ISTWorld_D1.1_CentralDataStructure.pdf (2005)
6. Jörg, B., Ferlež, J. and Grabczewski, E.: Public IST World Deliverable 1.2 – Data Model for Knowledge Organisation, 18 p., available at: http://ist-world.dfki.de/downloads/deliverables/ISTWorld_D1.2_DataModelForKnowledgeOrganisation.pdf (2005)
7. Jörg, B., Ferlež, J., Grabczewski, E. and Jermol, M.: IST World: European RTD Information and Service Portal. 8th International Conference on Current Research Inforation Systems: Enabling Interaction and Quality: Beyond the Hanseatic League (CRIS 2006). Bergen, Norway, 10 p., available at: http://epubs.cclrc.ac.uk/bitstream/905/ISTWorld01.pdf (2006)
8. Jeffery, K., Lopatenko, A. and Asserson, A.: Comparative Study of Metadata for Scientific Information: The place of CERIF in CRISs and Scientific Repositories. Proceedings of the 6th International Conference on Current Research Information Systems. University of Kassel, August 29 - 31, 2002, 77-86. (2002)
9. Jeffery, K.: An architecture for grey literature in a R&D context. The International Journal on Grey Literature, Vol. 1, No 2, 64-72. (2000)
10. Ivanović, D., Surla, D. and Konjović, Z.: CERIF compatible data model based on MARC21 format. The Electronic Library, Vol. 29, No. 1, 52-70. (2011)
11. Dimić, B. and Surla, D.: XML Editor for UNIMARC and MARC21 cataloguing. The Electronic Library, Vol. 27, No. 3, 509-528. (2009)
12. Dimić, B., Milosavljević, B. and Surla, D.: XML schema for UNIMARC and MARC 21 formats. The Electronic Library, Vol. 28, No. 2, 245-262. (2010)
13. Ivanović, D., Milosavljević, G., Milosavljević, B. and Surla, D.: A CERIF-compatible research management system based on the MARC21 format. Program: Electronic libarary and information systems, Vol. 44, No. 3, 229-251. (2010)
14. Milosavljević, G, Ivanović, D., Surla, D. and Milosavljević, B.: Automated Construction of the User Interface for a CERIF-Compliant Research Management System. The Electronic Library, Vol. 29, No. 5, pp. 565 – 588, DOI: 10.1108/02640471111177035 (2011).
15. Kovačević, A., Ivanović, D., Milosavljević, B., Konjović, Z. and Surla, D.: Automatic extraction of metadata from scientific publications for CRIS systems. Program: Electronic libarary and information systems, pp.376 – 396, DOI: 10.1108/00330331111182094 (2011).

16. Ivanović, D., Surla, D. and Racković, M.: A CERIF data model extension for evaluation and quantitative expression of scientific research results. Scientometrics, Vol. 86, No. 1, 155-172. (2011)
17. Glänzel, W and Moed, H.F.: Journal impact measures in bibliometric research. Scientometrics, Vol. 53, No. 20, 171–193. (2002)
18. Garfield, E.: Use of Journal Citation Reports and Journal Performance Indicators in measuring short and long term journal impact. Croatian Medical Journal, Vol. 41, No. 4, 368–374. (2000)
19. Garfield, E.: The history and meaning of the journal impact factor. Journal of the American Medical Association, Vol. 295, No. 1, 90–93. (2006)
20. Bensman, S. J.: Distributional differences of the impact factor in the sciences versus the social sciences: An analysis of the probabilistic structure of the 2005 journal citation reports. Journal of the American Society for Information Science and Technology, Vol. 59, 1366–1382, doi: 10.1002/asi.20810 (2008)
21. Bensman, S. J., Smolinsky, L. J. and Pudovkin, A. I.: Mean citation rate per article in mathematics journals: Differences from the scientific model. Journal of the American Society for Information Science and Technology, Vol. 61, 1440–1463, doi: 10.1002/asi.21332 (2010)
22. Bordons, M., Fernández, M.T. and Gómez, I.: Advantages and limitations in the use of impact factor measures for the assessment of research performance in a peripheral country. Scientometrics, Vol. 5392, 195–206. (2002)
23. van Leeuwen, T.N. and Moed, H.: Development and application of journal impact measures in the Dutch science system. Scientometrics, Vol. 5392, 249–266. (2002)
24. Buela-Casal, G.: Evaluating quality of articles and scientific journals. Proposal of weighted impact factor and a quality index. Psichothema, Vol. 15, No. 1, 23–35. (2002)
25. Bollen, J., Rodriguez, M.A. and van de Sompel, H.: Journal status. Scientometrics, Vol. 69, No. 3, 669–687. (2006)
26. Sombatsompop, N., Markpin, T. and Premkamolnetr, N.: A modified method for calculating the Impact Factors of journals in ISI Journal Citation Reports—Polymer Science Category in 1997–2001. Scientometrics, Vol. 60, No. 2, 235–271. (2004)
27. Frandsen, T.F. and Rousseau, R.: Article impact calculated over arbitrary periods. Journal of the American Society for Information Science and Technology, Vol. 56, No. 1, 58–62. (2005)
28. van Leeuwen, T.N. and Moed, H.: Characteristics of Journal Impact Factors—The effects of uncitedness and citation distribution on the understanding of journal impact factors. Scientometrics, Vol. 63, No. 2, 357–371. (2005)
29. Egghe, L.: A rationale for the Hirsch-index rank-order distribution and a comparison with the impact factor rank-order distribution. Journal of the American Society for Information Science and Technology, Vol. 60, 2142–2144, doi: 10.1002/asi.21121 (2009)
30. Bergstrom, C.: Eigenfactor: Measuring the value and prestige of scholarly journals. College & Research Libraries News, Vol. 68, No. 5, 314-316 (2007)
31. González-Pereira, B., Guerrero-Bote, V. P. and Moya-Anegón, F.: A new approach to the metric of journals's scientific prestige. Journal of Informetrics Vol. 4, No. 3, 379-391 (2010)
32. Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Hassan-Montero, Y., González-Molina, A., Moya-Anegón, F.: New approach to the visualization of international

scientific collaboration. Information Visualization, Vol. 9, No. 4, 277-287, doi: 10.1057/ivs.2009.31 (2010)

Dragan Ivanović has worked at the Department of Computing and Automatics, Faculty of Technical Sciences, Novi Sad on the position of assistant professor since 2010. Mr. Ivanović received his Master degree in 2006 and Ph. D. degree in 2010 in Computer Science from the University of Novi Sad, Faculty of Technical Sciences. He is the corresponding author and can be contacted at: chenejac@uns.ac.rs

Dušan Surla is a professor emeritus at the University of Novi Sad, Serbia since 2010. Mr. Surla received his Bachelor degree in Mathematics from the University of Novi Sad, Faculty of Philosophy in 1969, Master degree in Robotics from the University of Novi Sad, Faculty of Mechanics in 1976, and Ph. D. degree in Robotics from the University of Novi Sad, Faculty of Technical Sciences in 1980. Since 1976 he is with the Faculty of Science in Novi Sad. He published more than 200 scientific and professional papers.

Miloš Racković is a full professor at the Department of Mathematics and Informatics, Faculty of Sciences, Novi Sad, Serbia since 2006. Mr. Racković received his Bachelor degree in Informatics in 1989, Master degree in 1993, and Ph. D. degree in 1996, all in Informatics at Faculty of Sciences, Novi Sad. Since 1989 he is employed at the Faculty of Sciences in Novi Sad. He published more than 80 scientific and professional papers.

Received: August 01, 2011; Accepted: February 06, 2012.

DOI: 10.2298/CSIS101222001W

Ontology-Based Home Service Model

Moji Wei1, Jianliang Xu2, Hongyan Yun2,
and Linlin Xu2

[1]Information Research Institute, Shandong Academy of Sciences,
19th Keyuan Road, Jinan, Shandong Province, China
weimoji@126.com
[2]Department of Computer Science, Ocean University of China,
238 Songling Road, Qingdao, Shandong Province, China
{xjl9898, yunhongyan, gxl0216}@gmail.com

Abstract. This paper researches Home Service retrieval and invocation for smart home. We represent ontology-based Home Service Model to retrieve and invoke services according to user’s needs automatically. Firstly by analyzing the context of home service, we differentiate key concepts in the field and analyze the relations among them, and as a result, an upper ontology as a fixed viewpoint for further more detailed conceptualization is achieved. Then by reifying the concepts of the upper ontology, we construct two domain ontologies which are Function Concept Ontology and Context Concept Ontology to annotate the semantic of Home Service from different facets. The Function Concept Ontology is constructed with the guidance of Maslow's Hierarchy of Needs to annotate the goal of service for automating service retrieval. For service invocation, the Context Concept Ontology is constructed by analyzing the contents that services operate. Finally two scenarios for different types of services are given to demonstrate the usage of the model.

Keywords: Smart Home, Home Service, Domain Ontology, Need, Function

1. Introduction

The convergence of Telecommunication Network, Internet and CATV Network enhances Home Area Network (HAN). The enhanced HAN connects almost all the home appliances and integrates various services which were restricted in one single network. The central of the HAN is a residential gateway which bridges the HAN and the Internet. It manages multiple devices to connect to the Internet simultaneously. With the services provided by various devices, in home users could access Internet facilely, and outside home they could control the devices remotely. Accordingly invoking proper services to drive the devices performing suitable functions becomes the key of satisfaction of needs.

For communication within HAN, various standards have been given by different alliances from different perspectives. According to the alliances, existed standards could be classified into four categories. The first kind of standard is supported by network equipment manufacturers, such as OSGi (Open Service Gateway Initiative) [1, 2], IHA (Internet Home Alliance), etc. The second kind of standard is proposed by IT vendors, like DLNA (Digital Living Network Alliance) [3], UPnP (Universal Plug and Play) [4], etc. The third kind of standard is advocated by home appliance manufacturers, such as ECHONET (Energy Conservation and Homecare Network) [5] etc. The last kind of standard is lead by manufacturers of automatic control, like LonMark [6] etc. Besides above alliances, many International Organizations for Standardization like CEA (Consumer Electronics Association), ITU (International Telecommunications Union), ETSI (European Telecommunications Standards Institute), etc found institutions for HAN framework either. The standards offered by the alliances are different in technical layer because of different perspectives, but the models of HAN are similar. Figure 1 depicts the general HAN model. In Home Area Network, all the appliances are connected to residential gateway directly or indirectly, and access information from Internet through the gateway. Furthermore, in order to manage home appliances remotely, users would employ controller or other devices which could access Internet and send instructions to the gateway, then the gateway would distribute instructions to relevant appliances. The residential gateway-centered HAN forms the infrastructure of smart home.

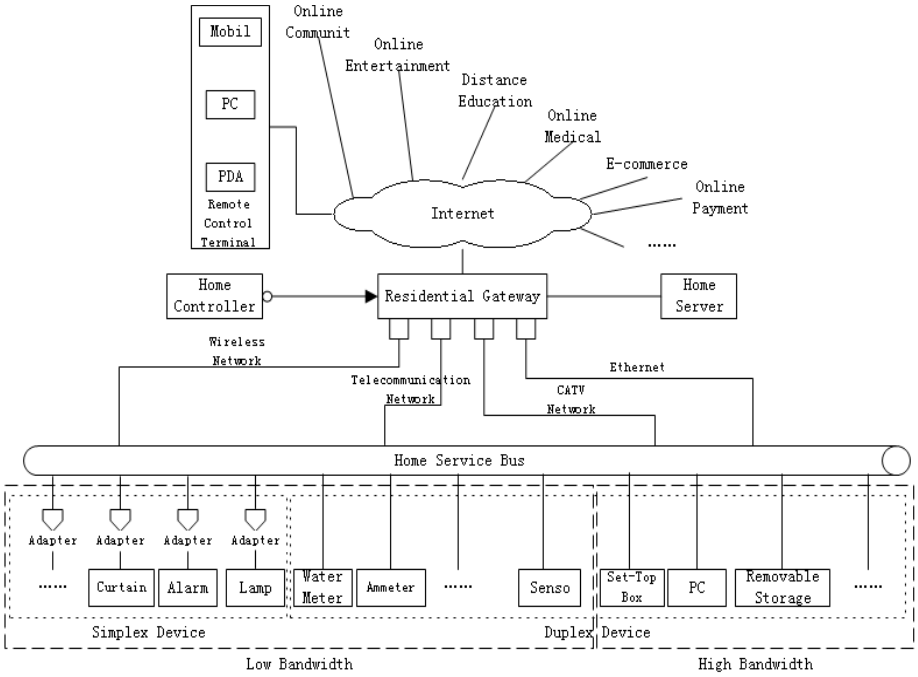


Fig. 1. Home Area Network Model

Home Services, which are implemented by home appliances, have narrow sense and broad sense. In a narrow sense, Home Services just mean the services used to control home appliances, thus the interactions among the services are restricted in HAN. While in a broad sense, Home Services also contain the services that interact with applications running on the Internet, for instance online medical services and distance education services are such services. Correspondingly researches about Home Services mainly come from two fields. One is field about automatic control of home appliance, which focuses on automation and responding rate of devices. It researches how to operate the devices correctly in different situations according to users' expectation, and rarely concerns the ability of network access. The other field originates from Web Service, and it focuses on the communication among services.

In order to facilitate Home Service accessing, the alliances mentioned above provide APIs for users or use the sophisticated protocols like HTTP, SOAP, etc. However these interfaces or protocols are informal expressions, therefore the service retrieval and invocation are based on syntactic matching. As other Web sources retrieving, the syntactic approach which uses keywords to match service has several drawbacks, including problems with synonyms (semantically similar but syntactically different services) and homonyms (syntactically equivalent but semantically different services).

Firstly in service discovery and selection, keyword-based mechanism, which does not consider the semantics of the requested services, usually illuminates the services with WSDL [7] which just describes the interface features. As a result the services selection can only retrieve service references that exactly match the service interface name in the query or whose properties include the same syntactic keywords when using a filter. It might result in a list of retrieved services which have the same interface but different functions.

Secondly in services invocation, keyword-based solution mandates that the consumer should know the name of service and the number and types of parameters. This is clearly an obstacle in a pervasive environment, because it prevents unaware services from dynamically invocation.

The automation of service retrieval and invocation demands computer-interpretable description which could not be tackled in syntactic layer. Therefore the service should be semantically annotated with a formal language. Ontology which is a formal, explicit specification of a shared conceptualization provides a common understanding by annotating the knowledge with concept. So in home field, domain-specific ontologies that semantically organize typical Home Services are critical to automate service retrieval and invocation.

Researches have proposed an upper ontology called OWL-S [8] for Web Service based on OWL (Web Ontology Language). It provides three essential classes for service: ServiceProfile, ServiceModel and ServiceGrounding. ServiceProfile is used for advertising and discovering services. ServiceModel gives a detailed description of a service's operation. ServiceGrounding provides details on how to interoperate with a service, via messages. In OWL-S, both the ServiceProfile and the ServiceModel are thought of as abstract representations; only the ServiceGrounding deals with the concrete level of specification, it could be thought of as a mapping from an abstract to a concrete specification of those service description elements that are required for interacting with the service. OWL-S, in conjunction with domain-specific ontologies for smart home field, provides standard means of specifying declaratively descriptions for Home Services that would enable retrieval and invocation automatically.

Home Service is an extension of Web Service for smart home. However comparing with Web Service [9, 10] on the Internet, Home Service owns inimitable features.

Firstly, most Web Services on the Internet are run on the computers, while the devices which run Home Services are various home appliances not limited to computers, therefore the equipments Home Services use are much more abundant.

Secondly, Web Services on the Internet focus on the communication among services rather than the interaction between services and environment. While an important function of Home Services is to control the devices adjusting states of environment. They play an important role in physical effects, hence the interaction between Home Services and environment could not be ignored.

According to comparison between Home Service and Web Service on the Internet, it can conclude that OWL-S could formally describe the semantics of Web Services, however it is still not sufficient for the description of device, context and human’s need. First OWL-S supposes that all the services are run on the computers, accordingly it has no concept for device. But in smart home the service is implemented on multifarious devices, thus the device concept could not be ignored. Second OWL-S just describes the effect which means the change of the context brought by service implementation, and the effect is treated as expression. However, in smart home, the context of Home Service is much more complicated than the context of Web Service. So in order to guarantee service invocation correctly, the context should present semantic information that context-aware services can use. It means that just using expressions to describe context is not enough, we should provide ontology rather than expression for context description. Finally human plays an important role in smart home and human’s need is a key factor for service retrieval and invocation, but the need is not involved in OWL-S.

Many literatures [11-17] have proposed relative domain ontologies of smart home for the concepts like device, place, policy, etc. The domain ontologies provide semantic information for Home Service from different facets. However to the best of our knowledge no attempts have been made to construct a domain upper ontology for smart home. Namely literatures independently construct various ontologies for the classification of concepts of smart home field without defining the concepts themselves and the relations among them formally. The absence of upper ontology would lead to concepts misuse. The current state of the literatures of home domain ontologies lacks the understanding of the correlation and inter-dependency of concepts within the field. In turn, this obscurity is hindering the advancement of this research field. As a result, the introduction of an organization of the domain knowledge and their relations is necessary to help the research community achieve a better understanding of this field.

The contribution of this paper is three-fold. First it presents Home Service Ontology which is domain upper ontology of smart home. The ontology explicates the key concepts abstracted from home context and presents the relations among the concepts formally, especially it differentiates the concepts of need and function which are two easily-confused concepts. Second with the domain upper ontology as a fixed viewpoint, we reify the concepts in it and propose several domain ontologies for further more detailed conceptualization. Finally we propose an ontological model for Home Services to organize the ontologies and services layered.

The rest of this paper is organized as follows. Section 2 extracts the concepts from the field of smart home and describes the relations among them based on the analysis of Home Service context, and then proposes our model for Home Service. In section 3 the Home Services are classified into two types according to the property of HAN. Then according to the types of services, section 4 represents two relevant scenarios to demonstrate the usage of the model proposed in the second section. Section 5 shows how to search the Home Services which are mentioned in above two scenarios in details. Related works about Home Service Model and Function Concept Ontology are introduced in section 6. The final section concludes the paper and presents the future work.

2. Home Service Model

2.1. Environment Analysis of Smart Home

Home Service is a software component that is implemented on the computer. Here the computer is a broad sense computer including PC (Personal Computer), SCM (Single-Chip Microcomputer), etc. In order to achieve specified function that user expects, the service should control home appliance to interact with the context. Consequently, to extract service-related concepts from the field of smart home, we should analyze the context where Home Service is run first.

Rosenman and Gero [18, 19] extend function-behavior-structure framework for improving intelligence of design system in mechanical manufacture, they analyze the relations among techno-physical environment, nature environment and socio-culture environment which have close relation with artifacts. For concept explanation they distinguish the context into three parts: External World, Interpreted World and Expected World.

Home Service is also an artifact. However, unlike other artifacts, Home Service as a software component is not substantial. It means that the implementation of service requires device as its carrier. So, Computer World as the carrier of service should also be involved in Home Service context.

Therefore we distinguish the home context into four parts: External World, Interpreted World, Expected World and Computer World. Figure 2 depicts the relations among the four worlds.

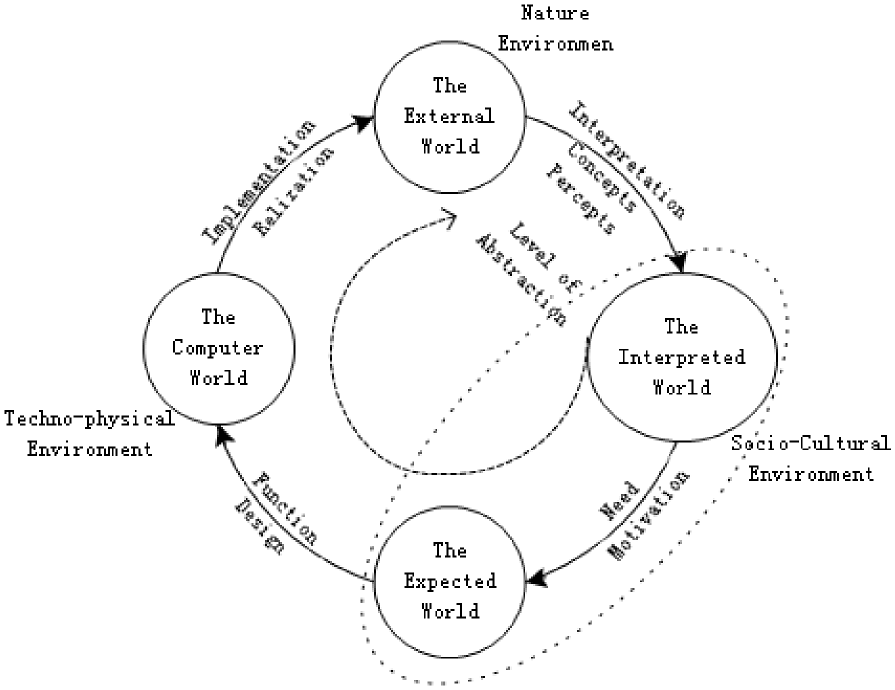


Fig. 2. Four Worlds

The External World pictures the current states of the physical world.

The Interpreted World is the one built up inside the human beings in terms of sensory experiences, percepts and concepts. That is, the internal representation of the External World that human interacts with. It triggers lots of unsatisfied needs.

The Expected World is the one being produced by using devices so as to satisfy the needs of human beings. The functions of devices can be imagined according to current goals of human and interpretations of the current states of the world.

The Computer World is the one consisting of devices which are designed to provide specific functions. It is responsible for transforming the expected world into the external world.

The four worlds are recursively linked together by four classes of processes. The Interpretation process transforms variables sensed in the External World into interpretations of sensory experiences, percepts and concepts which in turn become part of the Interpreted World. This process is done by the interaction of sensation, perception and conception processes. Then in Motivation process the Interpreted World is transformed into the Expected World motivated by needs which are generated by human experiences. The needs are expressed as goals in the Expected World. The devices in the Computer World are designed to achieve the goals proposed in the Expected World. The Design process transforms the goals which are proposed in the Expected World into the functions that could be provided by devices, and then the devices in the Computer World would be designed according to the functions. Finally the process of Implementation is an effect which brings about a change in the External World according to the function provided by the Computer World. The four worlds from the Interpreted World to the External World materialize conceptions of human beings gradually, and the level of abstraction is decreased as well. The Home Service which is a kind of artifact just sites in the Computer World and would satisfy users' needs by controlling the device to realize specified function.

2.2. A Domain Upper Ontology

By distinguishing home context into four worlds, we could abstract six service-related concepts including “Need”, “Context”, “Device”, “Function”, “Content” and “Service”. The following would explicate these concepts respectively.

Home Service would influence the external world by device operation. The effect of service is recognized as function which implies the goal of service. We say that the service satisfies need, when the effect plays a positive role for human. Need and function are two easily-confused concepts. So we distinguish the two concepts from origination first. As figure 2 shown, they are abstracted from different processes. Need originate from the Motivation process where the Interpreted World is transformed into Expected World; while function sites in Design process where the Expected World is transformed into Computer World. The levels of abstraction of two concepts are different as well. According to above section analysis the need concept is more abstract than the function concept.

In addition, the service as software may operate some data which is defined as content in the paper. The service would be recognized as different applications when cooperating with different contents. Take the questionnaire service as example, when the content consists of simple choice questions of some course, the service becomes an examination; while the service becomes a psychological test if the content consists of choices to certain circumstances one meets in daily life.

2.2.1 Home Service Ontology

To clarify the concept relations among need, function, service, content etc, this paper constructs Home Service Ontology as an upper ontology for smart home field. The ontology is expressed in the form of RDF diagram in figure 3.

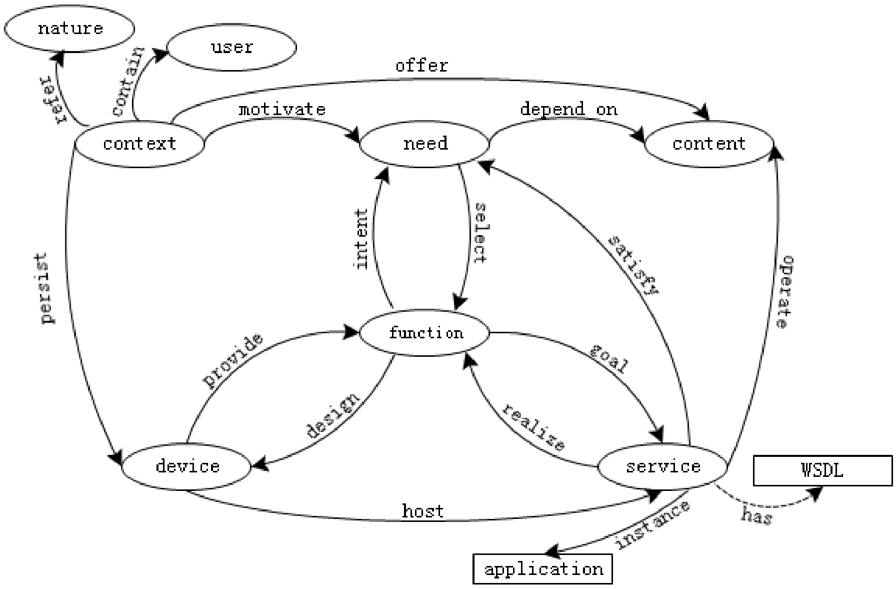


Fig. 3. Home Service Ontology

As shown in figure 3, human's *need* is *motivated* by *context*. Since the context varies dynamically, the needs are changeful and diverse in daily life. On the one hand a need could be satisfied by various services; on the other hand one service could satisfy a variety of needs motivated by different contexts. For example, an alarm informing need is motivated when insecure situation occurs in smart home. Then for residential users, the need could be satisfied by audio-playing service which is run on audio device, and for outworkers, they could be informed the alarm by using SMS service provided by cell phone. Besides the alarm informing need, other needs like note reminding need and music playing need could also be satisfied by the audio-playing service. So it is not proper to annotate the audio-playing service with alarm informing need, namely the need concept could not explicate the semantic of service. Psychology researches that with the same need the goals may be such differences for kinds of people, and on the other side to the same goal the needs from different people are inconsistent as well [20, 21]. If we examine carefully the average needs that we have in daily life, we find that they have at least one important characteristic, i.e., that they are usually means to a goal rather than goals in themselves. Usually when a conscious need is analyzed we find that we can go behind it, so to speak, to other, more fundamental intentions of the individual. It is characteristic of this deeper analysis that it will always lead ultimately to certain intentions behind which we cannot go; that is, to certain intention-satisfactions that seem to be goals in themselves and seem not to need any further justification or demonstration. The fundamental or ultimate intentions of all human beings do not differ nearly as much as do their conscious everyday needs [22]. The intentions are always implied in the needs. Therefore to satisfy a need, consumer usually has to select a proper *function* which reflects the *intention* implied in the *need*. Moreover the function presents the *goal* of the *service*, and meanwhile it also guides the *design* of *device*, hence the device *provides* the function exactly. The *realization* of given function relies on service implementation, and the precondition is that the device which *hosts* the service should be persisted in the context. Additionally the service would be implemented as different *applications* when cooperates with different *contents*, so we can conclude that the need *depends on* specific *content* as well. All the contents that the service *operates* are *offered* by the context. The home context consists of *device*, *user* and *nature* environment. Thus the content comes from above three sources.

2.2.2 The functional Semantic of Service

The Home Service Ontology explicates our viewpoint and formally defines service-related concepts from the unified viewpoint. In the ontology we differentiate function from need, which are usually confused with each other in other literatures. For instance other literatures recognize “illumination” and “guard” as the functions of “turn-on light” service. However from our viewpoint described above, the intention of “turn-on light” service is to adjust the brightness state of the External World, accordingly the service influences human vision, and then the function of the service is vision-related. The service which realizes vision-related function would actually alter the states of External World, and in different contexts the altered states may lead to different effects which eventually satisfy various needs such as “illumination” or “guard”. In our opinion, the function of a service is unique and fixed, while the needs which a service could satisfy are diverse with different contexts. Thus, annotating service with function could explicate the functional semantic of service.

The Home Service Ontology as domain upper ontology defines key concepts in the field, and formalizes relations among them. It formally clarifies the whole process from need motivation to service invocation, therefore annotating the service with the concepts in the ontology could promote the automation and intelligence of smart home. However, the concepts described above are highly abstracted, they can hardly be used to annotate services for exactly service searching. Thus we construct domain ontologies for smart home by reifying the concepts in Home Service Ontology.

2.3. Function Concept Ontology

As Home Service Ontology analysis shown in figure 3, function which indicates the goal of service and the intention of need bridges the need and the service. Annotating service with function would automate service retrieval based on human’s need. Therefore to classify service in semantic layer we should research the category of function.

As interpreted in figure 2, people perceive the External World with multi-sensory organs such as eyes, ears, nose etc. As a result in Interpreted World the conception model which represents External World that the human interacts with is also combined with many factors. Then the ideal model of Expected World evolves from conception model is also composed with multi-factors. It means that the satisfaction of need often refers to several factors.

On the other hand service which is implemented in Computer World is software component. The essence of component determines that service could only adjust limited states of External World, i.e. the functions realized by service are limited. According to software engineering the module should be highly cohesive and loosely coupled. Based on the principle an optimal service should just realize one function and alter one factor only. Consequently to automate service retrieval for need, the need should be decomposed according to factors. The decomposed factor should coincide with the service. So the factors decomposition is vital for automation.

The need situates in the Motivation process from Interpreted World to Expected World, namely need stimulates conception model to ideal model. While the target of our research is how to reuse existed services to satisfy need with matured ideal model, therefore our focal issue situates in the Design phase that transforms Expected World to Computer World. Accordingly we should abstract factors for service annotation in Design phase.

From relations of concepts shown in figure 3, it is known that the device in techno-physical environment is designed for specific function which denotes the goal of service. By annotating the goal of service with function, the service could be searched semantically. We differentiate function from need and annotate service with function rather than need, however the function still reflects intention which is implied in need and function selection relies on need as well. It means that the function can not isolate from need absolutely.

In conclusion, we should construct Function Concept Ontology to decompose factors with the guidance of a need theory, and then annotate the service with the classified function to present functional semantic information.

2.3.1 Maslow’s Hierarchy of Needs

Maslow’s hierarchy of needs [23] is layered need theory in psychology. Many researches have advanced Maslow’s theory and proposed several need theories [24-27] for different social formations and technology domains. However, to the best of our knowledge there is still no need theory is provided for smart home. Hence in this paper we would use Maslow’s hierarchy of needs as guidance and consult existed services to construct our Function Concept Ontology for service annotation.

Maslow's hierarchy of needs is often portrayed in the shape of a pyramid, with the largest and most fundamental levels of needs at the bottom, and the need for self-actualization at the top. It has five layers: Physiological need, Safety need, Belongingness and love need, Esteem need and Self-actualization need.

In Maslow’s theory, Physiological needs are the literal requirements for human survival. If these requirements are not met, the human body simply cannot continue to function. Physiological needs are usually related to body senses. With their physical needs relatively satisfied, the individual's Safety needs take precedence and dominate behavior. These needs have to do with people's yearning for a predictable orderly world in which perceived unfairness and inconsistency are under control, the familiar frequent and the unfamiliar rare. In the world of work, these Safety needs manifest themselves in such things as a preference for job security, grievance procedures for protecting the individual from unilateral authority, savings accounts, insurance policies, reasonable disability accommodations, and the like. After Physiological and Safety needs are fulfilled, the third layer of human needs is social and involves feelings of belongingness. This aspect of Maslow's hierarchy involves emotionally based relationships in general. Humans need to feel a sense of belonging and acceptance, whether it comes from a large social group, or small social connections. They need to love and be loved by others. After that all humans have a need to be respected and to have self-esteem and self-respect. Also known as the belonging need, esteem presents the normal human desire to be accepted and valued by others. People need to engage themselves to gain recognition and have an activity or activities that give the person a sense of contribution, to feel accepted and self-valued, be it in a profession or hobby. Finally the Self-actualization need pertains to what a person's full potential is and realizing that potential. Maslow describes this desire as the desire to become more and more what one is, to become everything that one is capable of becoming.

2.3.2 The Method of Function Extraction

Referring to Maslow’s hierarchy of needs, our Function Concept Ontology is also classified into five layers. In each layer we collect relative needs and then extract functions from the needs. The following would give the method of how to extract functions from needs.

As Home Service Ontology defined, both of need and function which reflect people’s intention are subjective. The difference between them is that function is context-independent while need interprets the intention additional with special context. Intuitively as shown in figure 3, the “need” concept is directly connected with “context” concept by the object property “motivate”; while the “function” concept has no relation with “context” concept. Since the context varies dynamically, the needs are changeful and diverse in daily life. Extracting from specific context, the function directly depicts the intention which is more fundamental.

In conclusion the difference between function and need is whether the interpretation of intention depends on the context. For instance the alarm informing need for insecure situation, the music appreciation need for entertainment, and the voice reminding need for important message are different needs motivated by different contexts. Eliminating the context, the intention of the needs is the same, that is an audio play. Then the “play” is the very function concept extracted from the needs. Hence we would extract concepts of function from various needs by eliminating the contexts which the needs rely on, then organize the concepts according to layers of function denoted by Maslow’s hierarchy of need. Figure 4 gives a portion of function concepts.

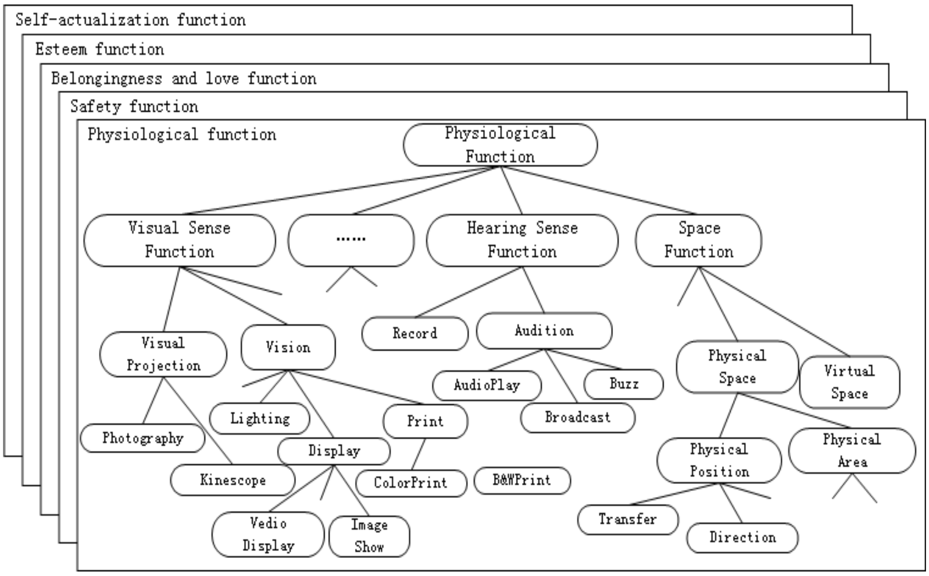


Fig. 4. Part of Function Concept Ontology

2.4. Context Concept Ontology

According to the former sections analysis, function presents the intention of need and the goal of service. It bridges the need and the service. By annotating service with the concepts in Function Concept Ontology, candidate services could be automatically retrieved for need. Then to decide which service would finally be invoked, there exist numerous rules, such as utility rule, cost rule, etc. Among the various rules the essential principal and the least constraint rule is that the context should offer all the contents that the service requires. So for a service invocation we should check whether the context offers suitable contents that could be operated by service. Therefore analyzing the elements of the context and expressing them with formal language to provide semantic information for context would assist automating service invocation.

Figure 1 shows that the boundary of the smart home model is device which would interact with user and nature environment. Consequently Home Service would communicate with device, user and nature environment only, and then the content that Home Service operates could only be derived from above three sources either. Consulting relative domain ontologies of other literatures, we construct the Context Concept Ontology by extracting concepts from device, user and nature environment. Figure 5 shows portion of the ontology and the hierarchy of concepts. With the Context Concept Ontology annotation, whether the least constraint rule is satisfied could be inferred automatically.

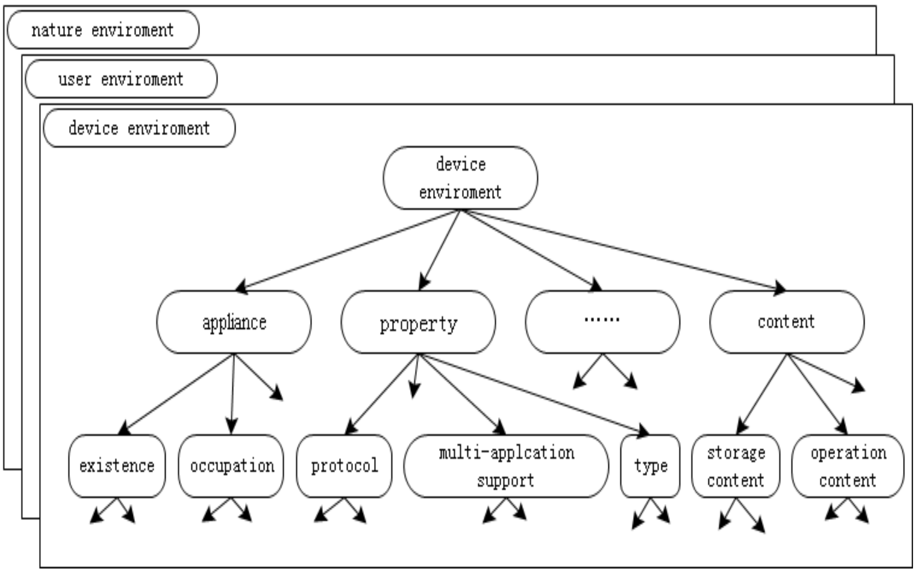


Fig. 5. Context Concept Ontology

2.5. Ontology-Based Home Service Model

Above three sections present three ontologies respectively: Home Service Ontology, Function Concept Ontology and Context Concept Ontology which would be scattered into two different layers. These three ontologies in conjunction with top-level ontology, rules and service repository compose our Home Service Model. Figure 6 represents the framework of the model.

As shown in figure 6, the Home Service Model has 5 layers.

The first layer is top-level ontology which is similar to other ontology-based models such as Eco-ontology [28], SWEET [29], MMI [30], etc. The top-level ontology defines the concepts like time, space, matter, energy, etc.

The second layer of Home Service Model is domain upper ontology which is used to formalize the concepts extracted from the smart home field. It provides theoretical support for the following layers. The ontologies in the first two layers are used to clarify the fundamental of the field of smart home.

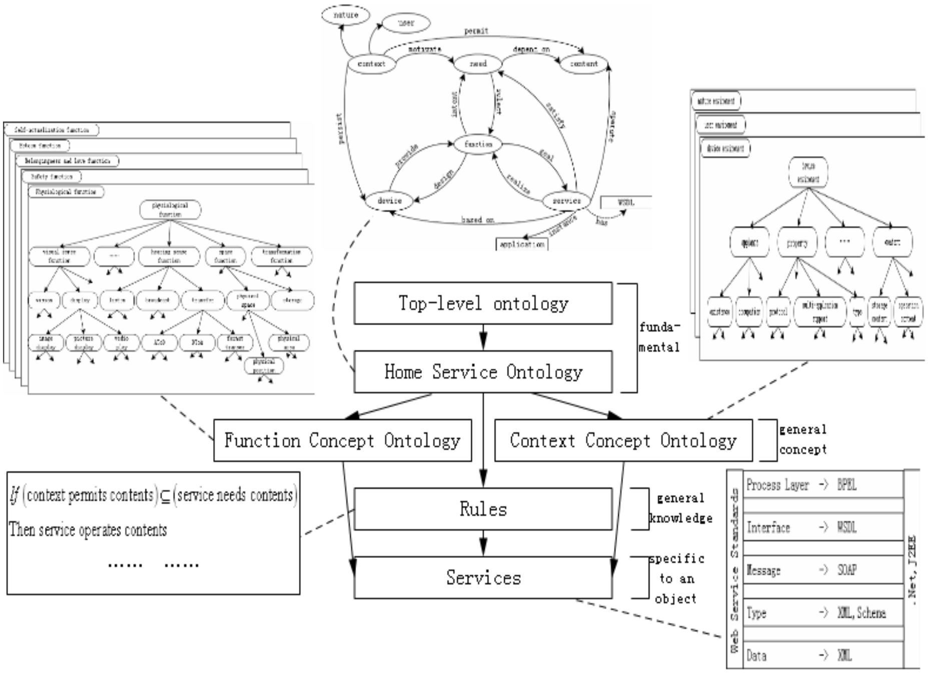


Fig. 6. Home Service Model

In the third layer there are two domain ontologies: Function Concept Ontology and Context Concept Ontology. They provide general concepts for the field from different facets. The Function Concept Ontology is constructed with the guidance of Maslow's hierarchy of needs, and the function concepts in it present the goals of services. With the annotation of Function Concept Ontology, the service could be automatically retrieved for need. Then for automating Home Service invocation, the Context Concept Ontology is constructed. It extracts concepts from device, user and nature environment. With the annotation of Context Concept Ontology, whether the context offers compatible content for service operation could be inferred.

The rules in the fourth layer which is maintained by user express user's general knowledge and exhibit user's preference. Initially the set of rules in Home Service Model provides just one default rule that is the least constraint rule, the formal expression of the rule is shown as Rule-1. Then users could customize rules set by appending new rules according to their preference.

$$\begin{array}{l} IF \quad (\text{context offers contents}) \supseteq (\text{service requires contents}) \\ \text{THEN service operates contents} \end{array} \qquad \text{Rule-1}$$

Finally the service repository in the bottom layer is labeled with criterions from SOA (Service-Oriented Architecture) which is a well adopted architecture for Web Service organization. For instance the interface of Home Service is described by WSDL (Web Services Description Language) and the communication protocol uses SOAP (Simple Object Access Protocol). With these universal criterions the Home Services in our repository could be invoked by other users with other models.

3. Home Services Analysis

Before demonstrating the usage of the model, we discuss the types of Home Services first, and then give one scenario for each type to simulate the solution.

As shown in figure 1, in HAN model there are two kinds of bandwidths: high bandwidth and low bandwidth and two kinds of devices: simplex device and duplex device. The HAN could be classified into 3 categories logically according to bandwidths and devices. The first one is Control Net. The devices in this kind of net just receive instructions from residential gateway, therefore the bandwidth required is very low. Light and alarm bell are such devices that are used in Control Net. The second category is Information Net. The devices in Information Net exchange a little information with the gateway, consequently the bandwidth is not high. The devices like sensor, ammeter are used in Information Net. They have to response the query from the gateway. The third category is Service Net. The devices in Service Net would exchange large amounts of data with the gateway, thus the bandwidth is very high. Such as VOD (Video On Demand) application, the TV has to receive large amounts of video data from the gateway.

According to the categories of network, the service could be classified into two types. First the services implemented in Control Net concentrate on controlling home appliances rather than content operation, so we call this type of services Control Service. Second the services implemented in Information Net or Service Net not only control devices but also operate contents, so we call this type of services Content Service. Table 1 presents the relations among service type, network, bandwidth, device and content.

Table 1. Service Type

Service type	Network	Bandwidth	Device	Require content
Control Service	Control net	Low	Simplex device	N
Content Service	Information net	Low	Duplex device	Y
	Service net	High	Duplex device	Y

4. Scenarios

As last section analysis, the Home Services could be classified into two types: Control Service and Content Service according to property of HAN. In this section we would display two scenarios for above two types of services respectively. The scenarios show the entire process from need generation to service invocation to demonstrate the usage of the Home Service Model.

4.1. Scenario for Control Service

This subsection presents a scenario that displays the process of solution of Control Service. The essential service that smart home should supply is security service. The model should inform the insecurity when dangerous affairs occur. Figure 7 displays the scenario to simulate the solution.

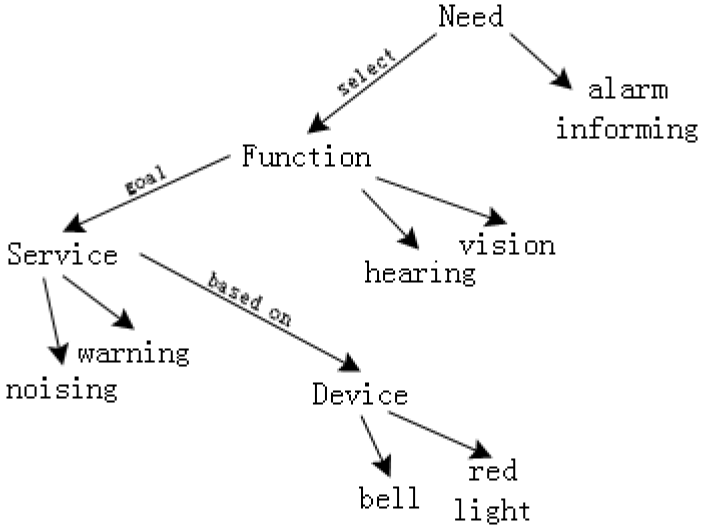


Fig. 7. Scenario for Control Service

It is obvious that insecure affairs would motivate “alarm informing” need. The model could select “vision” function and “hearing” function for the intention of the need. Then according to annotation of Function Concept Ontology, the “warning” service for “vision” goal and “noising” service for “hearing” goal could be retrieved. Afterward the model would reason Context Concept Ontology to check whether the devices “bell” and “red light” which implement two services separately are offered by the context. Finally the need could be satisfied, if the context does offer at least one device.

4.2. Scenario for Content Service

This subsection presents a scenario that displays the process of solution of Content Service. In smart home besides the security related services, the entertainment services should also be provided. After busy working, people may listen to the music for relaxation. Figure 8 displays the scenario to simulate the solution.

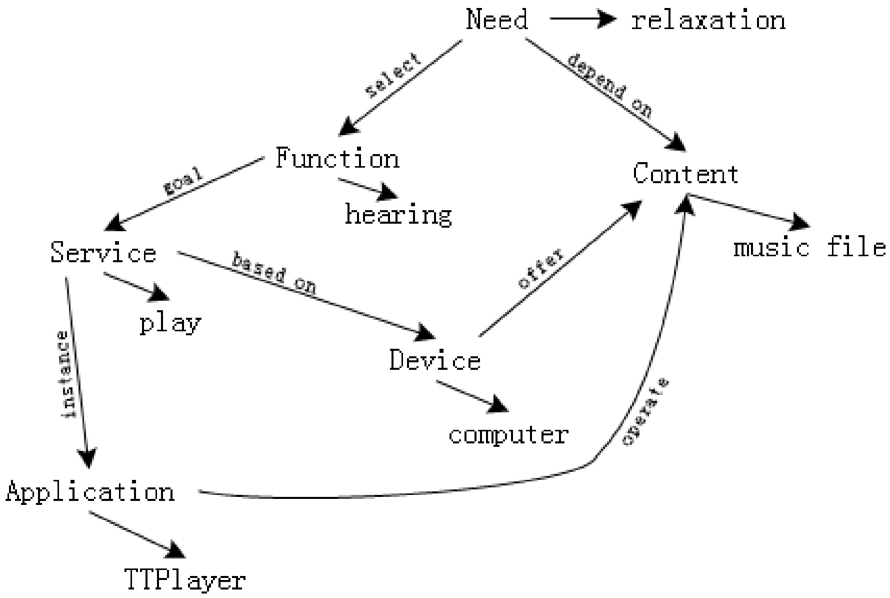


Fig. 8. Scenario for Content Service

It is obvious that the need in the scenario is “relaxation”. The intention could be achieved by “listen to music”. Thereby the selected function is “hearing” and the corresponding content is “music file”. Then the model may get “play” service for the “hearing” goal according to annotation of Function Concept Ontology. The model would locate the device “computer” which implements the “play” service by reasoning Context Concept Ontology, and then check whether the required content is offered by the device. If the device does store the content “music file”, the model may use “TTPlayer” which is an application of “play” service to operate the content.

In a word, in the Home Service Model, Home Service Ontology formally represents the whole process from need motivation to service invocation by explicating the relations among fundamental concepts. The relations in the ontology are defined by object properties of which domain and range connect two concepts respectively. Therefore by object property reasoning in DL (Description Logic), the model could automatically progress the phases of process. Then for various highly abstract concepts in the process, we reify them with the concepts from two domain ontologies to annotate the semantics of service.

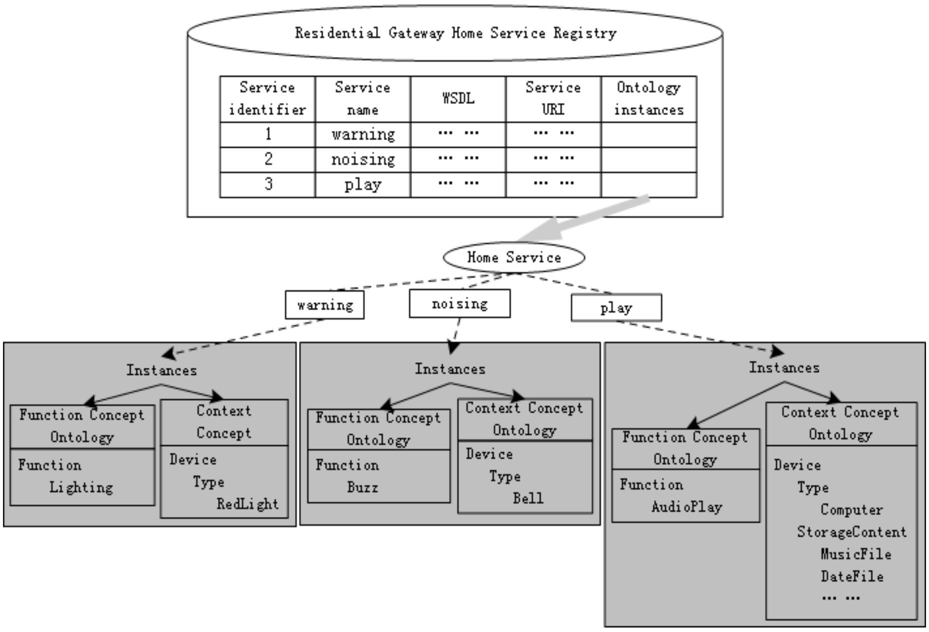
5. Ontology-Based Home Service Search

Last section has shown two scenarios about how to accomplish process with the guidance of Home Service Model roughly. In this section we would give more details about the process of service search.

As figure 1 shown, the residential gateway controls all the appliances in the smart home, and it is the center of HAN. Therefore we embed Home Service Model into the gateway to manage Home Services.

5.1. Home Service Register

When the Home Services are deployed into smart home, they should also be registered to Home Service registry which lies on the residential gateway. The current service registry of residential gateway stores a table-based representation of the Home Service. Then in service discovery phase, the residential gateway would search services by matching the keyword with service name syntactically. For appending addition semantic information to service, the service registry should be expanded as shown in figure 9.



Service identifier	Service name	WSDL	Service URI	Ontology instances
1	warning	… …	… …	
2	noising	… …	… …	
3	play	… …	… …	

Fig. 9. Instances of Ontology

We add Ontology instances item to the service registry. The new item which stores the instances of the concepts provides the semantic information for service. Figure 9 declares three instances for the services which are mentioned in the scenarios of last section. Referring to the Function Concept

Ontology, the function of “warning” service is annotated by “Lighting” concept, and the functions of “noising” service and “play” service are respectively annotated by “Buzz” concept and “AudioPlay” concept which are two sub-concepts of “Hearing” concept. That is declaring instances for the relevant function concepts. The function instances are declared as follow:

```
<Lighting   rdf:id=” Lighting_WarningService ”/>
<Buzz       rdf:id=”Buzz_NoisingService ”/>
<AudioPlayrdf:id=” AudioPlay_PlayService”/>
```

Then, as figure 9 shown, instances of Context Concept Ontology are declared to present the device information for services. For example the device type of “play” service is annotated with “Computer”, and the storage content is annotated with “MusicFile”, “DataFile”, etc.

5.2. Home Service Retrieval

In current service retrieval phase, the residential gateway searches Home Service registry by syntactically comparing keyword with service name. If there are some services matched, then the gateway returns the URI addresses where the services are deposited. However, the keyword-based approach would encounter synonyms and homonyms problems. For instance there is another service whose name is “play” and the function is to play video. Unfortunately the keyword-based approach could not distinguish two “play” services because of the same service name. Accordingly both “play” services would be retrieved by the keyword “play”, and apparently the video related service is improper service for user’s intention.

By enhancing residential gateway with Home Service Model, above syntactical problems could be solved by the appended semantic information. As figure 9 shown, each service in the registry is annotated with a function concept, namely an instance of specific function is declared to present functional information for the service. Therefore in our semantic approach the residential gateway would search services by reasoning the instances of the function which user requests.

For example when registering above video “play” service to Home Service registry, an instance of “VideoDisplay” concept would be declared for the service according to the Function Concept Ontology shown in figure 4. Then if an audio “play” service is wanted, the gateway would reason the instances of “AudioPlay” concept in Function Concept Ontology. All the services which are annotated with “AudioPlay” concept could be retrieved, and above video related “play” service which is declared as instance of “VideoDisplay” concept could be expelled.

In addition in DL instance reasoning the instances of sub-concepts are also instances of parent concept. As shown in figure 4 and figure 5, one function of two scenarios proposed in last section is “hearing”. Consulting Function Concept Ontology, “Buzz” concept and “AudioPlay” concept are two sub-concepts of “Hearing” Concept. Therefore by reasoning instances of “Hearing” concept both of “noising” service and “play” service could be retrieved. The names of above two retrieved services are such different and the functions are similar to some degree, which could hardly be achieved in keyword-based approach. Similarly “warning” service could be retrieved by reasoning instances of “Vision” concept.

Afterwards by searching Home Service registry, the URI addresses of retrieved services could be returned.

5.3. Home Service Invocation

Once candidate services are retrieved, the residential gateway should invoke the proper services automatically. As former sections analysis, the least constraint rule for service invocation is that the context should offer all the contents that service requires. Therefore the residential gateway has to check whether the device, user and nature environment are suitable for services invocation. That is to examine the compatibility between the instances of Context Concept Ontology and specific context.

For example both of “noising” service and “play” service could be retrieved by reasoning the instances of “Hearing” concept. The gateway should invoke proper service for specific context. As figure 7 for the first scenario depicted, the insecurity context requires “bell” device to inform alarm. Then the gateway would reason the relative instances of retrieved services. According to figure 9, the Device Type of “noising” service is annotated by “Bell”. It means that the “noising” service is compatible with the insecurity context, and then the gateway would invoke “noising” service to satisfy “alarm informing” need. Similarly with the first scenario, by reasoning Device Type instance and StorageContent instance for the second scenario, it can be drawn that the “play” service is compatible with the entertainment context and could be invoked to satisfy “relax” need.

6. Related Work

6.1. Research of Home Service Model

In Home Service invocation related field, several researchers have proposed several approaches for smart home. Choonhwa Lee et al. [31] propose using OSGi (Open Services Gateway initiative) as an infrastructure for integrating various devices and sensors to provide pervasive computing environments. However, they don’t resolve the service search and invocation problems. Dobrev et al. [32] tackle these two problems directly, but not from a semantic perspective. They import services from and export them to Jini and UPnP technologies using OSGi to bridge multiple discovery protocols. Tao Gu et al. [11] represent ontology-based context model. The model presents contextual information that can be used by context-aware applications. However it dose not describe physical effects by services execution. Daniel Retkowitz et al. [12] construct context model by extracting concepts from user and nature environment. The model ignores the information produced by devices. Diaz Redondo. et al. [13, 14] propose OWL-OS for Home Service by extending OWL-S, in their model the approach could discover, select and invoke services semantically.

Some initiatives have addressed the research issues of ubiquitous computing and user modeling for the ontological description of the smart home. Harry Chen et al. [15] describe a shared ontology called SOUPA - Standard Ontology for Ubiquitous and Pervasive Applications to represent intelligent agents with beliefs, desires and so on, and discuss the application in intelligence space. Dominik Heckmann et al [16, 17] present General User Model Ontology (GUMO) for the distributed user models in intelligent Semantic Web enriched environments.

The drawback of the models mentioned above is that the function concept and the need concept are confused. They annotate the functional semantic of service with need. The needs are diverse with various contexts. Accordingly in their model for different contexts one service could have multiple semantics, which would hinder service retrieval. In this paper we differentiate function from need, and adopt function as the functional semantic of service. The function is context-independent concept and is fixed for different contexts. Therefore annotating the service with function could facilitate service retrieval semantically.

6.2. Research of Function Concept Ontology

The Home Service Ontology discussed in Section 2 defines the concept of functionality strictly from the need-centered viewpoint. We extract function concepts by eliminating context factors from needs. Such definition is done intended to prescribe guidelines to functional modeling. We focus on purpose function rather than device function. Other types of function, however, still remain to be investigated. Consulting function categories proposed by Yoshinobu Kitamura [37] as shown in figure 10, this section discusses descriptive definitions of other kinds of function. Note that figure 10 shows an “is-a” hierarchy only for readability, because some distinctions are independent from each other.

Firstly, environmental function which represents changes outside of the device boundary is related to users or user actions. For example an electric fan performs cooling function for human body as environmental function where the cool-down effect by wind is on human body and thus outside of the system boundary. The environmental function could divide into physical environmental function and interpretational function. The physical environmental function represents physical changes of the environment in which device works, while an interpretational function sets up one of necessary conditions of human's cognitive interpretation. Chandrasekaran and Josephson [33] discuss a similar kind of function called environment function as effect on environment. They conceptualize “mode of deployment”, which is configuration of device and environment. Some researchers distinguish purpose from function [34, 35], where the purpose represents human-intended goal in the similar sense to the environmental function or interpretational function. Hubka [36] distinguishes the purpose function as effects from the technical function as internal structure. Rosenman and Gero [18] investigate purpose in socio-cultural environment. The situated FBS framework treats change of requirements [19]. Our function is kind of interpretation function, and we have differentiated function from need.

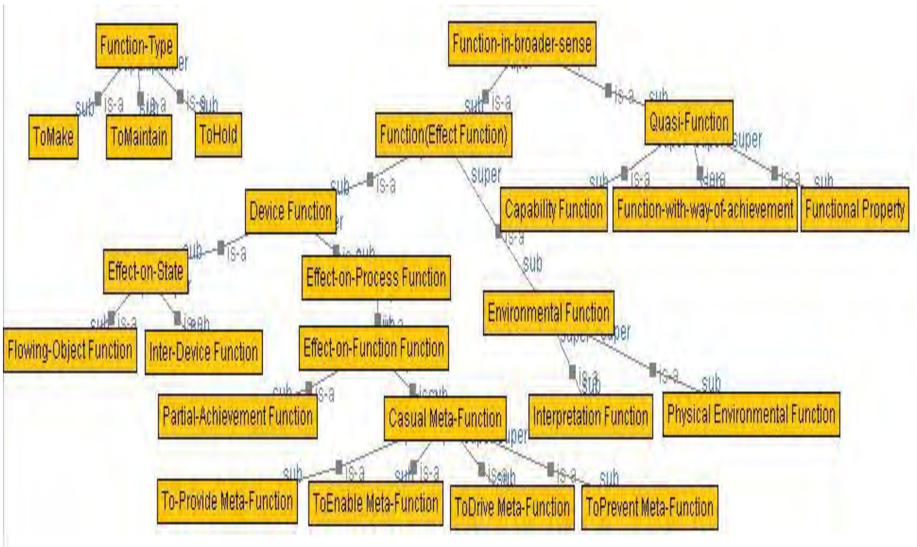


Fig. 10. Descriptive Categories of Function

Secondly, contrary to environmental function, device function considers changes within the device boundary. One of device function is effect-on-state function which focuses on temporal changes of physical attributes. The flowing-object function [37, 38] refers to temporal changes of physical attributes of objects which flow through the device and inter-device function [39] refers to changes of another device.

Another device function is effect-on-process function [40] which represents effect on a process or its changes. Behavior as basis of function can be regarded as a kind of a process. Thus, as a subtype of the effect-on-process function, effect-on-function function represents a role of a function for another function. It includes partial-achievement function and causal-meta function. The former is performed by a method function for a goal function in the is-achieved-by relation. The latter represents a role for another method function

Thirdly, we recognize the following three kinds of quasi-functions. Although the authors do not consider them as kinds of function, it is found that a quasi-function is confused with a function. Firstly, a function-with-way-of-achievement implies a specific way of function achievement as well as a function. Its examples include washing, shearing, adhering (e.g., glue adheres A to B). Because meaning of this type of function is impure, we regard it as quasi-function. Secondly, a functional property represents that an artifact (usually material) has a specific attribute-value which directly causes functionality. This is found in material science domain where a material whose function is dependent on its electronic, optical or magnetic property is called functional material [41]. For example, if an electrical conductivity of a material is high (i.e., it has high conductivity property), the material can perform the “to transmit electricity” function. There is direct relationship between the high-conductivity property and the transmitting function. Lastly, a capability function represents that an entity can perform an activity which is not effect on others. For example, people say that “a human has walking function”.

7. Conclusion

This paper presents an ontology-based Home Service Model which contains 5 layers for automatically progressing the phases of the process from need motivation to service retrieval and invocation. The model presents a domain upper ontology to clarify the relations among the concepts, and two domain ontologies to provide semantic information for service annotation from different facets. With assistant of above three ontologies, the model could retrieve and invoke feasible services automatically. Considering personality the model also offers a layer for rules which are customized according to user’s preference. Finally two scenarios, which are displayed for two types of services, are given to demonstrate the usage of model.

All the services discussed in this paper are atomic services. Along with the complication of task, several atomic services should be composed to satisfy one need. So in the future work we will research how to compose the services by functions. For this purpose we will adopt function decomposition tree to organize the functions in complex tasks.

References

1. Wu, C., Liao, C., Fu, L.: Service-Oriented Smart-Home Architecture Based on OSGi and Mobile-Agent Technology. IEEE Transactions on Systems, Man and Cybernetics 37(2). (2007)
2. Ishikawa, H., Ogata, Y., Adachi, K., Nakajima, T.: Building smart appliance integration middleware on the OSGi framework. Object-Oriented Real-Time Distributed Computing, 139–146. (May 2004)
3. Jung-Tae Kim, Yeon-Joo Oh, Hoon-Ki Lee, Eui-Hyun Paik, Kwang-Roh Park.: Implementation of the DLNA Proxy System for Sharing Home Media Contents. In IEEE Transactions on Consumer Electronics, vol. 53, 139–144. (2007)

4. Kang, D. O., Kang, K., Choi, S., Lee, J.: UPnP AV Architectural Multimedia System with a Home Gateway Powered by the OSGi Platform. IEEE Transactions on Consumer Electronics, Vol. 51, No. 1, 87-93. (February 2005)
5. Tajika, Y., Saito, T., Teramoto, K., Oosaka, N., Isshiki, M.: Networked home appliance system using Bluetooth technology integrating appliance control/monitoring with Internet service. IEEE Transactions Consumer Electronics, vol. 49, no. 4, 1043-1048. (Nov 2003)
6. Kastner, W., Palensky, P., Rausch, T., Roesener, CH.: A Closer Look on Today's Home and Building Networks. 7th AFRICON Conference in Africa, Vol. 2, 1239 – 1244
7. Chinnici, R., Gudgin, M., Moreau, J. J., Schlimmer, J., Weerarana, S.: Web Services Description Language (WSDL) Version 2.0. w3c.org, (March 2004)
8. Martin, D., Burstein, M., McDermott, D.: OWL-S 1.2 Release. Available at: http://www.daml.org/services/owl-s/1.2/
9. McIlraith, S., Son, T. C., Zeng, H.: Semantic Web Services. IEEE Intelligent Systems, 16(2):46–53. (March/April 2001)
10. Paolucci, M., Kawamura, T., Payne, T., Sycara, K.: Semantic matching of Web Services capabilities. In I. Horrocks and J. Hendler (eds.), Proceedings of the First International Semantic Web Conference (ISWC). Sardinia: Springer, 333–347. (2002)
11. Gu, T., Pung, H. K., Zhang, D. Q.: Toward an OSGi-Based Infrastructure for Context-Aware Applications. IEEE Pervasive Computing, vol. 3, no. 4, 66-74. (2004)
12. Retkowitz, D., Pienkos, M.: Ontology-based Configuration of Adaptive Smart Homes. Proceedings of the 7th Workshop on Reflective and Adaptive Middleware (ARM'08), ACM 11-16. (2008)
13. Redondo, R. P. D., Vilas, A. F., Cabrer, M. R., Arias, J. J. P., Lopez, M. R.: Enhancing Residential Gateways : OSGi Services Composition. IEEE Transaction Consumer Electronics, vol. 53, no. 1, 87-95. (2007)
14. Diaz Redondo, R. P., Vilas, A. F., Cabrer, M. R., Arias, J. J. P., Duque, J. G., Solla, A. G.: Enhancing Residential Gateways: A Semantic OSGi Platform. IEEE Intelligent Systems, vol. 23, 32 – 40. (2008)
15. Chen, H, Perich, F., Finin, T., Joshi, A.: SOUPA: Standard Ontology for Ubiquitous and Pervasive Applications. Mobile and Ubiquitous Systems: Networking and Services, 258-267. (2004)
16. Heckmann, D.: Integrating Privacy Aspects into Ubiquitous Computing: A Basic User Interface for Personalization presented at the Artificial Intelligence in Mobile System Workshop at Ubicomp (AIMS). http://w5.cs.uni-sb.de/~krueger/aims2003/camera-ready/heckmann-14.pdf. (2003)
17. Heckmann, Dominik., Schwartz, Tim., Brandherm, Boris., Schmitz, Michael., Margeritta Wilamowitz-Moellendorff.: GUMO-The General User Model Ontology. User Modeling, LNCS 3538, Springer, 428-432. (2005)
18. Rosenman, M. A., Gero, J. S.: Purpose and function in design: from the socio-cultural to the techno-physical. Design Studies, 19, 161-186. (1998)
19. Gero, J. S., Kannengiesser, U.: The Situated Function-Behaviour-Structure Framework. In Proceeding of Artificial Intelligence in Design, 89-104. (2002)
20. Maslow, A.: Motivation and personality. New York: Harper. (1954)
21. Maslow, A.: The farther reaches of human nature. New York: The Viking Press. (1971)
22. Maslow, A., Lowery, R.: Toward a psychology of being (3rd ed.). New York: Wiley & Sons. (1998)

23. Maslow, A.: A theory of human motivation. Psychological Review, 50, 370-396. Retrieved June 2001, from http://psychclassics.yorku.ca/Maslow/motivation.htm. (1943).
24. Jifeng, X., Hanning Z.: The analyzing on design model of modern Chinese furniture based on Maslow's hierarchy of needs. Computer-Aided Industrial Design & Conceptual Design, 2009. CAID & CD 2009. IEEE 10th International Conference, 175 – 178. (2009)
25. Robert U., Frolick, M. N.: The IT Value Hierarchy: Using Maslow's Hierarchy of Needs as a Metaphor for Gauging the Maturity Level of Information Technology Use within Competitive Organizations. Information Systems Management, (Feb. 2008)
26. Le Guen Herve, Moga Sorin: A model of cooperative agent based on imitation and Maslow's pyramid of needs. Proceedings of the 2009 international joint conference on Neural Networks, (Jun. 2009)
27. Richard J., Holden, Ben-Tzion Karsh: A theoretical model of health information technology usage behavior with implications for patient safety. Behavior & Information Technology, (Jan. 2009)
28. Fonsecaa, F., Martinb, J.: Space and time in eco-ontologies. AI Communications, 17(4) : 259～269. (2004)
29. Raskin, R.: Guide to SWEET ontologies. NASA/Jet Propulsion Lab, Pasadena, CA. (2006)
30. Bermudez, L., Graybeal, J., Arko, R.: A marine platforms ontology: Experiences and lessons. Citeseer. (2006)
31. Lee, C., Nordstedt, D., Helal, S.: Enabling Smart Spaces with OSGi. IEEE Pervasive Computing, vol. 2, no. 3, 89-94. (2003)
32. Dobrev, P., Famolari, D., Kurzke, C., Miller, B.A.: Device and Service Discovery in Home Networks with OSGi. IEEE Communications Magazine, vol. 40, no. 8, 86-92. (2002)
33. Chandrasekaran, B., Josephson, J. R.: Function in Device Representation. Engineering with Computers, 16(3/4):162-177. (2000)
34. Chittaro, L., Guida, G., Tasso, C., Toppano, E.: Functional and Teleological Knowledge in the Multi-Modeling Approach for Reasoning about Physical Systems: A Case Study in Diagnosis. IEEE Transactions on Systems, Man, and Cybernetics 23(6):1718-1751. (1993)
35. Lind, M.: Modeling Goals and Functions of Complex Industrial Plants. Applied artificial intelligence, 259-283. (1994)
36. Hubka, V., Eder, W.E.: Theory of Technical Systems. Berlin: Springer-Verlag, (1998)
37. Yoshinobu, K., Riichiro, M.: A Device-oriented Definition of Functions of Artifacts and Its Perspectives. Functions in Biological and Artificial Worlds: Comparative Philosophical Perspectives, Ulrich Krohs, Peter Kroes (eds.), MIT Press, pp. 203-221. (March 2009)
38. Riichiro, M., Yoshinobu, K.: A Functional Ontology of Artifacts. The Monist - An Int'l Quarterly J. of General Philosophical Inquiry, Vol. 92, No.3, 387-402. (July, 2009)
39. Kitamura, Y., Mizoguchi, R.: Ontology-based systematization of functional knowledge. Journal of Engineering Design 15(4): 327-351. (2004)
40. Kitamura, Y., Sano, T., Namba, K., Mizoguchi, R.: A Functional Concept Ontology and Its Application to Automatic Identification of Functional Structures. Advanced Engineering Informatics, 16(2):145-163. (2002)
41. Engineering and Physical Sciences Research Council, http://www.epsrc.ac.uk/ResearchFunding/Programmes/Materials/ResearchPortfolio/EngineeringFunctionalMaterials.htm

Moji Wei, born in 1981, received the B.E. and M.E. degrees in Information Management and Information System from Shandong Economic University, China, in 2005 and 2008 respectively, and the Ph.D. degree in Department of Computer Science from Ocean University of China in 2011. His current research interests are knowledge representation, ontology theory and application, smart home. (weimoji1981@126.com)

Jianliang Xu, born in 1969, received the B.E. and M.E. degrees in software engineering from Shan-Dong University, China, in 1991 and 1994 respectively, and the Ph.D. degree in system engineering from Yamaguchi University in 1998. He is presently a professor in the Department of Computer Science and Technology, Ocean University of China. His research interests include automata theory, computational complexity and semantic web. (xjl9898@gmail.com)

Hongyan Yun, born in 1971, Ph. D. candidate in Department of Computer Science, Ocean University of China. Associate professor of the College of Information Engineering, Qingdao University. She graduated from North West University (Xi’an, China) in 1993 with a Bachelor Degree in computer software, and then she earned a Master Degree in computer software and theory from North West University in 2000. Her current research interests are Semantic Web, knowledge representation and reasoning, ontology engineering and intelligent information system. (yunhongyan@gmail.com)

Linlin Xu, born in 1989, MS. candidate in Department of Computer Science, Ocean University of China. She graduated from Inner Mongolia Finance and Economics University in 2010 with a Bachelor Degree in computer science and technology. Her current research interests are Semantic Web and semantic-based web serivices discovery and compostion.(gxl0216 @gmail.com)

Received: December 22, 2011; Accepted: February 20, 2012.

DOI: 10.2298/CSIS110615005S

Automatic Generation of E-Courses Based on Explicit Representation of Instructional Design

Goran Savić, Milan Segedinac, and Zora Konjović

Faculty of Technical Sciences, Trg D. Obradovića 6,
21000 Novi Sad, Serbia
{savicg, milansegedinac, ftn_zora}@uns.ac.rs

Abstract. This paper presents the system for automatic generation of IMS LD compliant E-Course from three components: machine readable explicit representation of instructional design, ontology of learning goals, and IMS Content Packaging compliant learning resources. For the explicit representation of instructional design, we have created a new domain-specific language named ELIDL which is aimed primarily at assistance to implementation of various pedagogical approaches in the course. Using ELIDL a teacher defines instructional design template which is one of the input parameters for the course generation. The system is verified by generating examples of the six different instructional designs for the Web Programming e-course via templates written in ELIDL.

Keywords: e-learning, instructional design, DSL, automatic course generation, IMS Learning Design.

1. Introduction

When a learning environment is switched from classical face-to-face to e-learning, an unavoidable task is development of electronic courses which are compliant with e-learning standards and shaped according to appropriate instructional design model(s). There are two globally adopted specifications that provide such course description – SCORM [1] and IMS Learning Design [2]. Creation of a new course or modification of an existing course compliant with any of these two standards is, at least, a time-consuming process. In addition, the realistic scenarios very often are those when existing courses should be modified. Most available tools for e-course creation make this task unduly hard because the sequences of learning activities and learning resources are interwoven and represented as a monolithic unit. While changing the course in some segment, it is necessary to deal with this monolithic representation, although only one of these segments which very often is instructional design, is usually the subject of the change.

Course development efforts can be significantly reduced by using systems for automatic course generation. The focus of this paper is an electronic course description which gathers learning resources and learning activities. It presents a system for automatic course generation with a focus on implementation of different instructional design strategies.

The rest of the paper is organized as follows. Concrete motivation for developing the system is described in the *Motivation* Section. The Section *Related Work* analyzes other systems for automatic course generation and it is particularly concerned about the role and implementation of instructional design in these systems. Our system is in detail presented in Sections *System Components* and *System Architecture and Functioning*. We have used the system to automatically generate six electronic courses in Web Programming differing in instructional strategies. The results are described in *Case Study* Section. At the end, conclusions and future research directions are given.

2. Motivation

The first-hand motivation for the research presented in this paper is demand to switch relatively large number of courses on the Faculty of Technical Sciences in Novi Sad (Serbia) to blended learning environment in a short time. Additionally, every semester brings some changes in the existing courses. Very often these changes are related to the course instructional design. There are two main reasons for applying changes to course's instructional design: gradual improvement of an ongoing teaching process, and studying/creating different instructional techniques. For these reasons it is useful to have a system that enables creating/changing the course instructional design easily. Therefore, we have decided to create a system that automates the course creation process by using a formal, machine readable description of instructional design as a base. By using this system, accompanied with ontology of learning goals and IMS Content Packaging compliant learning resources, we can automatically generate an electronic course in IMS LD format.

3. Related Work

In this section we analyze other systems for automatic generation of e-learning courses. These systems using different inputs generate an e-learning course as an output. For a course creation, two aspects have to be considered: (1) technical aspect which is about system functionalities, input/output formats, standard compliance and other characteristics relevant from the technical point of view, and (2) pedagogical aspect which includes ability to apply different instructional principles in a generated course.

Paper [3] describes a system for automatic course generation. The system generates the sequence of learning objects in IMS Content Packaging format on the basis of learning goals and student's knowledge state. Planning mechanism and PDDL (Planning Domain Definition Language) language are used for the creation of the sequence. Hernandez et al. in [4] also use PDDL plan for a course generation in IMS Learning Design format. A concrete learning activity that will be presented to a learner is chosen in real-time depending on the student's profile and learning goals. An approach to course generation based on learning goals ontology where learning resources are mapped to ontology nodes is presented in [5]. Nodes are hierarchically ordered and mastering one learning goal may be a precondition for other learning goals. A student's model contains her/his learning preferences and cognitive state. Ontology data and the student's model are inputs for generating a learning path. A similar approach is used in [6], but there is no defined learning goals ontology. Relations among learning objects are defined directly in the set of objects. The result is an e-learning course compliant with IMS Learning Design specification. All mentioned systems generate a course in one of globally accepted formats, which is done in our system, too. Additionally, our system has borrowed the idea of presenting learning goals by ontology and mapping learning objects to this ontology. Still, in contrast to our system, there is no explicit representation of the instructional design in these papers. The role of PDDL is to define a sequence of learning activities which is similar to the role of the instructional design in an e-learning course. But, PDDL is intended for the general planning and it is not sufficiently expressive to describe a variety of instructional designs in the e-learning course.

The pedagogical aspect is considered in paper [7], which presents a course generation system based on learning resources, learning goals, learner's profile and instructional design. Instructional design is defined separately from learning resources which is similar to our approach. But, it is defined by concrete learning activities in the course, not as a general template like in our paper. In our paper instructional design is defined abstractly and can be applied to any course. Paper [8] describes a web portal for defining pedagogical scenario and graphical interface of an e-learning course. A teacher defines these two components by answering an online questionnaire. The pedagogical scenario is represented by an XML file in IMS Learning Design format and this file is used for automatic course generation. Although, a pedagogical scenario is formally described in a machine-readable format, there is only a predefined set of pedagogical scenarios available.

Since we have decided to create a new domain-specific language for representing instructional design, it is important to analyze other similar researches focused on the formal representation of instructional design and the development of a domain-specific language for this purpose. Instructional design may be implicitly defined in the formal definition of the e-learning course. IMS LD specification introduces an XML-based language that enables an implicit definition of instructional design in a course by specifying learning activities and its organization. Similarly, LAMS [9] uses a specific language for specifying learning activities used in the system. These two languages show how instructional design may be implicitly defined altogether with all other course components. For creating an explicit representation of instructional design (independently from the concrete learning activities) we have to consider a teaching process more abstractly. It means that it is necessary to identify and model specific patterns in instructional design that are widely used in courses. One of the important researches in this area is *Pedagogical Patterns* project [10]. The project collects standard pedagogical patterns used in the teaching process. By this, teachers may find the most appropriate pedagogical pattern for a concrete teaching situation. Although these patterns are very useful for teachers and pedagogues, they are not intended for computer interpretation. The patterns are represented by textual descriptions, so they are not machine-readable. The Pedagogical Patterns project is mostly focused on the classical face-to-face learning. An idea for creating templates that can be used in an e-learning environment is presented in [11]. A template presents a course structure and it is presented graphically. Such a representation is not machine-readable too. Paper [12] also introduces the concept of templates in an e-learning course. Template is defined as a set of learning resources that contains specific pedagogical principles and these templates are copied in their original sequence into the course web site. Such an approach doesn't make an explicit distinction between learning resources and instructional design. The instructional design is not defined as a distinct component, but rather interwoven with learning resources. In [13], an adaptive e-learning environment SAP LSO is improved by involving templates in the system. Templates define a micro-strategy – a rule for organizing atomic knowledge items into the larger learning objects. Our paper is focused on a macro-strategy – how a sequence of learning objects is organized. There are only two predefined macro strategies in [13]. It is not possible to create a new macro-strategy as in our system. The paper [14] presents an approach that formalizes the representation of instructional design using SMID – the semantic model of instructional design. The model contains data about learner's knowledge state, learning goals, learning resources and an instructional strategy. The instructional strategy is defined using if-then rules. The sequence of learning actions is generated on the basis of learner's knowledge state and metadata about learning goal and learning object. Actions are composed of Gagne's Nine Events of Instruction. Here again, there is only a predefined set of instructional strategies available. The paper [15] presents an approach for converting pedagogical patterns described in [10] from textual format to the machine-readable IMS LD language. The result of this conversion is not a concrete IMS LD course, but a template that represents a set of abstract learning activities. Still, IMS LD is not sufficient to describe a machine-readable instructional design template since the purpose of the IMS LD is to describe concrete learning process, not an abstract template. Another approach that formalizes the representation of pedagogical patterns is described in [16]. Authors present a structured description of a pedagogical pattern. Although the pattern description is structured, it still consists of textual field, so it is not machine-readable. A term "pedagogical pattern" is in their paper used for describing a set of learning activities that are performed by a student or teacher. This definition is quite similar to the definition of the “instructional design template” used in our paper.

By analyzing instructional design description in all these papers, we have derived two conclusions about instructional design treatment: (1) instructional design is not represented as a machine readable format and/or (2) expressiveness and flexibility are not sufficient for describing a variety of instructional techniques that may be used in an e-learning course (in most papers these techniques are predefined or the underlying planning mechanisms are of general type which cumbers defining a new instructional techniques).

The system described in our paper is aimed at an automatic course generation which allows combination of pedagogical and technical aspects. The pedagogical aspect is in our system implemented using a distinct component written in our domain-specific language for describing instructional design. The proposed language enables formal representation of instructional design in a machine-readable format, independently of concrete learning goals and used learning resources and poses expressiveness and flexibility sufficient for the specific domain – instructional design. The language principles and other details are described in Section 4.3.

4. System Components

Our approach is based on a curriculum model presented in [17]. The model is based on Tyler’s taxonomy and defines four components in a course:

- learning goals,
- learning resources,
- instructional design, and
- assessment strategy.

Within the context of learning process automation, first three components are of interest. Learning goals are defined by course curriculum and they are not subject of a frequent change. For learning resources, a need for change is slightly bigger; commonly, a teacher changes some learning resources, adds new tests, etc. Finally, instructional design is the most dynamic course component in practice. In a teaching process, there is a continuous need to apply different instructional strategies for the same learning goals and the learning resources as an attempt of improving the learning process.

For the sake of flexibility in a course modification, where a teacher commonly changes only some of course’s components, we chose to define learning goals, learning resources and instructional design as three distinct components. The system organized in this way enables application of different instructional strategies for achieving defined learning goals where each strategy is consist of a particular set of learning activities. Likewise, since learning resources are defined as a separate component, the system enables simple alteration of learning resources. The result is an automatically generated IMS Learning Design compliant course which applies defined pedagogical approach. Hereby, a course creation process is easier and faster and a teacher gets a flexible environment for a course preparation. Next to it, detachment of a learning goals/content from an instructional design provides for easier testing of various instructional strategies to find the best one, as well as for creation/analysis of instructional design strategies.

4.1. Learning Goals

Learning goals can be defined as learning outcomes that a learner has to achieve. While considering a formal description of learning goals we set two basic requirements:

- the structure must be sufficiently expressive to describe learning goals and appropriate relations among them, and
- description must be machine-readable to allow automatic processing of learning goals

Following these requirements we decided to use ontology for representing learning goals. Similar approach is used in [3], [5] and [18]. Each ontology node represents one learning goal. We have defined two types of relations among our ontology nodes:

- `is-part-of`. Since learning goals are hierarchically organized, a learning goal may contain subgoals, which are defined by this relation type.
- `is-precondition`. This relation type defines a condition between learning goals: a goal c1 is a precondition for a goal c2 if, from knowing that a student has achieved c2, we can infer that he/she has achieved c1.

An ontology node can be annotated with the `hierarchical level` annotation. Predefined hierarchical levels are: `course`, `chapter`, `lesson` and `topic`. This is an optional annotation. A goal may be defined only for the purpose of grouping other learning goals. The goal doesn't necessarily represent a lesson, topic or any other course item. Furthermore, nodes can be annotated with the `orderNumber` attribute. This attribute is also not mandatory and specifies the suggested order of learning goals when the goals share the same hierarchical level and there is no relation `is-precondition` among them. Figure 2 shows the part of the learning goals ontology for the Web programming course.

4.2. Learning Resources

In this paper we use the term *learning resource* to refer any textual, graphical or multimedia digital content that a learner consumes during a learning process. For learning resources, it is necessary to specify formats of digital content, metadata and packaging. Regarding the format of digital content, although there are numerous globally used electronic formats (PDF and Office documents, images, video files), recent trends are to create learning resources in a browser readable format. The reason is that most e-learning systems are web applications, so a learner uses the internet browser to interact with the system. There are different formal specifications for defining learning resources. One widely used is IMS Content Packaging (IMS CP) [19] specification. IMS CP organizes learning resources into packages. A package consists of learning resources and a manifest file. The manifest file in XML format specifies resources, their description, organization and other elements defined by IMS CP standard. Learning resources are described by metadata and IMS consortium suggests IEEE LOM [20] specification for this purpose. Considering global acceptance of IMS CP specification, we have decided to use it to define learning resources in our system.

Learning resource is always related to one or more learning goals. A student uses a learning resource to achieve a specific learning goal. If learning resource is a test, then it is used to evaluate student’s knowledge. Since learning resources and learning goals are separated in our system, it is necessary to define a component that links learning resources with learning goals. Our system contains an intermediate component that maps learning resources to the ontology of learning goals. Mapping is defined as an XML document. The XML schema of this XML document is shown in Figure 1.

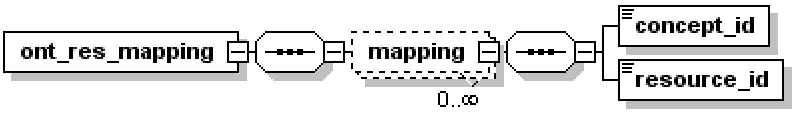


Fig. 1. The XML schema of the mapping document

Listing 1 shows a part of the mapping XML document for the Web programming course. One can notice that the learning resource “HTML tables” is related to the learning goal “Tables”.

```
<ont_res_mapping  ...>
...
   <mapping>
      <concept_id>Tables</concept_id>
      <resource_id>HTML_tables</resource_id>
   </mapping>
...
</ont_res_mapping>
```

Listing 1. XML document for mapping learning goals to learning resources in the Web programming course

An example of the learning goals ontology for the Web programming course and links among learning resources and learning goals are shown in Figure 2. Within the text some names are translated to English for the sake of clarity. This applies to all figures and listings in the paper.

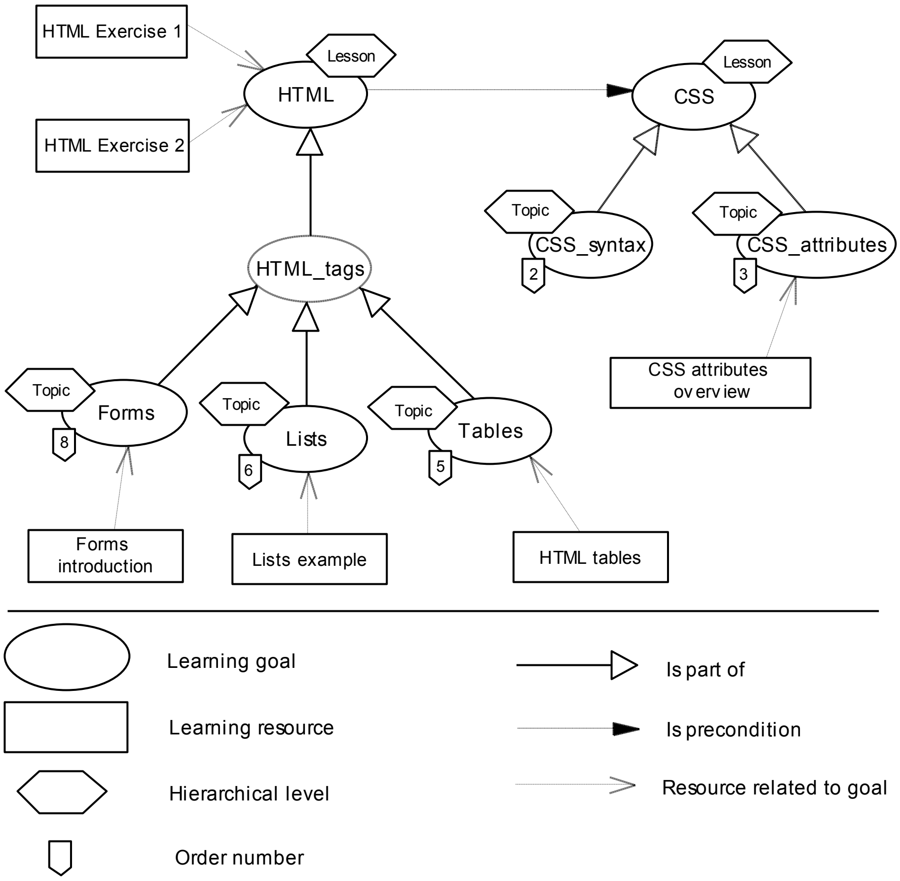


Fig. 2. Mapping of learning resources to the learning goals ontology

Using this structure, the ontology becomes an instrument that defines relations among learning resources. Thus, the relation is not defined as a part of the resource, but it is implicitly defined through the ontology. Using this approach it is easier to change the relation between two resources. In order to do that, a teacher should only map learning resource to another learning goal. Likewise, when adding a new resource, its relationships with other resources will be defined by simply mapping it to a specific learning goal(s).

4.3. E-Learning Instructional Design Language (ELIDL)

In our system, the idea is to represent instructional design detached from concrete learning goals and learning resources used for achieving these goals since the instructional design, in general, is not learning goals/resources specific [21]. Therefore, an instructional design applied to a course is represented by a detached component. The formal description of instructional design is implemented using templates that may be applied to any learning goals and learning resources. The templates are written in ELIDL (E-Learning Instructional Design Language) - our domain-specific language for describing instructional design.

ELIDL is developed following to the main principle that it should enable teacher to completely specify all components of instructional design in the e-learning environment. Thus, it is of interest to analyze how a teacher specifies instructional design in an e-learning course. Learning experience in an e-learning environment is always connected with consuming a specific digital learning resource using the computer. Learning resources can be different (a web page which explains a lesson, an internet forum for discussion with other students, an online test for knowledge evaluation, etc.), but from the technical point of view learning experience has always to do with a student's usage of some electronic resources. Therefore, instructional design in an e-learning environment is defined by the selection of appropriate learning resources and by ordering these resources. In other words, instructional design is specified by learning activities and paths through them. Given that learning activity in e-learning corresponds to the usage of some learning resource, in this paper we use the terms "learning resource" and "learning activity" interchangeably.

Willey in [22] examines instructional design for learning objects and he proposes a new instructional design theory – LODAS (Learning Object Design and Sequencing Theory). The author states that instructional design for learning objects is defined by two components:

- the scope and design of learning objects,
- the sequencing or combination of learning objects

Our research is not concerned with the scope and design of learning objects. In our work created resources are assumed as input parameters. Therefore, our language is aimed at defining the second component: the sequencing or combination of learning objects. Using ELIDL a teacher is able to formally specify instructional design that defines the order of learning activities and a criterion for selecting learning resources (number of resources, their type and priority). By this, instructional design in an e-learning course is explicitly and formally represented. Instructional design defined in ELIDL syntax is machine-readable, which is utilized here for the automatic generation of an e-learning course based on instructional design template written in ELIDL. After analyzing other languages [2, 7, 9, 15] from this domain we have decided that our language should also be XML-based.

The XML schema for ELIDL is shown in Fig. 3 and 4.

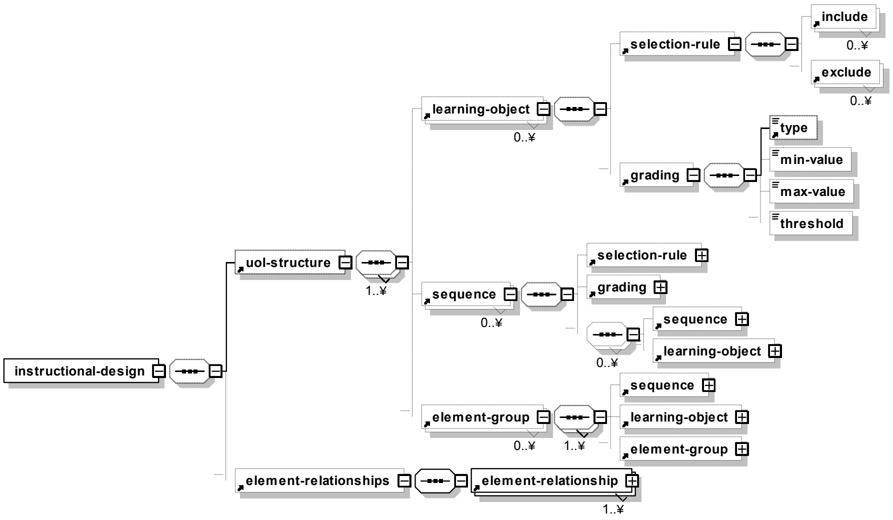


Fig. 3. The XML schema for an instructional design template – part 1

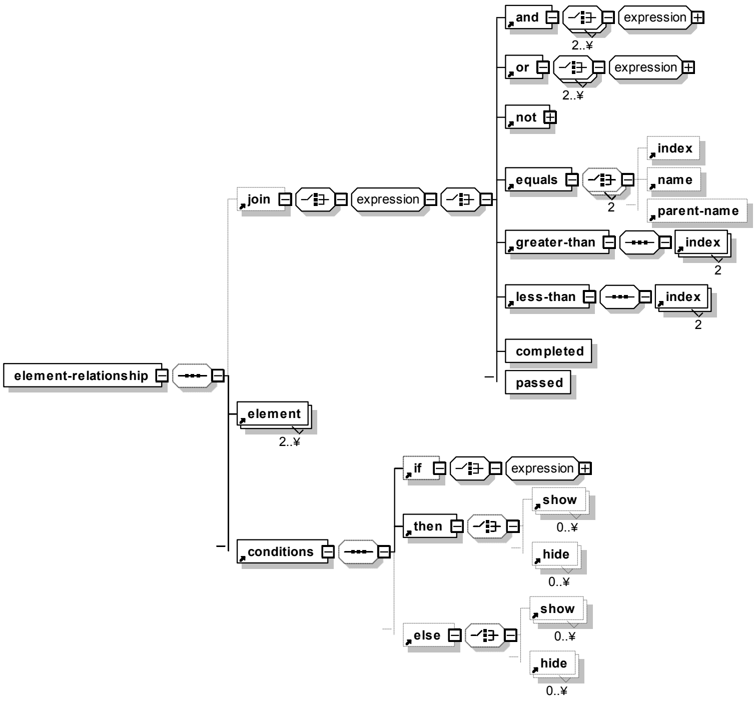


Fig. 4. The XML schema for an instructional design template – part 2

The most important schema elements are described in the following. Root element `instructional-design` contains two subelements:

– `uol-structure` - for defining learning activities in the course
– `element-relationships` – for defining relations among learning activities

Learning activities are defined by `sequence` or `learning-object` elements. The element `sequence` represents a sequence of other elements. The role of this element is similar to the role of “loop” statements in programming languages. The element `sequence` directs the system to pass through all learning goals annotated with hierarchical level defined in attribute `element`. On this way one can, for example, define the sequence of lessons, topics or learning objects. The element `learning-object` represents a concrete learning object in the course. Both `learning-object` and `sequence` elements contains the `selection-rule` element. This element specifies which learning resources and in which order will be selected from the set of learning resources mapped to a specific learning goal. The selection and priority of learning resources is defined by elements `include` and `exclude`, respectively. These elements are subelements of the `selection-rule` element. For the selection of a learning resource, it is necessary to define its metadata name and value. In this way our system provides the usage of any set of metadata for describing learning resources. The element `grading` specifies grading strategy for the learning object (the element is used for tests, projects, etc.). In order to facilitate managing learning elements, we can group them using ELIDL element `element-group`. The purpose of this element is to be a container for other elements.

The element `element-relationships` is a container for elements that define relations among learning activities in a course, e.g. there may be a strong relationship between a theory test and a project task. The project task is not available for students who didn’t pass the theory test. ELIDL defines such relation using `element-relationship` element. Its subelements named `element` specifies course elements that are parts of the relation. From all specified elements, relations will be formed only for the elements that satisfy a condition defined by the `join` element. The element `conditions` has a set of `if-then-else` elements which specify concrete actions that will be applied to the learning elements. Currently ELIDL supports actions for displaying and hiding learning resources (ELIDL elements `show` and `hide`).

The element `expression` groups elements that represent a logical expression. The elements `and`, `not` and `or` represent logical operators: conjunction, negation and disjunction, respectively. Logical operators for comparison are defined in the elements `equals`, `greater-than` and `less-than`. We can define the comparison of order numbers of the learning goals (ELIDL element `index`). Also, the `equals` element can compare string values. A string value that can be used as a comparison operand is the name of the learning goal to which a learning object is connected (ELIDL element `name`). Similarly, the element `parent-name` can be used to specify the name of the parent learning goal. Completing a learning activity or passing a test is evaluated using the ELIDL elements `completed` and `passed`, respectively.

The description of all schema elements is publicly available at www.informatika.ftn.uns.ac.rs/GoranSavic/IDTemplates.

The usage of our domain-specific language is illustrated by an example presented in Listing 2.

Listing 2 shows the part of an instructional design template for describing instructional strategy called "competency assessment". This strategy is described in [23] as follows: "*The learner is first presented with the introduction (lesson overview). He is then presented with an assessment that internally evaluates the learner's mastery of each of the module objectives. The learner is presented with the instructional material (modules) related to unsatisfied objectives. After the learner has completed all the required instructional materials, an exam is presented that re-tests the objectives the learner has not satisfied.*"

The XML document shown in Listing 2 has a root element `instructional-design-template` whose attribute `root` specifies the initial node in the ontology of learning goals. Thus, learning resources that will be presented to a learner are chosen from the learning objects mapped to the subgoals of the root goal. Concretely in this example, the initial goal is the one at the highest hierarchical level annotated with "course". The first `sequence` element defines the sequence of learning goals on the hierarchical level "lesson". For each goal (lesson), we choose a learning object annotated with the label "pre-test". Each test is graded with a passed/failed mark (boolean type of grading). The second `sequence` element specifies doing the lessons once again. This time, the sequence of lesson topics is represented by the learning goals annotated with the hierarchical level "topic". For each topic, the sequence of learning objects is created. The element `selection-rule` first chooses theoretical content, then examples and all other learning resources at the end. Finally, resources annotated with the "post-test" label are added to the course. This is defined by the last `sequence` element.

As noted, this instructional strategy specifies that a learner should learn only lessons related to the learning goals unsatisfied on the pre-test or the post-test. So, we have to define the relation between the lesson, pre-test and post-test using the `element-relationship` element. The element `join` specifies that relations are created only among lessons, pre-tests and post-tests which belongs to the same learning goal. Element `if` defines the condition that a learner has to satisfy to pass the pre-test or the post-test. If they have passed one of the tests, the `then` element specifies that the corresponding lesson shouldn't be available to the learner. Otherwise, the lesson will be available which is defined in the `else` element.

```
<instructional-design root = "course">
 <uol-structure>
  <sequence element="lesson" element-id="PRETEST_LESSON_SEQUENCE">
   <learning-object element-id="PRE_TEST">
    <selection-rule>
     <include att-name="label" att-value="pre-test"/>
    </selection-rule>
    <grading>
     <type>boolean</type>
    </grading>
   </learning-object>
  </sequence>
  <sequence element="lesson" element-id="L_SEQ"
           sequence-element-id="LESSON">
   <sequence element = "topic" element-id="T_SEQ">
    <sequence element = "learning-object">
     <selection-rule>
      <include att-name="type" att-value="explanation-content"
               priority="1"/>
      <include att-name="type" att-value="example" priority="2"/>
      <include att-name="type" att-value="*" priority="3"/>
     </selection-rule>
    </sequence>
   </sequence>
  </sequence>
  <sequence element = "lesson" element-id="PT_L_SEQ">
   <learning-object element-id="POST_TEST">
    <selection-rule>
     <include att-name="label" att-value="post-test"/>
    </selection-rule>
    <grading>
     <type>boolean</type>
    </grading>
   </learning-object>
  </sequence>
 </uol-structure>
<element-relationships>
  <element-relationship>
   <element element-ref="LESSON" alias="lesson"/>
   <element element-ref="PRE_TEST" alias="pretest"/>
   <element element-ref="POST_TEST" alias="posttest"/>
   <join>
    <and>
     <equals>
      <name element="lesson"/>
      <parent-name element="pretest"/>
     </equals>
     <equals>
      <name element="lesson"/>
      <parent-name element="posttest"/>
     </equals>
    </and>
   </join>
```

```
     <conditions>
      <if>
       <or>
        <passed element="pretest"/>
        <passed element="posttest"/>
       </or>
      </if>
      <then>
       <hide element="lesson"/>
      </then>
      <else>
       <show element="lesson"/>
      </else>
     </conditions>
    </element-relationship>
   </element-relationships>
  </instructional-design>
```

Listing. 2. Instructional design template for “competency assessment” instructional strategy

5. System Architecture and Functioning

This section presents system architecture and its functioning. System components and their relations are shown in the Fig. 5.

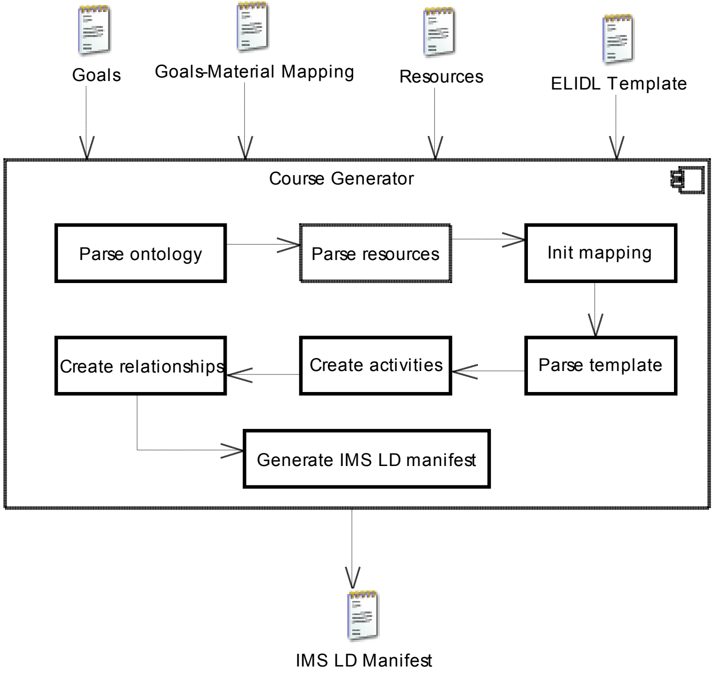


Fig. 5. System architecture

The files on the top of the figure are the input parameters to our system. The system is neutral with respect to how these files are created. In the Section 6 we have specified the software we used to create the input files. Ontology of learning goals can be created using some ontology editor, such as Protege. The goals-resources mapping is an XML file and can be created using any text or XML editor. Learning resources should be defined in the IMS CP format. There are some specialized applications, named IMS CP editors, which can facilitate this task. ELIDL template still has to be created manually using some text or XML editor, but our plan is to develop a graphical editor for creating ELIDL templates.

The system functioning is presented through the process of the course generation. Firstly, the system parses the OWL file that represents the ontology of learning goals. Then, the IMS CP manifest file which specifies learning resources is parsed. The system links resources with learning goals by parsing the XML file which defines the mapping of learning resources to the learning goals ontology. Then, the system parses ELIDL instructional design template. On the basis of the template, the system traverses the learning goals. For each learning goal, specified learning resources are taken. Then, the hierarchy of learning activities with relations among them is created. Each relation contains `if`, `then` and `else` section created according to the content of the `element-relationship` element in the template. Finally, the IMS LD compliant unit of learning is generated.

IMS LD unit of learning is generated by converting learning activities and their relations to the corresponding tags of the IMS LD manifest file. IMS LD specifies users and their roles in the manifest element `roles`. Our system creates two types of roles - `learner` for learners and `staff` for teachers within the `roles` element in the manifest file. The tag `play` is in IMS LD the root element when interpreting the learning design and it represents the flow of activities during the learning process. So, during the course generation, learning activity on the highest hierarchy level is mapped to the tag `play` in the manifest file. In IMS LD, play consists of `act` elements. In our system, tags `act` are generated on the basis of the activities on the second hierarchy level. On the third hierarchy level in IMS LD are `role-part` elements. Our system generates `activity-structure` tags for learning activities on the third hierarchy level. For each of these activities the system generates the `role-part` tag which references an appropriate `activity-structure` tag. Finally, `learning-activity` tags are created from the activities on the lowest hierarchical level, because in IMS LD `learning-activity` elements are on the lowest hierarchical level.

The flow of a learning process depends on relations among learning activities. So, relations among activities are converted to the `conditions` section of the IMS LD manifest file. Logical operators *and*, *or* and *not* are mapped to the tags with the same name in the IMS LD manifest file. ELIDL operator `completed` defines a certain action that will be performed after the completion of a learning activity. In IMS LD an activity contains the tag `on-completion` for this purpose. In this tag we can change the value of some property using the tag `change-property-value`. Properties in IMS LD are defined using tags `locpers-property`. Depending on the property value, IMS LD allows performing specific action in the `conditions` section of the manifest. So, ELIDL operator `completed` causes adding a new `locpers-property` tag within the `properties` element of the manifest. Likewise, sub tags `on-completion` and `change-property-value` are added to the `learning-activity` tags. The purpose of these elements is to define that the value of the `locpers-property` attribute will be changed if the learning activity is completed. At the end, for the defined property, the tag `greater-than` is added to the `conditions` section. Listing 3 shows a segment of the IMS LD manifest file that is generated on the basis of the template operator `completed`.

```
<imsld:locpers-property identifier="HTML_and_Java-Is-Done">
 <imsld:datatype datatype="integer"/>
 <imsld:initial-value>0</imsld:initial-value>
</imsld:locpers-property>
…
<imsld:complete-activity>
 <imsld:user-choice/>
</imsld:complete-activity>
<imsld:on-completion>
 <imsld:change-property-value>
  <imsld:property-ref ref="HTML_and_Java-Is-Done"/>
  <imsld:property-value>6</imsld:property-value>
 </imsld:change-property-value>
</imsld:on-completion>
…
<imsld:if>
 <imsld:greater-than>
  <imsld:property-ref ref="HTML_and_Java-Is-Done"/>
  <imsld:property-value>5</imsld:property-value>
 </imsld:greater-than>
</imsld:if>
```

Listing. 3. Generated segment of the manifest file for the template operator `completed`

`HTML_and_Java-is-Done` represents the property whose value will be changed on the completion of a learning activity. The initial value is 0 (defined in the `initial-value` tag). After the completion of the activity, the value is changed to 6 (in the `change-property-value` tag). A specific action will be performed if the value is greater than 5, which is defined in the `greater-than` tag. Hence, the action will be executed after the completion of the activity.

We are now going to consider the ELIDL operator `passed`. A student passes the test if he/she gets a passing grade. Therefore, it is necessary to define the grade for each operator `passed`. A new tag `locpers-property` that represents the grade is created in the `properties` section of the manifest file. In order to grade a learning activity, it is necessary to add a new web page into the unit of learning. The page has an input field for entering the grade. This page is automatically created and added to the list of learning resources. The page is created on the basis of the template grading page, part of which is shown in the Listing 4.

```
<ld:set-property ref="${PROPERTY_NAME}" property-of="self"
                view="value"/>
```

Listing. 4. Template input field for entering the grade in a grading web page

Before the grading page is added, the template field `${PROPERTY_NAME}` is substituted with the name of a concrete attribute that represents the grade. A new `greater-than` tag is added to the `conditions` sections indicating whether a student has passed the learning activitiy. Listing 5 shows the segment of the manifest file generated for the template operator `passed`.

```
<imsld:locpers-property identifier="Html_1_grading-Grade">
 <imsld:datatype datatype="boolean"/>
 <imsld:initial-value>false</imsld:initial-value>
  </imsld:locpers-property>
...
<imsld:if>
...
  <imsld:is>
   <imsld:property-ref ref="Html_1_grading-Grade"/>
   <imsld:property-value>true</imsld:property-value>
  </imsld:is>
...
</imsld:if>
```

Listing. 5. Generated segment of the manifest file for the template operator `passed`

The `locpers-property` named `Html_1_grading-Grade` represents the grade. The shown activity is graded using a boolean value – passed/not passed. In the `if` tag we evaluate this value in order to perform an action. Hence grading is done by the teacher, grading activity is defined by the tag `support-activity` in the manifest.

Regarding `if` and `then` tags in an instructional design template, the system converts them to `if` and `then` manifest tags, respectively. ELIDL actions `show` and `hide` are mapped to the manifest tags with the same name and they are added to the `conditions` section. These two tags represent actions for displaying and hiding a learning activity, respectively.

Finally, generated manifest file with all learning resources is packed into a ZIP file and this file represents the unit of learning compliant with IMS LD standard.

The system is implemented in Java programming language. Given that all input parameters are XML files, we have used Java DOM parser for parsing these files. Other technical details about the system can be found in [24].

6. Case Study

Case study shows automatic creation of the Web Programming course using our system. According to defined course curriculum, we created the ontology of learning goals in OWL [25] language using ontology editor Protege 3.4.1. Learning resources for the course are converted to HTML web pages, packed into the IMS CP package and each resource is mapped to the corresponding learning goal. In order to improve the quality of the selection of learning resources, the resources are annotated with metadata. Firstly, we describe the type of a learning resource with metadata `learningResourceType` defined by IEEE LOM specification. Values for this element are chosen from extended vocabulary defined in CLEO [26] specification which is created as an extension to the IEEE LOM. Furthermore, resources are annotated with the IEEE LOM metadata `description` that closely describes the resource.

We have created 6 instructional design templates that formally describe different instructional strategies. The templates are created for the following instructional strategies: *No sequencing*, *Linear*, *Knowledge Paced*, *Remediation*, *Competency Assessment* (strategies description adopted from [23]) and *Project-based learning* (description adopted from [27]). All templates written in ELIDL and accompanied with brief descriptions are publicly available at www.informatika.ftn.uns.ac.rs/ GoranSavic/IDTemplates/ Examples. Using these templates as inputs to our system, 6 courses of Web Programming in IMS LD format are automatically generated. These units of learning are also publicly available at www.informatika.ftn.uns.ac.rs / GoranSavic/IDTemplates/ Examples.

As an illustration, we are going to present two generated courses: for linear and knowledge paced instructional strategy. Linear strategy is described in [23] as follows "... *the learner must progress through the contents in a pre-determined order. The learner will start with the introduction first, then do all the modules and lessons in a linear order, directed by the LMS. The learner cannot proceed forward with the lessons until he has completed the current lesson. Each module is complete when he has finished all lessons in the module. The student will be presented with a comprehensive exam or quiz after he has completed all the modules*." Linear strategy for the Web programming course is presented graphically in Fig. 6.

One can notice that a student has to learn the lesson HTML firstly. After completing this lesson, he/she can learn CSS. After it, the lesson JavaScript becomes available and so on.

Using ELIDL we have written an instructional design template that formally describes this instructional strategy. Listing 6 presents a fragment of the template.

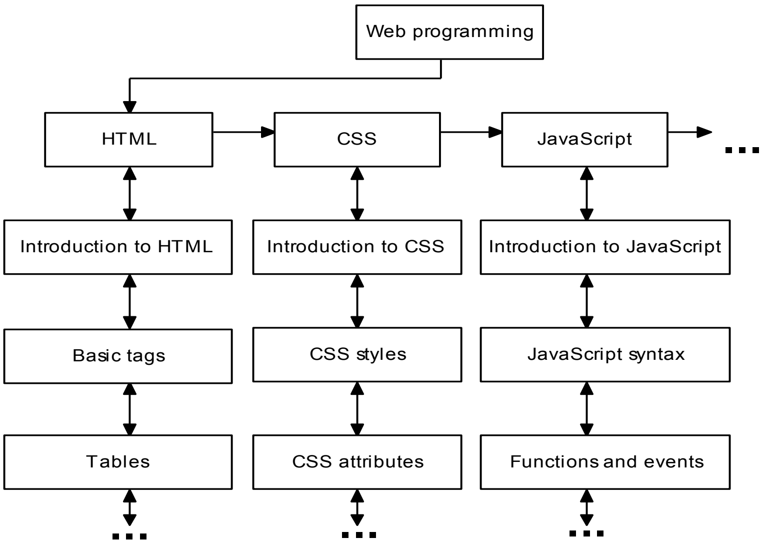


Fig. 6. Linear instructional strategy for the Web programming course

```
<uol-structure>
 <sequence element = "lesson" element-id="L_SEQ"
   sequence-element-id="LESSON">
  <sequence element = "topic" element-id = "T_SEQ"
    sequence-element-id="TOPIC">
   <sequence element = "learning-object">
    <selection-rule>
...
     <include att-name="type" att-value="*" priority="3"/>
    </selection-rule>
   </sequence>
...
<element-relationships>
   <element-relationship>
      <element element-ref="LESSON" alias="l1"/>
      <element element-ref="LESSON" alias="l2"/>
      <join>
         <less-than>
            <index element="l1"/>
            <index element="l2"/>
         </less-than>
      </join>
      <conditions>
          <if>
             <not> <completed element="l1"/> </not>
          </if>
          <then>
             <hide element="l2"/>
          </then>
          <else>
             <show element="l2"/>
          </else>
...
```

Listing. 6. ELIDL template for “Linear” instructional strategy

A student gets all learning resources in linear order. The tag include defines the selection of all types of learning resources. This is defined using the attributes att-name and att-value. The character * in the att-value attribute indicates the selection of all learning resources.

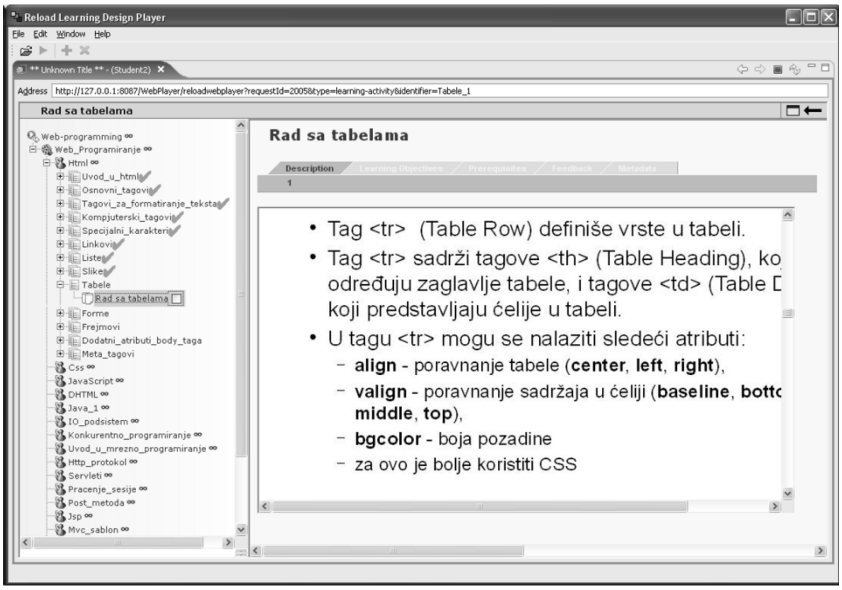
The tag conditions in the template document defines when the learning resource should be shown to the student. The lesson is available only when the previous lesson has been completed. Using this ELIDL template, our system has generated an IMS LD manifest file that is shown in Listing 7.

```
<imsld:learning-activity identifier="Tables_1">
 <imsld:title>Working with tables</imsld:title>
 <imsld:activity-description>
  <imsld:item identifierref="tables_res">
   <imsld:title>Working with tables</imsld:title>
  </imsld:item>
 </imsld:activity-description>
 <imsld:complete-activity>
  <imsld:user-choice/>
 </imsld:complete-activity>
 <imsld:on-completion>
  <imsld:change-property-value>
   <imsld:property-ref ref="Tables_1-Is-Done"/>
   <imsld:property-value>6</imsld:property-value>
  </imsld:change-property-value>
 </imsld:on-completion>
</imsld:learning-activity>
...
<imsld:if>
...
 <imsld:not>
  <imsld:greater-than>
   <imsld:property-ref ref="Tables_1-Is-Done"/>
   <imsld:property-value>5</imsld:property-value>
  </imsld:greater-than>
 </imsld:not>
...
</imsld:if>
<imsld:then>
 <imsld:hide>
  <imsld:activity-structure-ref ref="as_CSS_attributes"/>
 </imsld:hide>
</imsld:then>
```

Listing. 7. IMS LD manifest file generated for “Linear” instructional strategy

Listing shows a generated learning activity for learning HTML tables (learning activity with the identifier Tables_1). Tag if defines that the topic “CSS attributes” (activity structure as_CSS_attributes) is not available if a student hasn’t completed learning of HTML tables. The property Tables_1-Is-Done indicates whether the learning of the HTML tables is completed.

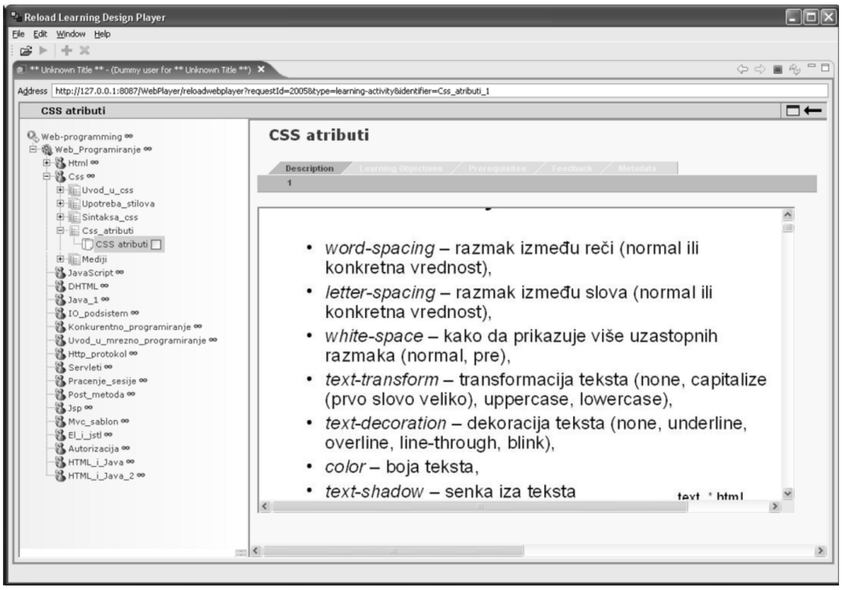
Figures 7 and 8 present screenshots of the Web programming course generated using ELIDL template for “linear” instructional strategy. The screenshots are taken from the Reload LD Player [28].



Rad sa tabelama

- Tag <tr> (Table Row) definiše vrste u tabeli.
- Tag <tr> sadrži tagove <th> (Table Heading), ko određuju zaglavlje tabele, i tagove <td> (Table D koji predstavljaju ćelije u tabeli.
- U tagu <tr> mogu se nalaziti sledeći atributi:
  - **align** - poravnanje tabele (**center**, **left**, **right**),
  - **valign** - poravnanje sadržaja u ćeliji (**baseline**, **bottc middle**, **top**),
  - **bgcolor** - boja pozadine
  - za ovo je bolje koristiti CSS

Fig. 7. HTML lesson in the Web programming course generated for linear instructional strategy



CSS atributi

- *word-spacing* – razmak između reči (normal ili konkretna vrednost),
- *letter-spacing* – razmak između slova (normal ili konkretna vrednost),
- *white-space* – kako da prikazuje više uzastopnih razmaka (normal, pre),
- *text-transform* – transformacija teksta (none, capitalize (prvo slovo veliko), uppercase, lowercase),
- *text-decoration* – dekoracija teksta (none, underline, overline, line-through, blink),
- *color* – boja teksta,
- *text-shadow* – senka iza teksta

Fig. 8. CSS lesson in the Web programming course generated for linear instructional strategy

The screenshots show the course before and after the learning of HTML is completed. We can notice from the Figure 7 that the item “CSS” cannot be expanded and its subitems are not displayed to the user. The reason is that the lesson “CSS” is not available if the lesson “HTML” hasn’t completed. The lesson HTML is completed when all its topics are completed (as shown in the Figure 7, completed topics are checked). The topics can be completed in any order. When the user completes all topics in the lesson HTML, the lesson CSS becomes available. As we can see from the Figure 8, CSS’s topics are now displayed to the user.

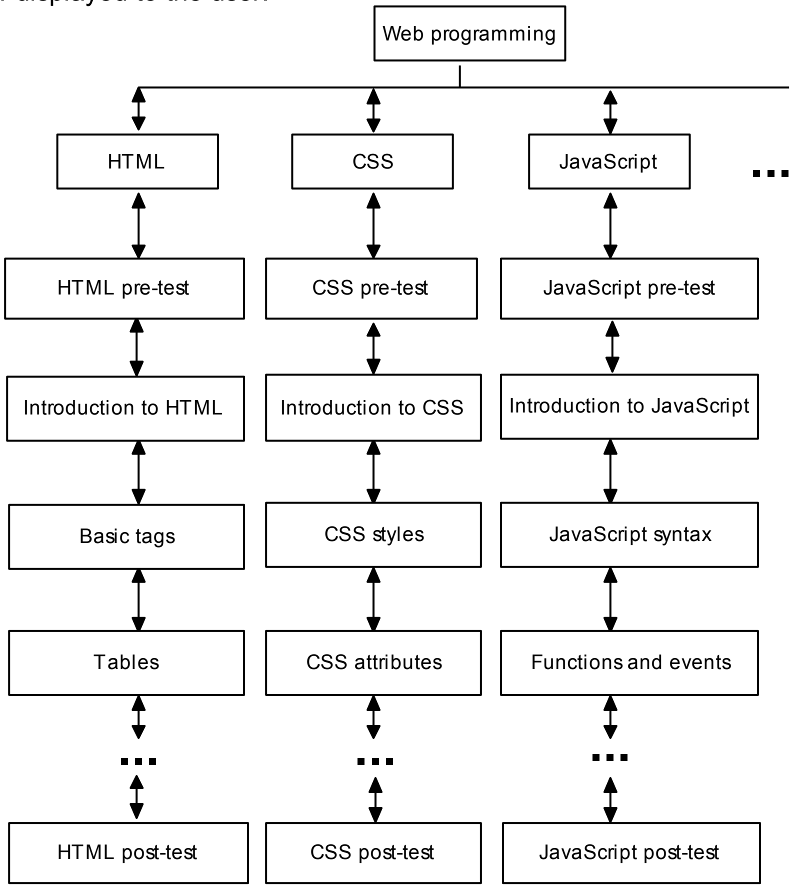


Fig. 9. “Knowledge Paced” instructional strategy for the Web programming course

The second course that will be presented here is the one generated for “Knowledge Paced” instructional strategy. This strategy is described in [23] as follows: “*In this mode, the learner must go through and complete the introduction first. After that he may proceed to the module 1 pre-test, select another module pre-test, or select a lesson. The learner may "jump" between modules, selecting pre-tests or lessons in any order. The learner cannot select the Module post-tests. These are only encountered after the learner "flows" through the modules lessons. If the learner passes an exam (pre- or post-), the module’s learning objective has been satisfied and the module’s* *post-test becomes disabled. The learner may continue to select individual lessons for the duration of the course, even after a module's objective has been satisfied. If the learner does not pass the exam, the learner is directed to that module's instructional content, and once completed, must retake the module exam (post-test).*" Figure 9 shows "Knowledge Paced" instructional strategy for the Web programming course.

In contrast to "linear" instructional strategy, in "knowledge paced" strategy a student is not limited only to linear path. He/she may arbitrary choose next learning unit. A learning unit is satisfied if a student passes the unit pre-test or post-test. Listing 8 shows ELIDL template created for this instructional strategy.

```
<uol-structure>
...
 <sequence element = "lesson" element-id="L_SEQ">
  <learning-object element-id="PRE_TEST">
   <selection-rule>
    <include att-name="label" att-value="pre-test"/>
   </selection-rule>
...
  <sequence element = "topic" element-id="T_SEQ">
   <sequence element = "learning-object">
    <include att-name="type" att-value="*" priority="3"/>
...
  <learning-object element-id = "POST_TEST">
   <selection-rule>
    <include att-name="label" att-value="post-test"/>
   </selection-rule>
...
<element-relationships>
 <element-relationship>
  <element element-ref="PRE_TEST" alias="pre"/>
  <element element-ref="POST_TEST" alias="post"/>
  <join>
   <equals>
    <parent-name element="pre"/>
    <parent-name element="post"/>
   </equals>
  </join>
  <conditions>
   <if>
    <or>
     <passed element="pre"/>
     <passed element="post"/>
    </or>
   </if>
   <then>
    <hide element="pre"/>
    <hide element="post"/>
   </then>
   <else>
    <show element="pre"/>
    <show element="post"/>
   </else>
  </conditions>
 </element-relationship>
</element-relationships>
```

Listing 8. ELIDL template for "Knowledge Paced" instructional strategy

As we can see, the template iterates through all lessons (ELIDL element sequence). For each lesson it firstly selects a pre-test using the tag include and its attribute att-value whose value is “pre-test”. After the pre-test, in the second sequence element, the template iterates through lesson topics and their learning resources. It selects all learning resources in the lesson (the attribute att-value in the tag include has the value “*”). At the end, a user gets a post-test in a similar way as for the pre-test.

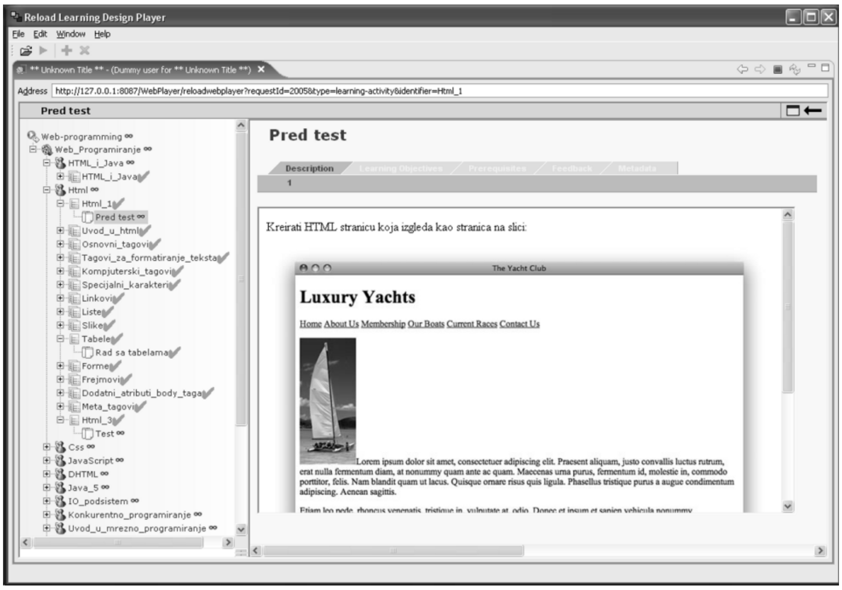
The tag element-relationship defines a relation between pre-test and post-test. If a student passes pre-test or post-test (defined in passed ELIDL tags), both tests become unavailable (ELIDL elements hide). Otherwise, the tests are available which is defined in the ELIDL elements show. Listing 9 shows a part of the generated IMS LD manifest file for “Knowledge Paced” instructional strategy.

```
<imsld:learning-activity identifier="Html_3">
 <imsld:title>Post test</imsld:title>
 <imsld:activity-description>
  <imsld:item identifierref="html_posttest_res">
   <imsld:title>Post test</imsld:title>
  </imsld:item>
 </imsld:activity-description>
</imsld:learning-activity>
...
<imsld:activity-structure identifier="as_Html_3"
                         structure-type="sequence">
 <imsld:title>Html_3</imsld:title>
 <imsld:learning-activity-ref ref="Html_3"/>
</imsld:activity-structure>
...
<imsld:if>
...
 <imsld:or>
  <imsld:is>
   <imsld:property-ref ref=" Html_1_grading-Grade"/>
   <imsld:property-value>true</imsld:property-value>
  </imsld:is>
  <imsld:is>
   <imsld:property-ref ref=" Html_3_grading-Grade"/>
   <imsld:property-value>true</imsld:property-value>
  </imsld:is>
 </imsld:or>
...
</imsld:if>
<imsld:then>
 <imsld:hide>
  <imsld:activity-structure-ref ref=" as_Html_3"/>
 </imsld:hide>
</imsld:then>
```

Listing. 9. IMS LD manifest file generated for “Knowledge Paced” strategy

A post-test for lesson HTML is represented with a learning activity “Html_3” and a corresponding activity-structure element. The post-test should be hidden if a student has passed the pre-test or post-test. The attributes “Html_1_grading-Grade” and “Html_3_grading-Grade” represent grades on the pre-test and post-test, respectively. The test grade is a logical value –

true/false (passed/failed). When the grades values are `true` (defined in the IMS LD tag `if`), the post-test is hidden (defined using IMSL LD element `hide`). Figure 10 shows the screenshot of the Web Programming course generated from “Knowledge Paced” ELIDL template. The screenshot is taken before the pre-test for the lesson “HTML” is passed. We observe appearance of the items Html_1 and Html_3 inside the lesson HTML. These items represent pre-test and post-test, respectively. Fig. 11 shows the screenshot of the same course after the pre-test is passed. One can notice that the Html_1 and Html_3 items do not appear in the lesson, because the pre-test and post-test are hidden now.



Address http://127.0.0.1:8087/WebPlayer/reloadwebplayer?requestId=2005&type=learning-activity&identifier=Html_1

Kreirati HTML stranicu koja izgleda kao stranica na slici:

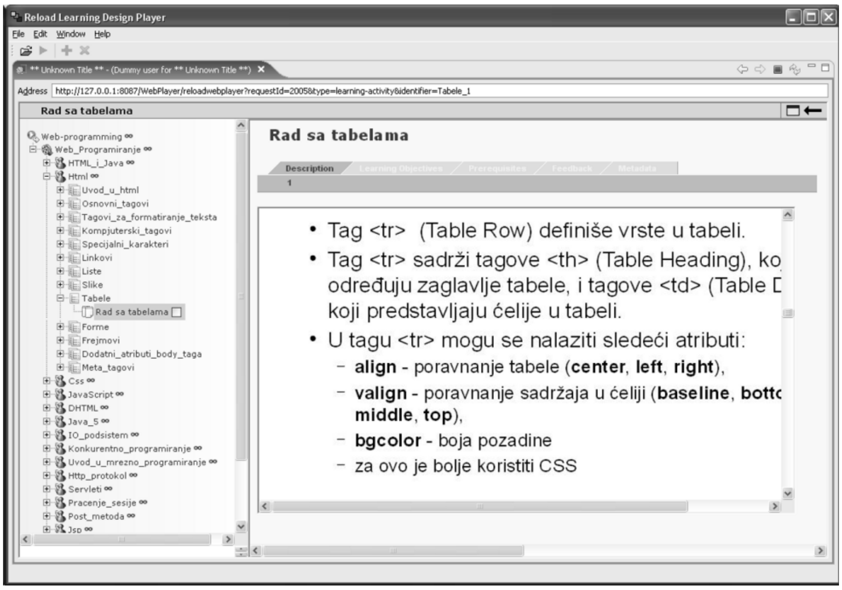
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Praesent aliquam, justo convallis luctus rutrum, erat nulla fermentum diam, at nonummy quam ante ac quam. Maecenas urna purus, fermentum id, molestie in, commodo porttitor, felis. Nam blandit quam ut lacus. Quisque ornare risus quis ligula. Phasellus tristique purus a augue condimentum adipiscing. Aenean sagittis.

Etiam leo pede, rhoncus venenatis, tristique in, vulputate at, odio. Donec et ipsum et sapien vehicula nonummy.

Fig. 10. HTML lesson in the Web programming course generated for “Knowledge Paced” instructional strategy (before the pre-test is passed)

All 6 generated units of learning for Web Programming course are analyzed and the results are presented in Table 1. For each unit of learning, the table shows applied instructional strategy and the following characteristics of the manifest file: number of learning activities, relations, `locpers-property` tags and lines of XML code.

Three of the six instructional strategies are described earlier in the text: *Competency Assessment, Linear and Knowledge Paced*. Before the analysis of the generated courses, we are going to describe briefly three other strategies. *No-sequencing* is the simplest strategy and it allows unrestricted access to all learning resources in any order. *Remediation* strategy is the combination of the *Linear and Knowledge Paced strategy.* A learner has to learn lessons in linear order. For each lesson there is a post-test (there is no pre-tests as in case of *Knowledge Paced* strategy). If a student passes the post-test, the lesson is completed and he/she doesn't need to learn the lesson any more. *Project-based learning* strategy is directed to create a specific product as the result of the learning process. A student is free to explore all learning content in any order without restrictions. During the course there are milestones for evaluating the progress of the project. Each milestone defines a project-part which is evaluated.



Rad sa tabelama

- Tag <tr> (Table Row) definiše vrste u tabeli.
- Tag <tr> sadrži tagove <th> (Table Heading), ko
  određuju zaglavlje tabele, i tagove <td> (Table D
  koji predstavljaju ćelije u tabeli.
- U tagu <tr> mogu se nalaziti sledeći atributi:
  - **align** - poravnanje tabele (**center**, **left**, **right**),
  - **valign** - poravnanje sadržaja u ćeliji (**baseline**, **botto**
    middle, **top**),
  - **bgcolor** - boja pozadine
  - za ovo je bolje koristiti CSS

Fig. 11. HTML lesson in the Web programming course generated for "Knowledge Paced" instructional strategy (after the pre-test is passed)

Table 1. Comparison of the generated units of learning

The strategy applied	# of				Processing time
	Activi ties	Relations	locpers-property	Manifest lines	
No-sequencing	85	0	2	1942	2 s
Linear	84	138	83	47109	219 s
Knowledge Paced	131	210	95	7663	6 s
Remediation	107	186	94	58332	456 s
Competency Assessment	131	210	104	19673	22 s
Project-based learning	88	170	7	5305	4 s

The strategies *No-sequencing* and *Linear* are characterized by a small number of learning activities because they contain only learning the lessons and work on a project. Since the students have to take pre-tests and post-tests, *Knowledge-paced* and *Competency Assessment* strategies have the largest number of activities. Regarding the number of relations, the *No-sequencing* strategy has no relation at all, as it means a completely free learning path. The largest numbers of relations have *Knowledge Paced* and *Competency Assessment* strategies since they have the largest number of activities and these activities are conditioned by each other (availability of some activities depends on the result of completed and passed actions of other activities). Elements `locpers-property` comply with the number of relations among learning activities. *Competency Assessment* and *Knowledge Paced* strategies have the same number of relations, but for *Competency Assessment* strategy three activities participate in some relations, while for *Knowledge Paced* strategy all relations are defined between only two activities. Therefore, the generated manifest file has more `locpers-property` elements for the *Competency Assessment* instructional strategy. The *No-sequencing* strategy doesn't have any relations, so it has only two `locpers-property` elements. These two elements represent the student's grades on the projects and they are the consequences of the grading element in the instructional design template. The least lines of XML code is generated for *No-sequencing* strategy, since this strategy has a small number of activities and there are no relations in it. For this strategy, the instructional design template doesn't define any condition element at all. Manifests for the *Linear* and *Remediation* strategies have most lines of code, since these strategies require linear path through lessons. This means that a lesson is not available until the previous one is completed which causes many `if-then` elements in the manifest. In addition, the *Remediation* strategy contains tests, which produces a manifest file containing most lines of code. Regarding the processing time, it is proportional to the number of the generated lines in the manifest. *No-sequencing* strategy requires least time, while the most time the system spends to generate a course based on the *Remediation* strategy.

An automatically created course can be imported into some IMS LD Player. For our purposes, we have used Reload LD Player [28] and CopperCore IMS Learning Design Engine [29].

7. Conclusion

The paper presents the system for automatic generation of IMS LD compliant courses based on explicitly expressed machine-readable instructional design templates. The courses are generated from three inputs: the ontology of learning goals, learning resources and instructional design template. The output result is an IMS LD unit of learning that applies defined instructional design.

Instructional design templates provide the means for explicit and formal specification of instructional design, which most of existing systems for the automatic course generation do not support. Detaching instructional design eases change of an applied instructional strategy in the course. For the purpose of formal templates' specification and representation, we have developed a new domain-specific language named ELIDL.

It should be mentioned that the generated course is a static one – it represents the sequence of predefined learning activities. The quality of the teaching process and utilization of an e-learning environment could be improved if a learning activity is chosen in real-time. These would enable that a student gets a learning activity dynamically depending on various parameters (instructional design, student's knowledge state, personal preferences etc.). However, given that most current LMS's support only static courses, in this phase of the research we decided to generate only a static sequence of learning activities.

The system is verified on the example of generating Web Programming course at the Faculty of Technical Sciences in Novi Sad. For the defined learning goals and learning resources, we created 6 instructional design templates that describe following instructional strategies: *No sequencing*, *Linear*, *Knowledge Paced*, *Remediation*, *Competency Assessment* and *Project-based learning*. As a result, the system generated 6 Web Programming courses compliant with IMS LD specification. All templates and generated courses are publicly available.

Our future work is mainly concerned with using the generated courses in a teaching process. Currently we are measuring students' achievements and motivation in order to evaluate different instructional strategies and to determine which one is the best for the given course.

Another direction of the further research is oriented towards development of the software tool aimed at assisting teachers in instructional design specification. This tool will enable simple and intuitive specification of instructional design templates. There are ongoing activities aimed at creating a graphical editor for ELIDL that would enable creating templates without knowing ELIDL syntax. This editor should be only a part of the integrated GUI application that we are planning to develop. This application will provide a graphical interface for our system. Currently, our system is used from the command line.

Although in our course some activities are done by students and some by teachers, our system doesn't have a component for an explicit specification of users and their roles. Our long-term goal is to create such a component. It would enable collaborative learning, which is not currently supported.

Acknowledgements. Results presented in this paper are part of the research conducted within the Grant No. III-47003 financed by the Ministry of Education and Science of the Republic of Serbia.

References

1. Advanced Distributed Learning (ADL): SCORM 2004 4th edition - Sequencing and Navigation Version 1.1 (2009). [Online]. Available: www.adlnet.gov (current June 2011)
2. IMS Global Learning Consortium: IMS Learning Design Information Model (2003). [Online]. Available: www.imsglobal.org/learningdesign/ldv1p0 /imsld_infov1p0.html (current June 2011)
3. Kontopoulos, E., Vrakas, D., Kokkoras, F., Bassiliades, N., Vlahavas, I.: An Ontology-based Planning System for e-Course Generation, Expert Systems with Applications, Vol. 35, No. 1-2, 398-406. (2008)
4. Hernandez, J., Baldiris, S., Santos, O. C., Fabregat, R., Boticario, J. G.: Conditional IMS Learning Design Generation using User Modeling and Planning Techniques. In Proceedings of the 9th IEEE International Conference on Advanced Learning Technologies, IEEE Computer Society, Riga, Latvia, 228-232. (2009)
5. Capuano, N., Gaeta, M., Micarelli, A., Sangineto, E.: An Integrated Architecture for Automatic Course Generation. In Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT 02), Kazan, Russia, 322-326. (2002)
6. Morales, L., Castillo, L., Fernandez-Olivares, J., Gonzalez-Ferrer, A.: Automatic Generation of User Adapted Learning Designs: An AI-Planning Proposal. In Proceedings of the 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Springer-Verlag, Hannover, Germany, 324-328. (2008)
7. Arapi, P., Moumoutzis, N., Mylonakis, M., Theodorakis, G., Christodoulakis, S.: A Pedagogy-Driven Personalization Framework to Support Adaptive Learning Experiences. In Proceedings of the 7th IEEE International Conference on Advanced Learning Technologies (ICALT 2007), Nigata, Japan, 96-97. (2007)
8. Pacurar, E. G., Trigano, P., Alupoaie, S.: Knowledge base for automatic generation of online IMS LD compliant course structures. Educational technology & Society, Vol. 9, No. 1, 158-175. (2006)
9. LAMS: Learning Activity Management System. [Online]. Available: www.lamsinternational.com (current June 2011)
10. The Pedagogical Patterns Project. [Online]. Available: www.pedagogicalpatterns.org (current June 2011)
11. McAlpine, I., Allen, B.: Designing for active learning online with learning design templates. In ICT: Providing choices for learners and learning, Proceedings ascilite Singapore 2007, Centre for Educational Development, 639-651. (2007)
12. Heathcote, E. A.: Learning design templates – a pedagogical just-in-time support tool. In: Minshull, Geoff & Mole, Judith (Eds.) Designing for Learning, JISC Development Group, Bristol, UK, 19-26. (2006)
13. Abbing, J., Koidl, K.: Template Approach for Adaptive Learning Strategies. In Proceedings of Adaptive Hypermedia, Dablin, Ireland. (2006)
14. Guangzuo, C., Xinqi, R., Haitao, Z., Ronghuai, H.: SMID: A Semantic Model of Instructional Design. In Proceedings of the 2009 First International Workshop on Education Technology and Computer Science, Vol. 3, IEEE Computer Society, 130-134. (2009)
15. de Moura Filho, C.O., Derycke, A., Pedagogical Patterns and Learning Design: When Two Worlds Cooperate. In Proceedings of the UNFOLD-PROLEARN Joint Workshop, Valkenburg, The Netherlands. (2008)

16. Ljubojevic, D., Laurillard, D., A theoretical approach to distillation of pedagogical patterns from practice to enable transfer and reuse of good teaching. In Proceedings of the 2010 European LAMS & Learning Design Conference, Oxford, UK. [Online]. Available: http://lams2010.lamsfoundation.org/papers.htm (current June 2011)
17. Segedinac, M., Savić, G., Konjović, Z.: Knowledge representation framework for curriculum development. In Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Valencia, Spain. (2010)
18. Segedinac, M. T., Konjović, Z., Segedinac, M. D., Savić, G.: A formal approach to organization of educational objectives. Psihologija, In press. (2011)
19. IMS Global Learning Consortium: IMS Content Packaging Specification v1.2 (2007). [Online]. Available: www.imsglobal.org/content/packaging (current June 2011)
20. IEEE: IEEE 1484.12.1-2002 Learning Object Metadata Standard (2002). [Online]. Available: http://ltsc.ieee.org/wg12 (current June 2011)
21. Hlebowitsh, P. S.: Designing the School Curriculum. Pearson Education, Inc., USA. (2005)
22. Wiley, D. A., Learning object design and sequencing theory, PhD dissertation, Department of Instructional Psychology and Technology, Brigham Young University (2000). [Online]. Available: www.opencontent.org/docs/dissertation .pdf (current June 2011)
23. Chew, L. K., Hua, T. G., Instructional Strategies and Limitations of the SCORM 2004, In Proceedings of the 16th International Conference on Computers in Education (ICCE 2008), Taipei, Taiwan, 153-160. (2008)
24. Savić, G., Segedinac, M., Konjović, Z.: The Implementation of the IMS LD E-course Generator. In Proceedings of the 1st International Conference on Information Society, Technology and Management (ICIST 2011), Kopaonik, Serbia (2011)
25. W3C: OWL Web Ontology Language Overview (2004). [Online]. Available: www.w3.org/TR/owl-features (current June 2011)
26. CLEO Lab: CLEO Extensions to the IEEE Learning Object Metadata (2003). [Online]. Available: www.oasis-open.org/committees/ download.php/20490 /CLEO_ LOM_Ext_v1d1a.pdf (current June 2011)
27. Derntl, M., Motschnig-Pitrik, R., Patterns for Blended, Person-Centered Learning: Strategy, Concepts, Experiences, and Evaluation. In Proceedings of the 2004 ACM symposium on Applied computing, ACM, Nicosia, Cyprus, 916-923. (2004)
28. Reload: Learning design player v. 2.1.3 (2010). [Online]. Available: www.reload.ac.uk/ldeditor.html (current June 2011)
29. CopperCore: The IMS learning design engine v 3.3 (2008). [Online]. Available: http://coppercore.sourceforge.net (current June 2011)

Goran Savić graduated from the University of Novi Sad, Faculty of Sciences, in 2006. He is currently on PhD studies in Computer Science, from the University of Novi Sad, Faculty of Technical Sciences where he is a teaching assistant. He has authored papers in international and national journals and conferences. His research interest is e-learning.

Milan Segedinac was born in Novi Sad, on May 28, 1984. He received his M.Sc. degree from the Faculty of Technical Sciences at the University of

Novi Sad in 2008, where he entered the Ph.D. studies the same year. He has been employed at the Faculty of Technical Sciences since 2011 as a teaching assistant. His research interests are in the area of computer-enhanced education. He has authored papers in international and national journals and conferences.

Zora Konjović received her Bachelor degree in Mathematics from the Faculty of Natural Science Novi Sad (in 1973), the Master degree and PhD (both in Robotics, in 1985. and 1992 respectively) from the Faculty of Technical Sciences Novi Sad. She is a full professor at the Faculty of Technical Sciences, Novi Sad, Serbia since 2003. Prof Konjović participated in 30 research projects (as the project leader in 18). She published more than 180 scientific and professional papers. Her current research interests include artificial intelligence, web programming, digital libraries and archives, and geo-informatics.

Received: June 15, 2011; Accepted: February 21, 2012.

DOI: 10.2298/CSIS111129017C

Recommending Collaboratively Generated Knowledge

Weiqin Chen[1,2] and Richard Persen[1]

[1] Department of Information Science and Media Studies,
University of Bergen, POB 7802, N-5020 Bergen, Norway
{weiqin.chen, richard.persen}@@infomedia.uib.no
[2] Faculty of Technology, Art and Design
Oslo and Akershus university college of applied sciences
POB 4, St. Olavs plass, NO-0130 Oslo, Norway

Abstract. With the development and adoption of information technologies in education, learners become active producer of knowledge. There is an increasing amount of content generated by learners in their learning process. These emerging learning objects (ELOs) could potentially be valuable as learning resources as well as for assessment purpose. However, the potentials also give rise to new challenges for indexing, sharing, retrieval and recommendation of such learning objects. In this research we have developed a recommender system for emerging learning objects generated in a collaborative knowledge building process and studied the implications and added values of the recommendations. We conducted two evaluations with learners to assess and improve the system's design and study the quality and effects of the recommendations. From the evaluations, we received generally positive feedback and the results confirm the added values of the recommendations for the knowledge building process.

Keywords: Recommender systems, collaborative knowledge, knowledge building, emerging learning objects.

1. Introduction

Traditionally, learning objects refer to resources that are created mainly by teachers or course designers. These learning objects are mostly self-contained and vary in granularity. For example, a course, a simulation, or a piece of text can all be learning objects. Learning objects can be aggregated into a larger collection of content. Substantial effort has been made in annotating (with metadata), sharing, and reusing these learning objects. The benefits of reusing and sharing learning objects have been studied intensively in the last decade [27, 31, 35]. Recommender system for learning objects or learning materials have been developed [2, 42, 52, 54]. These systems are based on traditional learning objects which are either standard-compliant or have been annotated by teachers or learners. In addition, because of the educational potentials that information on the Internet presents, efforts have been made to take advantage of the online information. For example, some systems directly search the Internet using keywords for learning material recommendation [40]. Some use data mining techniques to automatically extract meta-data and relations among learning objects [54]. When it comes to evaluation of the systems, the main focus has been on the algorithm and technical measurement. Few of the systems for recommending learning objects have been evaluated through trials with human learners although the importance of such evaluation has been highlighted [34]. Subjective criteria (e.g. usability) in standard human-computer interaction research are rarely applied to evaluate adaptive systems including recommender systems [34, 53].

With the development and adoption of digital technologies in learning, there is an increasing amount of material generated by learners in their learning process. For example, in a knowledge building process, there are a large number of messages posted by learners and learning groups, including problems, hypothesis, and scientific evidence [48, 49]. In an inquiry learning process, learners generate many different materials (data collected, pictures taken, models created, and hypotheses generated) [19, 23]. In contrast to traditional learning objects, these learner-generated contents, also called emerging learning objects, are not pre-fabricated or authored [23]. They represent the growing knowledge and experience of learners and groups. The emerging learning objects could potentially be valuable as learning resources as well as for assessment purpose. For example, learners can build new knowledge based on existing learning objects created by themselves or other learners. Adaptive learning can be generated based on these objects. However, the potentials also give rise to new challenges for indexing, sharing, retrieval and recommendation of such learning objects which call for intelligent support for dynamic annotation and modelling of learning contexts on a semantic level [23]. In addition, it is essentially important to study the effect of these emerging learning objects on the learning process and outcome [17, 41, 55].

2. Recommender Systems

This section presents an overview of general recommender systems and recommendations in learning environments and describes the methods for creating recommendations. Special attentions are paid to the systems that recommend emerging learning objects.

2.1. General Recommender Systems

Recommender systems have been an important research area over the last decade [44]. The aim of recommender systems is to help users to deal with information overload and provide personalized recommendations, contents and services [1]. In recent years a number of recommender applications have been implemented and applied in different fields. For example, Amazon recommends books, CDs and other products to the users [30] and Google News recommends news stories to readers. Some systems recommend research papers to the users based on the users' profiles [5, 36].

In order to provide recommendations to users, it is crucial for the systems to represent the user behaviors and the information about the items to be recommended. Many of current recommender systems include a user profile (user model) that contains information about the user's tastes, preferences and needs. This information can be elicited from the users explicitly, e.g. through questionnaires or user's rating of items, or implicitly, e.g. learned from the user's activities over time. In addition, contextual information has also been incorporated into the recommendation process. Based on the methods used, recommender systems are usually classified into three categories: Content-based recommendation recommends items similar to those the user preferred in the past [4]. Collaborative recommendation recommends items that other people with similar tastes and preferences liked in the past [28, 45]. Hybrid recommendations combine the content-based and collaborative methods [10, 37].

Early research in adaptive hypermedia [9] focused mainly on content-based adaptation, which is to adaptively present a selection of links or content most appropriate to the current user. Part of this branch of study can be considered alternatively as content-based recommendations when the users' past reading items/pages are recorded and analyzed.

2.2. Recommendations in educational settings

Some of the techniques in recommender systems have been adopted for educational purposes. However, making recommendations in learning environments is different from that in other domains (e.g. movies, news) because of the particular pedagogical considerations [15, 51]. Learning objects/items liked by learners might not be pedagogically appropriate for them. Therefore a recommendation based purely on the learner's interest may not be suitable from pedagogical point of view. Other factors such as the learning context should be taking into consideration as well.

In learning environments, the recommendation mechanisms focus on providing personalized advice to the learners which are pedagogically suitable to the learner and learning activities. Some systems can recommend knowledgeable people or possible collaborators who share similar interests or are working on similar topics or problems [3, 6]. Some systems provide learners with navigation support or advice on suitable learning activities to follow [15, 39]. Some systems recommend learning resources [34]. Santos [47] defined eight types of recommendations in lifelong learning including motivation, learning style, technical support, previous knowledge, collaboration, interest, accessibility and scrutability. An educational recommender system normally keeps a learner model which includes the learner's preferences, learning history and current learning activities and context. Recommendations can be initiated by the learner implicitly or explicitly. For example, a learner is carrying on his/her learning activities in a certain location. The system presents context-based recommendations. The recommendations are triggered by the activity of the learner. A learner can also explicitly require recommendations from the system by e.g. initiating a query for possible collaborators. Among other elements in learner model, recommendations of learning objects are mainly based on three types of information about learning activities:

- context: Current location of the learner or learning activities. Many learning environments with handheld devices provide just-in time recommendations (learning objects and/or peers) based on the learner's location and activities [12, 26]. For example, descriptions of a painting are given to learners on their PDAs or mobile phones when they walk towards the painting in a museum.
- similarity: The ranking of the learning objects by the learner. The learning environment analyzes the rankings and builds a model of the learning objects based on the ranking [29, 51]. The environment then recommends relevant learning objects (based on clustering of objects) and possible collaborators (based on clustering of learners).
- topic: The current topic on which the learning activity is focused. By analyzing the current learning activities, the learning environment can identify the topics in focus. These topics are used to retrieve relevant learning objects as recommendations [20, 56]. The learning objects are classified into different categories using clustering techniques beforehand.

For an extensive review of recommender systems in technology-enhanced learning, see [34].

2.3. Recommending Emerging Learning Objects

With the rapid growing access of teachers and learners to the Internet, Web resources are becoming an important aspect in educational settings. In order for such resources to be used in a productive, educational relevant ways, recommenders systems have been developed to provide adaptive support based on web resources [16]. According to [9], the Web educational resources or dynamically expanding educational repositories fall into the “open corpus” category where it is not possible to manually structure and index them with domain concepts and metadata. Traditional learning objects which are annotated and maintained systematically by teachers, designers and institutions can be considered as in the category of “closed corpus” because the semantic description of these objects are predefined. Lying between the “open corpus” and “closed corpus” is the emerging learning objects which are created by learners or learning groups for a certain learning goal/task with a certain tool and in a certain learning environment. On the one hand, these objects are growing in volume and quality and they are not annotated or indexed at design time. On the other hand, they carry with them context information as part of its meta-data. Therefore emerging learning object can be considered as “semi-open corpus”.

Efforts have been made in providing recommendations based on learner generated content in a community setting [16, 25, 33, 43]. Tang and McCalla [51] describe a system that recommends research papers based on learners' explicit rating of the papers. ReMached [16] mashes data of learners from various Web2.0 services in order to recommend learning resources. The majority of the systems are based on tags and ratings that learners provide. This implies the use of collaborative filtering or hybrid method for recommendations and requires explicit feedback from the learners. Explicit feedback could be difficult to obtain, so data mining techniques have been developed to mine the usage data [25].

Another line of research in recommending learner generated content is based on the current topic on which the learning activity is focused. Ye [57] describes the CodeBroker which can predicate the learner's information needs and recommend code examples/software components created by other learners based on the task being performed, the knowledge of the learner performing it and the information used. Hartmann and colleagues [20] developed HelpMeOut, a social recommender system that aids the debugging of error message by suggesting solutions that peers have applied in the past.

2.4. Recommending Collaborative Knowledge in Educational Discussion Forums

Discussion forums and bulletin boards have been widely used in web-based education and computer supported collaborative learning (CSCL), in order to assist learning and collaboration. Learners use discussion forums to discuss course-related issues, such as topics in their courses, learning tasks, and projects, etc. These discussion forums include questions and answers, examples, articles posted by former learners, and thus they are potentially useful for future learners [22]. There are different variations of the educational discussion forum based on different pedagogies for collaborative learning. For example, some require learners to specify categories for their messages, and others use “sentence openers” to help learners with scientific thinking and message writing. By identifying relevant messages and reusing them as new learning resources, future learners can benefit from former learners' knowledge and experiences.

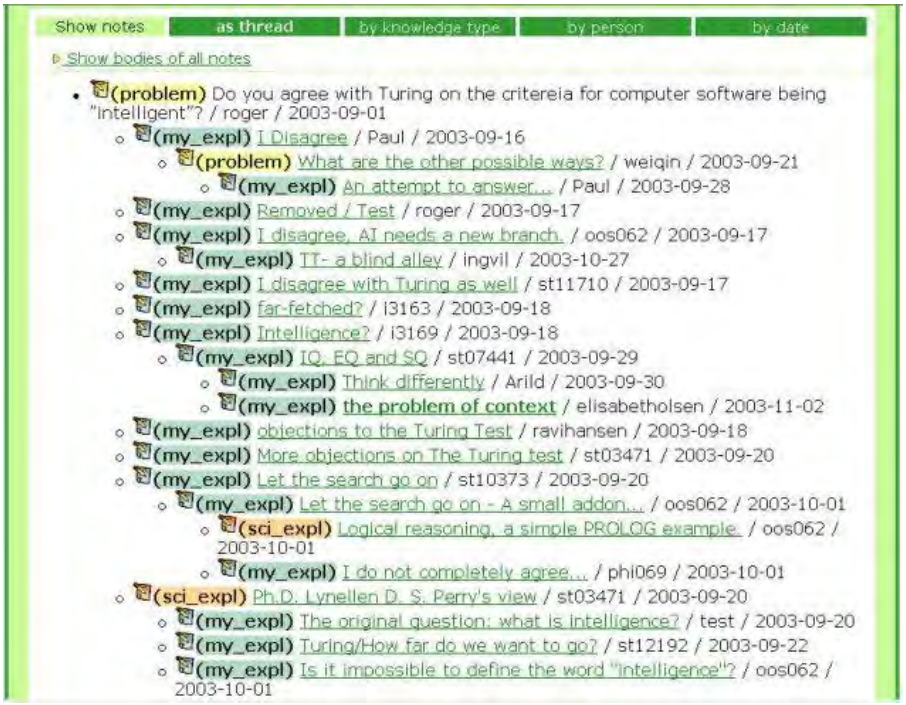
However, it is not an easy task to identify relevant information from discussion forums given the thread-based structure of them. Messages posted in a discussion forum are usually organized as a tree structure with each branch as a thread. In each thread, the messages are presented in a temporal sequence. It is usually not so easy to decide whether the message is relevant by looking at the title alone, because it is not always informative. It is possible to use a full text search within the discussion forum based on keywords. However, there are always some irrelevant messages that are included in the search result. The modern information retrieval techniques and methods [50] are rarely adopted in the search for information in discussion forums.

A few efforts on indexing and reusing the messages in educational discussion forums have been made. The main method is to create a predefined structure for a discussion forum, where the structure reflects a conceptual schema of the subject domain [14]. Helic and his colleagues [21, 22] described a tool to support conceptual structuring of discussion forums. They attached a conceptual schema to a discussion forum, and the learners had to manually assign their messages to the schema. Their study shows some limitations with this method. First, some messages could be assigned to more than one concept in the schema. Second, the learners were not motivated enough to make the extra effort in assigning their messages to concepts, although it may have been beneficial to those learners to do so. Our own experience confirms the second point. We developed a plug-in for FLE3 (see section below), where students could choose relevant topics when preparing their messages, but they could also chose to ignore this feature. Very few students made use of this function to specify relevant topics for their message.

In our research we have developed a recommendation system, AnnForum, which can identify relevant existing message from threaded discussion in FLE3 and recommend them to learners. For each message in current or previous discussion, the system associates it with a certain topic (defined in a domain model) using text analysis and clustering techniques, with each association a value describing the relevance between the message and the topic. This process can be automatic or semi-automatic where the teacher is provided with a tool to check and manually specify the association generated by automatic mechanism. When a learner is reading or writing a message, the system identifies the current topic and retrieves relevant messages with a matching mechanism. The relevant messages (each with a relevant value) are provided to the learner as recommendations. The intention is to provide just-in-time recommendations that are meaningful to the current problem so that the learners can look at the problem from a different viewpoint, evaluate their own thoughts, and be inspired by and build upon the ideas in the recommendations.

3. Collaborative Knowledge Building with FLE3

FLE3 [38] is web-based groupware for computer supported collaborative learning (CSCL). It is designed to support the collaborative process of progressive inquiry learning. The basic idea of progressive inquiry is that learners gain a deeper understanding by engaging in a research-like process where they generate their own problem, make hypotheses, and search out explanatory scientific information collaboratively with other learners.



- (problem) Do you agree with Turing on the critereia for computer software being "intelligent"? / roger / 2003-09-01
    - (my_expl) I Disagree / Paul / 2003-09-16
        - (problem) What are the other possible ways? / weiqin / 2003-09-21
            - (my_expl) An attempt to answer... / Paul / 2003-09-28
    - (my_expl) Removed / Test / roger / 2003-09-17
    - (my_expl) I disagree, AI needs a new branch. / oos062 / 2003-09-17
        - (my_expl) TT- a blind alley / ingvil / 2003-10-27
    - (my_expl) I disagree with Turing as well / st11710 / 2003-09-17
    - (my_expl) far-fetched? / i3163 / 2003-09-18
    - (my_expl) Intelligence? / i3169 / 2003-09-18
        - (my_expl) IQ, EQ and SQ / st07441 / 2003-09-29
            - (my_expl) Think differently / Arild / 2003-09-30
            - (my_expl) the problem of context / elisabetholsen / 2003-11-02
    - (my_expl) objections to the Turing Test / ravihansen / 2003-09-18
    - (my_expl) More objections on The Turing test / st03471 / 2003-09-20
    - (my_expl) Let the search go on / st10373 / 2003-09-20
        - (my_expl) Let the search go on - A small addon... / oos062 / 2003-10-01
            - (sci_expl) Logical reasoning, a simple PROLOG example. / oos062 / 2003-10-01
            - (my_expl) I do not completely agree... / phi069 / 2003-10-01
    - (sci_expl) Ph.D. Lynellen D. S. Perry's view / st03471 / 2003-09-20
        - (my_expl) The original question: what is intelligence? / test / 2003-09-20
        - (my_expl) Turing/How far do we want to go? / st12192 / 2003-09-22
        - (my_expl) Is it impossible to define the word "intelligence"? / oos062 / 2003-10-01

Fig. 1. Knowledge Building in FLE3

To support the collaborative progressive inquiry process, FLE3 provides several modules, such as virtual WebTop, a Knowledge Building module, and an Administration module. The Knowledge Building module is considered to be the scaffolding module for progressive inquiry, where learners post their messages to the common workspace according to predefined categories. The categories they can use are Problem, My Explanation, Scientific Explanation, Comment, and Summary. These categories are defined to reflect the different phases in the progressive inquiry process (Fig. 1).

FLE3 has been used as a knowledge building tool in our Introductory Artificial Intelligence course for discussing issues such as "what is intelligence" and "can machine think". Students follow the progressive inquiry process and collaboratively build knowledge and understand the concepts in artificial intelligence.

4. Design and Implementation of AnnForum

In order to recommend relevant messages from the knowledge building process, a conceptual domain model is constructed first. Based on this model, the messages posted in previous knowledge building processes are annotated and classified into different categories corresponding to different concepts in the model. The teacher is responsible for constructing and managing the domain model, as well as for validating (adding/removing) annotations. The messages (including those in the current and previous knowledge building process) that are relevant to the current topic under discussion are gathered and presented to learners. The learners can then read through the messages and rank them according to their degree of relevance. Fig. 2 shows the use cases for the recommender system.

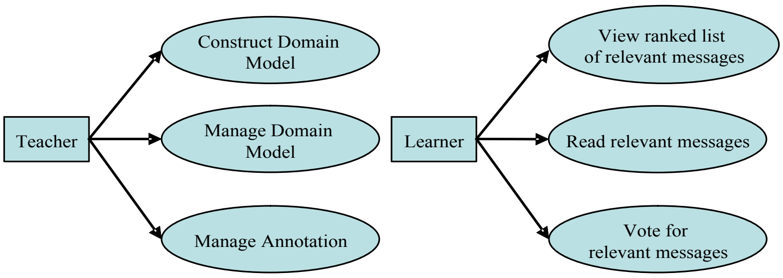


Fig. 2. Use cases

The following two scenarios explain how AnnForum is used by learners.

- Scenario 1: A learner writing a new message

After the learner has finished writing their message, it is submitted and appears in the knowledge building interface. AnnForum automatically annotates and classifies this message based on the domain model. In the meantime, it finds a list of existing messages that are relevant to the learner's message by computing the relevant values. The ranked list of relevant messages is presented to the learner. The annotation and relevant values are stored into AnnForum's database.

- Scenario 2: A learner reading an existing message in the knowledge building interface

When the learner is reading the message, they click on a button called "show relevant messages". The relevant interface appears, containing a ranked list of relevant messages. These messages are retrieved from the database.

4.1. Main Components

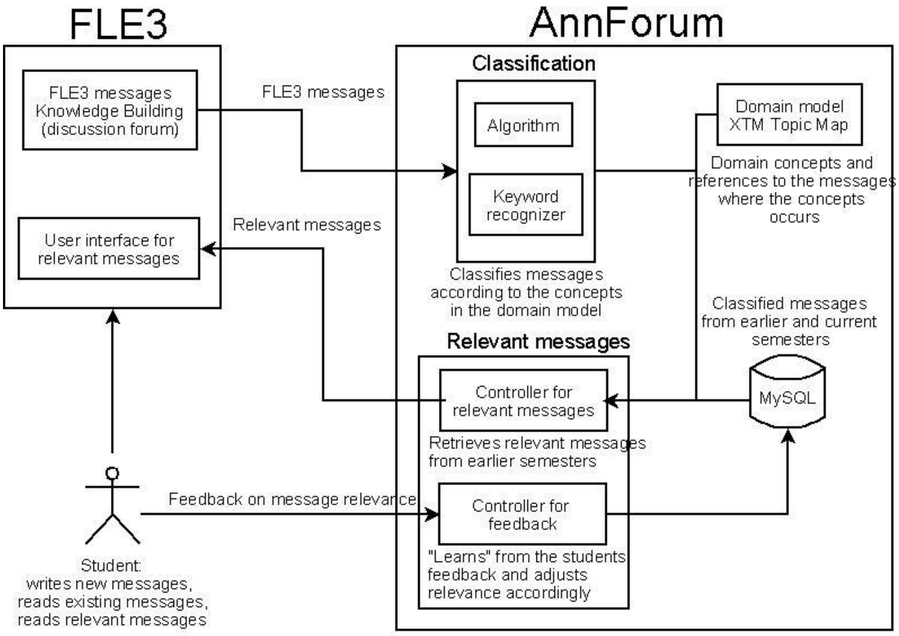
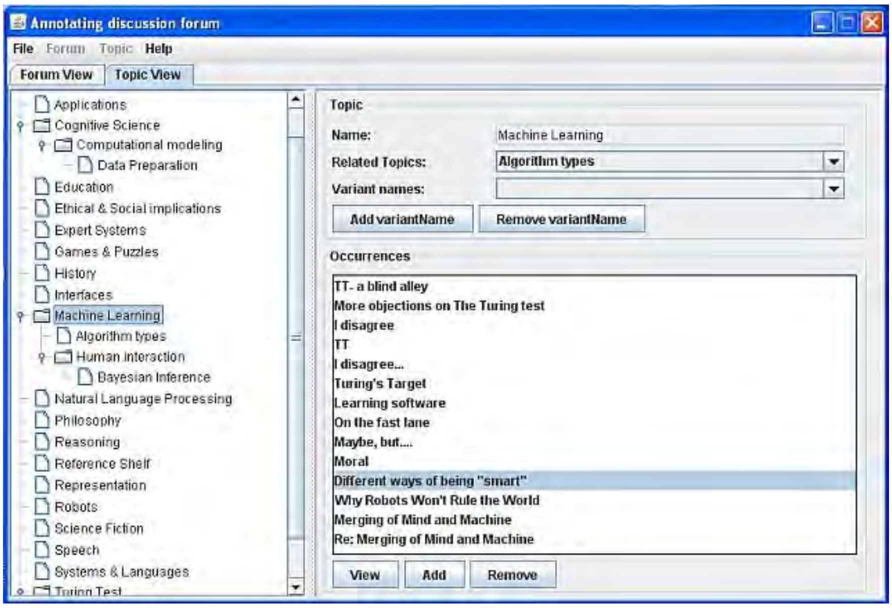

FLE3
AnnForum
Classification
FLE3 messages Knowledge Building (discussion forum)
FLE3 messages
Algorithm
Keyword recognizer
Domain model XTM Topic Map
Domain concepts and references to the messages where the concepts occurs
User interface for relevant messages
Relevant messages
Classifies messages according to the concepts in the domain model
Classified messages from earlier and current semesters
Relevant messages
Controller for relevant messages
MySQL
Retrieves relevant messages from earlier semesters
Feedback on message relevance
Controller for feedback
"Learns" from the students feedback and adjusts relevance accordingly
Student: writes new messages, reads existing messages, reads relevant messages

Fig. 3. Main Components in AnnForum

AnnForum is a plug-in to the FLE3 environment. It is a domain-independent tool. As shown in Fig. 3, AnnForum takes the domain model and the messages as input, and puts the classified messages into the database. When a new message comes, the Classification module decides its relevant topics. Then it searches for the relevant messages in the database, computes the relevant values based on the relevance of the messages, and stores them in the database. The Controller for relevant messages module retrieves the relevant messages and sends them to the interface in FLE3. The learners can then read through the messages and rank them according to whether they think it is relevant or not. In the current design, the learners do not need to explicitly associate their message to a topic in the domain model. However, if they are willing to do so, they can use the “choose topics” function when preparing a message, which is provided by the plug-in to FLE3. The Controller for feedback module learns from the feedback of the learners and adjusts the weights used in the matching algorithm accordingly.

4.2. Conceptual Domain Model and Annotation of Messages

A conceptual domain model is used to describe the domain concepts and the relationships among them, which collectively describe the domain space. This domain model is usually represented by an ontology. A simple conceptual domain model can also be represented by a topic map. Topic Maps ISO/IEC13250 [24] is an ISO standard for describing knowledge structures and associating them with information resources. It is used to model topics and their relations in different levels. The main components in Topic maps are topics, associations, and occurrences. The topics represent the subjects, i.e. the things, which are in the application domain, and make them machine understandable. A topic association represents a relationship between topics. Occurrences link topics to one or more relevant information resources. Topic maps provide a way to represent semantically the conceptual knowledge in a certain domain.



Topic

Name: Machine Learning

Related Topics: Algorithm types

Variant names:

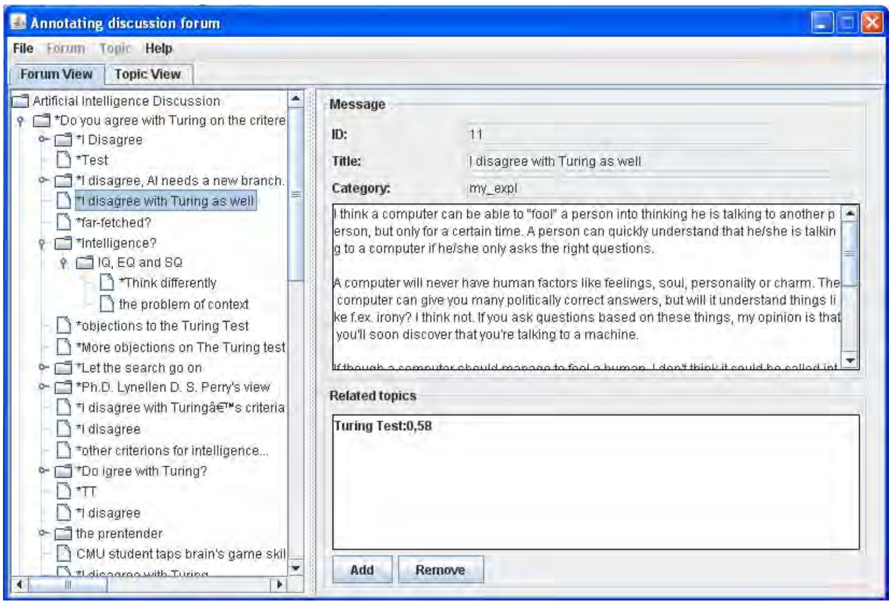
Occurrences

TT- a blind alley
More objections on The Turing test
I disagree
TT
I disagree...
Turing's Target
Learning software
On the fast lane
Maybe, but....
Moral
Different ways of being "smart"
Why Robots Won't Rule the World
Merging of Mind and Machine
Re: Merging of Mind and Machine

Fig. 4. Topic and association management

In AnnForum we use a topic map to represent the domain model of Artificial Intelligence. This domain model includes AI topics and their relations such as machine learning, agents, knowledge representation, searching algorithm, etc. These topics are described as topics in the topic map. Relations between these topics are represented as associations. The occurrence describes the links to the messages where the topic was discussed in the discussion forum. The occurrence is generated by the automatic classification algorithm presented in next subsection.

In the earlier prototypes of the system, teachers had to write XML in order to create Topic maps for their course domains and when a message is posted, associated topics to this message have to be selected manually by the contributors (learners/teachers). These have been proved rather tedious. In the newer versions AnnForum provides a graphical interface for teachers to create a domain topic map interactively (Fig. 4). Using AnnForum, teachers can create Topic maps for their course domain and load /reload them into FLE3. Because the topic map is written in XML format (XTM), it is easy for teachers to understand and maintain the topics, and the domain model can also be easily reused in other contexts. Fig. 4 also shows the associations between the messages and the related topic ("machine learning") using automatic classification techniques. Teachers can also use this tool to edit and verify the associations.



Artificial Intelligence Discussion
- *Do you agree with Turing on the critere
  - *I Disagree
  - *Test
  - *I disagree, AI needs a new branch.
  - *I disagree with Turing as well
  - *far-fetched?
  - *Intelligence?
    - IQ, EQ and SQ
      - *Think differently
      - the problem of context
  - *objections to the Turing Test
  - *More objections on The Turing test
  - *Let the search go on
  - *Ph.D. Lynellen D. S. Perry's view
  - *I disagree with Turingâ€™s criteria
  - *I disagree
  - *other criterions for intelligence...
  - *Do igree with Turing?
  - *TT
  - *I disagree
  - the prentender
  - CMU student taps brain's game skil
  - *I disagree with Turing

ID: 11

Title: I disagree with Turing as well

Category: my_expl

I think a computer can be able to "fool" a person into thinking he is talking to another p
erson, but only for a certain time. A person can quickly understand that he/she is talkin
g to a computer if he/she only asks the right questions.

A computer will never have human factors like feelings, soul, personality or charm. The
computer can give you many politically correct answers, but will it understand things li
ke f.ex. irony? I think not. If you ask questions based on these things, my opinion is that
you'll soon discover that you're talking to a machine.

If though a computer should manage to fool a human, I don't think it could be called int

Turing Test:0,58

Fig. 5. Manual annotation of messages

Fig. 5 shows an interface where teachers can manually associate messages to topics. The left panel shows the threads of the knowledge building forum. The * in front of the message title means that this message has been classified automatically. In the right panel, teachers can view the information for each message in the discussion thread (including title, knowledge building category and content) and the topics this message is related to. In the message content, the identified topics, such as "Turing Test" was listed with a relevance value. They can also add or remove the related topics by clicking on the buttons at the bottom of the right panel.

4.3. Message Classification

Since the messages can be seen as a kind of document collection, we investigate the methods for classifying documents from a document collection to a conceptual model. There are various approaches in information retrieval that deal with the problem of document classification. For example, ontology-based classification of unstructured/semi-structured documents goes beyond the use of keywords and classifies documents into categories that are meaningful to users [11, 13].

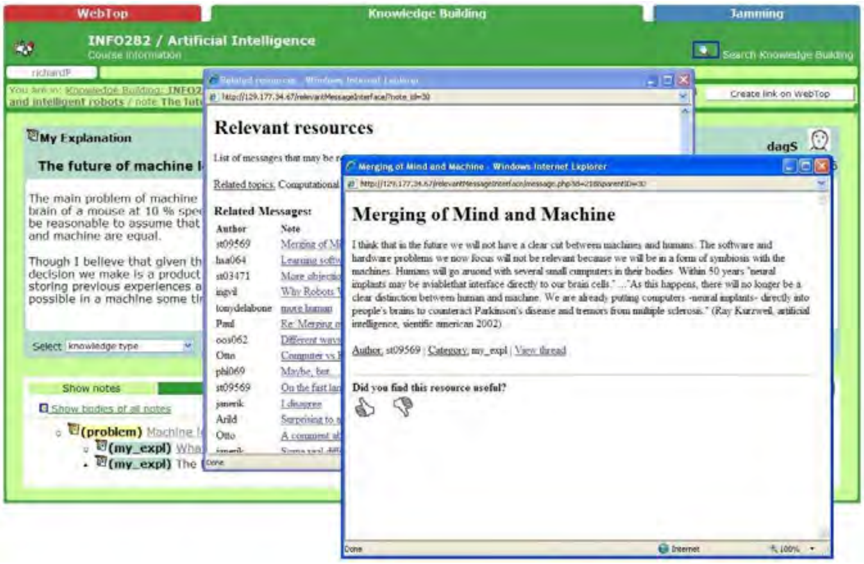
In AnnForum we use an approach that combines a conceptual model (represented by a Topic Map) and a Vector Space Model based classification to determine the relevance of a message to a concept in the domain model. It could be considered a simplified ontology-based text classification because in our Topic Map, the associations between concepts are limited to is-a relations. The core of our approach can be seen in three steps:

1. Generating term vector Vt: We use Apache Lucene to create the TF-IDF (term frequency-inverse document frequency) weighted term vector which consists of words and phrases in the document ordered by their relative importance [46]. For example, a message (id: m-26) in the knowledge building process has Vt: <Turing test: 0.043, production rules: 0.025, machine learning: 0.024, color: 0.009…>
2. Mapping the term vector Vt to Topic Map: For those terms that are in both Vt and the Topic Map, they form a new term vector Vtt. The basenames and variants of the basenames in the Topic Map are used to normalize the vector to account for synonyms. Vtt for message m-26 becomes <Turing test: 0.043, production rules: 0.025, machine learning: 0.024, …>. Color with its weight has disappeared from the vector because it is not in the Topic Map.
3. Enhancing the term vector by the associations in the Topic Map and adjusting the weights of terms in Vtt: For every term Ti in Vtt, we use this term to find the associated topics in the Topic Map. If an associated topic Tj is not already in Vtt, Vj will be added to Vtt and the weight for Tj is 25% of the weight of Ti. If an associate topic Tj exists in Vtt, both the weight of Ti and Tj will increase 50%. Vtt for message m-26 becomes <Turing test: 0.043, production rules: 0.025, machine learning: 0.024, knowledge representation: 0.013, …>. Knowledge representation has been added to the vector because it is associated with production rules in the Topic Map. Note that the weight adjustment factors (25% and 50%) are rather arbitrary in the current version and some heuristics and empirical data will be used in the future to fine tune these values.

The classification results are stored in a MySQL database. The database includes both the messages (title, author, timestamp, and thread information), and the domain topics they are related to, with values of relevance.

4.4. Relevant Message Recommendations and Learner's Feedback

The Controller for relevant messages module in AnnForum retrieves the relevant messages and sends them to the interface in FLE3. Fig. 6 shows the interface where learners get the relevant message recommendations. They can click the title link to read the message. After reading they can also vote for or against the messages using the Thumb up/down buttons. The voting will affect the relevance value later. The Controller for feedback module adjusts the weights based on the votes. Teachers can also use the vote function to change the relevance value.



The main problem of machine brain of a mouse at 10 % spe be reasonable to assume that and machine are equal.

Though I believe that given th decision we make is a product storing previous experiences a possible in a machine some ti

Relevant resources

List of messages that may be r

Related topics: Computational

Related Messages:

Author	Note
st09569	Merging of Mi
lsa064	Learning softw
st03471	More objectio
ingvil	Why Robots
tonydelabone	more human
Paul	Re: Merging o
oos062	Different wavs
Otto	Computer vs I
phi069	Maybe, but
st09569	On the fast lan
janerik	I disagree
Arild	Surprising to s
Otto	A comment ab

Merging of Mind and Machine

I think that in the future we will not have a clear cut between machines and humans. The software and hardware problems we now focus will not be relevant because we will be in a form of symbiosis with the machines. Humans will go aruond with several small computers in their bodies. Within 50 years "neural implants may be aviablethat interface directly to our brain cells." ..."As this happens, there will no longer be a clear distinction between human and machine. We are already putting computers -neural implants- directly into people's brains to counteract Parkinson's disease and tremors from multiple sclerosis." (Ray Kurzweil, artificial intelligence, sientific american 2002)

Author: st09569 | Category: my_expl | View thread

Did you find this resource useful?

Fig. 6. Recommendation interface

5. Evaluation

We have conducted two formative evaluations with improvement in between. In contrast to most recommender systems which evaluate the algorithms, our evaluations focus mainly on subjective criteria based on principles in human computer interaction.

5.1. Formative Evaluation 1

This evaluation focused mainly on usability issues. It had three main goals: to assess the extent of the system's functionality, to assess the effect of the system on the learner, and to identify any specific problems with the system. More specifically, the evaluation aims to answer the following three questions:

- How well does the system annotate and classify existing messages based on the domain model?
- How useful are the relevant messages to the learners?
- What improvements need to be made to the system?

Six university students taking an information science major participated in the evaluation. All of them were familiar with the AI domain and had experience with discussion forums. The evaluation was carried out in a controlled environment where only one participant and one researcher were present. After a short introduction about the system, the participant was given a set of tasks to carry out. Data were collected by observation and interview after the tasks. The questions in the interview reflected the three goals of the evaluation.

Before the evaluation, the researchers prepared two messages and posted them into the knowledge building module in FLE3. One of the messages concerned the Turing Test and the other concerned Machine Learning. Both are important topics in AI. These two messages were posted in the category of “Problem”, and served as the starting point for the discussion.

Each participant was asked to read existing messages, use the relevant message interface to check out relevant messages, and post at least two messages of their own responding to existing messages. This way, the number of existing messages grew as the evaluation progressed – the first participant had 2 existing messages to read and the last one had 10. This could be considered a simulation of a real knowledge building process, and it also made the dynamic nature of the system (annotating, classifying, and matching) more realistic. The current messages in the knowledge building module, as well as the 237 messages from previous knowledge building processes, were the source of the relevant messages.

The data from observation and interview show that all participants were very positive toward the system, and saw the added value of the relevant messages in their knowledge building process. They used the relevance value, or a combination of relevance value and the title of the messages, to decide which recommended relevant messages to read. After reading some of the messages, all six participants thought the message with the highest relevant value was the most relevant, while the one with the lowest relevant value was not quite relevant. Some also used the “thumb up” and “thumb down” buttons to vote for the recommended messages they read. Half of the participants responded by stating that the relevant messages they read affected the formulation of their own messages.

Perceived relevance of the recommended messages. The relevance values of recommended relevant messages were found to reflect the actual relevance to the current discussion. This indicates that the performance of the annotation, classification, and matching mechanism is acceptable. The automatic process is important because it saves learners from having to manually annotate their messages.

Perceived usefulness. The most positive aspects about providing the relevant messages include:

- The learners can build more in-depth knowledge about the discussion topic. The relevant messages provide them with different viewpoints.
- It gives the learners a feeling that the discussion is more alive, which motivates them to formulate good messages.
- The ranking list of the relevant messages is a better alternative than searching the discussion forum.
- It can reduce the possible duplication of information. Duplication of information is a problem in most big discussion forums.
- The dynamic nature of the relevant messages prevents earlier messages that are “buried” deep down in the thread from being ignored.

Improvements. The evaluation also resulted in a few ideas for improvements:

- Allowing learners to see the thread to which the recommended message belongs.
- Allowing learners to see more information about each relevant message. In this version the system shows the title and the relevant value. The feedback from the participants indicates that showing a few opening sentences of each relevant message would help the learners to make a better judgment before going on to read the whole message. This could be implemented as a mouse-over event, which means the opening sentences of the message will be shown in a floating box near the title whenever the learner moves the mouse over the title of the message, and that the floating box disappears when the mouse is moved away from the title.
- Making the relevant messages’ interface a part of the FLE3 interface with the same look-and-feel. In this version the relevant interface is implemented as a pop-up window, which, according to the participants, disturbed the workflow.

5.2. Formative Evaluation 2

This evaluation focused on the quality and effect of the recommendations in the context of learning. It was conducted after further development based on the feedback. The system was evaluated throughout the fall semester with 35 Information Science students in the Introductory AI course. Data were collected by an online questionnaire with open end questions after the course. The questions reflected the main focus of the evaluation. Some of the questions used in the questionnaire are as follows:

- Which relevant messages you read or voted for/against? Why did you choose these to read or vote/against?
- Do you think the relevant value in percentage is correct for the messages you read? Why?

- Have any of the relevant messages you read affect how you formulate your own messages?
- What do you think of the idea of recommending relevant messages?

Before the evaluation, the researcher prepared two questions about intelligence and Turing Test and posted them into the knowledge building module in FLE3. These two questions served as the starting point for the discussion. Each participant was asked to read the messages posted by others, use the relevant message interface to check out relevant messages, and post at least four messages (two in each question) of their own responding to others' messages. The current messages in the knowledge building module, as well as the 249 messages from previous semesters' knowledge building processes, were the source of the recommendations.

Perceived quality of the recommendations. The data from the questionnaire shows that the students used the relevance value, or a combination of relevance value and the title of the messages, to decide which recommended relevant messages to read. After reading some of the messages, about half (18) of the participants thought the message with the highest relevant value was the most relevant, while the one with the lowest relevant value was not quite relevant. This indicates that the performance of the classification mechanism is acceptable. About half of the participants responded by stating that the relevant messages they read affected the formulation of their own messages. Few (4 out of 36) used the “thumbs up” and “thumbs down” buttons to vote for/against the recommended messages they read because most of the students “didn’t notice” the buttons.

Effect of the recommendations. The data shows that more than half (20 out of 35) of the students were very positive toward the recommender, and saw the added value of the relevant messages in their knowledge building process. For example, to the question of “what do you think of the idea of recommending relevant messages”, here are some of the responses:

“This is a good idea, possibilities to inspire new ideas. Continuity is important both to access a first hand community history, as well as an opportunity to see what previous ideas and thought people had earlier.”

“A student can learn a lot about the views of others. It is very important because it is information and knowledge that is acquired”.

“The possibility to see relevant message is so useful…It can help you to find a new ideas, to see if your previous colleagues think more or less the same of you or completely different.”

To the question of “Have any of the relevant messages you read affect how you formulate your own messages”, here are some of the responses:

“I read this message and he (the author of the message) remembered an article which I read some years ago, so I read the article again and I was able to put my ideas.”

“Of course, I found anyone posted more or less the same as I intended, so I had to reevaluate my post and take in account the insights the post brings.”

One of the most positive aspects about the recommender is that the learners can build more in-depth knowledge about the discussion topic. The relevant messages provide them with inspirations, different viewpoints and additional resources.

6. Conclusion and Future Work

In this paper we have presented a recommender system for collaborative knowledge. It can be seen as a content-based recommender based on topics on which the learners are focusing. We have also conducted formative evaluations of the system. The general response is positive and provides a confirmation that the recommended messages may be useful. Further research is needed to judge the effect of the recommendations on learning. Content analysis based on logs of learners' activities with timestamps may provide evidence on whether the recommended messages actually influence the formulation of the new messages. Moreover, qualitative evaluation of the algorithm can also provide us insight on whether the technical decisions are appropriate.

Some possible improvements were also identified through the evaluation. One feedback is to allow learners to see the thread to which the recommended message belongs. This will give the learners context information regarding the message. Context information allows learners to have a feeling of presence, that is, that they are collaborating with previous learners [18]. In the current version in order to help the learners to make better judgment before going on to read the whole message, we show a few opening sentences of each recommended message.

The approach presented in this paper is primarily for recommending messages in educational discussion forums. But it has implications for searching in traditional discussion forums and for organizational knowledge management.

The problem with traditional discussion forum is that it is difficult to find useful information about a certain topic, especially when the number of messages grows. It becomes impossible to have an overview of threads. In addition, the titles of messages usually use “RE: XXX” and do not tell the users much about the content of the message. The results from keyword-based search are not always satisfactory. The method presented in this paper, including the dynamic annotation, classification and matching, will be able to help users in finding relevant information from traditional discussion forums by providing them a ranked list of relevant messages. This will also help to reduce the number of duplicated messages. Because the process is automatic, it does not give users overhead when they post messages.

One important research area in knowledge management is to look for better ways to handle large amount of organizational information and knowledge so that it is easy to represent, organize, maintain, search and reuse them. Annotation and information retrieval have played important roles in knowledge management. We believe that knowledge management can benefit from our research in two aspects:

- Dynamic annotation and classification allows each new piece of knowledge to be automatically annotated and classified immediately when it is stored in the organizational knowledge repository.
- Dynamic matching provides users with ranked list of relevant information and knowledge, which saves the users from having to formulizing queries by themselves.

The approach presented in this paper can also be used for retrieval and suggestion of any unstructured/semi-structured documents as learning resources. Recently wiki becomes popular as a pedagogical tool to support learning [7] where new knowledge is generated collaboratively. The approach we have used in our research can also be used to analyze wiki pages created by learners and provide recommendations to the learners based on the topics they are working on.

References

1. Adomavicius, G., Tuzhilin, E.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17, 734-749 (2005)
2. Al-Khalifa, H. S.: Building an Arabic learning object repository with an ad hoc recommendation engineProc. of the 10th international Conference on information integration and Web-Based Applications & Services (iiWAS'08). ACM, Linz, Austria (2008)
3. Anjewierden, A., Chen, W., Wichmann, A., Van Borkulo, S. P.: Process and product analysis to initiate peer-topeer collaboration on system dynamics modellingWorkshop on Intelligent and Innovative Support for Collaborative Learning Activities (WIISCOLA), Rhodes, Greece (2009)
4. Balabanovic, M., Shoham, Y.: Fab: Content-based, collaborative recommendation. Communications of the ACM 40, 66-72 (1997)
5. Basu, C., Hirsh, H., Cohen, W.: Recommendation as classification: using social and content-based information in recommendationthe Fifteenth National Conference on Artificial Intelligence (1998)
6. Behama, G., Kumpa, B., Leya, T., Lindstaedta, S.: Recommending knowledgeable people in a work-integrated learning system Procedia Computer Science (2010)
7. Bruns, A., Humphreys, S.: Wikis in teaching and assessment: the M/Cyclopedia projectthe 2005 international Symposium on Wikis. ACM, San Diego, California (2005)
8. Brusilovsky, P.: Adaptive hypermedia. User Modeling and User Adapted Interaction 11, 87-110 (2001)
9. Brusilovsky, P., Henze, N.: Open corpus adaptive educational hypermedia. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds) The adaptive web Springer-Verlag (2007)
10. Burke, R.: Hybrid recommender systems: survey and experiments. User-Modeling and User-Adapted Interaction 12, 331-370 (2002)
11. Castells, P., Fernandez, M., Vallet, D.: An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. IEEE Transactions on Knowledge and Data Engineering 19, 261-272 (2007)

12. Chen, G. D., Chang, C. K., Wang, C. Y.: Ubiquitous learning websites: scaffold learners by mobile devices with information-aware techniques. Computers & Education 50, 77-90 (2006)
13. Chen, W., Persen, W.: Reusing Collaborative Knowledge as Learning Objects -- The Implementation and Evaluation of AnnForum. In: Dillenbourg, P., Specht, M. (eds) The Third European Conference on Technology-Enhanced Learning (EC-TEL). Springer-Verlag (2008)
14. Cheng, C. K., Pan, X., Kurfess, F.: Ontology-Based Semantic Classification of Unstructured DocumentsAdaptive Multimedia Retrieval. Springer Berlin / Heidelberg (2004)
15. Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Migam, K., Slattery, S.: Learning to extract symbolic knowledge from the World Wide WebProf. of the 15th National Conference on Artificial Intelligence, Madison, Wisconsin (1998)
16. Drachsler, H., Hummel, H. G. K., Koper, R.: Personal recommender systems for learners in lifelong learning: requirements, techniques and model. International Journal of Learning Technology 3, 404-423 (2008)
17. Drachsler, H., Pecceu, D., Arts, T., Hutten, E., Rutledge, L., Van Rosmalen, P., Hummel, H., Koper, R.: ReMashed – Recommendations for Mash-Up Personal Learning Environments. In: Cress, U., Dimitrova, V., Specht, M. (eds) Proceedings of the Fourth European Conference on Technology-Enhanced Learning (EC-TEL 2009). Springer-Verlag (2009)
18. Falloon, G.: Learning objects and the development of students' key competencies: A New Zealand school experience Australasian Journal of Educational Technology 26, 626-642 (2010)
19. Garrison, D. R., Anderson, T., Archer, W.: Critical thinking in text-based environment: Computer conferencing in higher education. The Internet and Higher Education 2, 87-105 (2000)
20. Hanauer, D. I., Jacobs-Sera, D., Pedulla, M. L., Cresawn, S. G., Hendrix, R. W., Hatfull, G. F.: Inquiry Learning: Teaching Scientific Inquiry. Science 314, 1880-1881 (2006)
21. Hartmann, B., MacDougall, D., Brandt, J., Klemmer, S. R.: What Would Other Programmers Do? Suggesting Solutions to Error MessagesProceedings of the 28th international conference on Human factors in computing systems (CHI'10). ACM (2010)
22. Helic, D., Maurer, H., Scerbakov, N.: Learning to extract symbolic knowledge from the World Wide Web. ACTA Advanced Technology for Learning 1, 8-16 (2004)
23. Henze, P. B. a. N.: Open corpus adaptive educational hypermedia. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds) The Adaptive Web. Springer-Verlag (2007)
24. Hoppe, H. U.: Integrating Learning Processes Across Boundaries of Media, Time and Group Scale Research and Practice in Technology Enhanced Learning 2, 31-49 (2007)
25. Janssen, J., Tattersall, C., Waterink, W., Van den Berg, B., Van Es, R., Bolman, C., Koper, R.: Self-organising navigational support in lifelong learning: how predecessors can lead the way. Computers & Education 49, 781-793 (2007)
26. Khribi, M. K., Jemni, M., Nasraoui, O.: Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. Educational Technology & Society 12, 30–42 (2009)

27. Klamma, R., Spaniol, M., Cao, Y.: Community Aware Content Adaptation for Mobile Technology Enhanced Learning In: Nejdl, W., Tochtermann, K. (eds) Proc. of the First European Conference on Technology Enhanced Learning (EC-TEL 2006). Springer (2006)
28. Klebl, M., Krämer, B. J.: Editorial: Learning ojbects in progress. British Journal of Educational Technology 41, 869-872 (2010)
29. Kobsa, A., Koenemann, J., Pohl, W.: Personalized hypermedia presentation techniques for improving online customer relationships. The Knowledge Engineering Review 16, 111-155 (2001)
30. Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., Riedl, J.: GroupLens: Applying collaborative filtering to Usenet news. Communications of the ACM 40, 77-87 (1997)
31. Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing Jan./Feb.(2003)
32. Littlejohn, A.: Reusing Online Resources: A Sustainable Approach to e-Learning, Routledge (2003)
33. Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., Koper, R.: Recommender Systems in Technology Enhanced Learning In: Ricci, F., Rokach, L., Shapira, B., Kantor, P. B. (eds) Recommender System Handbook. Springer (2011)
34. Manouselis, N., Vuorikari, R., Assche, F. V.: Simulated Analysis of MAUT Collaborative Filtering for Learning Object RecommendationProceedings of the 1st Workshop on Social Information Retrieval for Technology-Enhanced Learning & Exchange (2007)
35. Manouselis, N., Vuorikari, R., Assche, F. V.: Collaborative recommendation of e-learning resources: an experimental investigation. Journal of Computer Assisted Learning 26, 227–242 (2010)
36. McGreal, R.: Online Education Using Learning Objects. Routledge2004)
37. McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., Konstan, J. A., Riedl, J.: On the recommending of citations for research papersACM International Conference on Computer Supported Collaborative Work (2002)
38. Melville, P., Mooney, R. J., Nagarajan, R. P., 2002.: Content-boosted collaborative filtering for improved Recommendationsthe 18th National Conference on Artificial Intelligence (2002)
39. Muukkonen, H., Hakkarainen, K., Lakkala, M.: Collaborative technology for facilitating progressive inquiry: future learning environment toolsProc. of International Conference on Computer Supported Collaborative Learning (CSCL'99), Palo Alto, CA (1999)
40. Mödritschera, F.: Towards a Recommender Strategy for Personal Learning Environments. Procedia Computer Science (Proc. RecSysTEL 2010) 1, 2775 - 2782 (2010)
41. Nitchot, A., Gilbert, L., Wills, G.: A Competence-based System for Recommending Study Materials from the Web (CBSR)Proc. 9th International Conference on Information Technology Based Higher Education and Training (2010)
42. Northrup, P. T.: Learning Objects for Instruction: Design and Evaluation. Information Science Publishing (2007)
43. Ochoa, X., Duval, E.: Use of contextualized attention metadata for ranking and recommending learning objectsProc. of the 1st international Workshop on

Contextualized Attention Metadata: Collecting, Managing and Exploiting of Rich Usage information (CAMA '06). ACM, Arlington, Virginia, USA (2006)

44. Recker, M. M., Walker, A., Lawless, K.: What do you recommend? Implementation and analyses of collaborative information filtering of web resources for education Instructional Science 31, 299-316 (2003)
45. Resnick, P., Varian, H. R.: Recommender systems. Communications of the ACM 40, 56-58 (1997)
46. Rucker, J., Polanco, M.: Siteseer: personalized navigation for the Web. Communications of the ACM 40, 73-76 (1997)
47. Salton, G., Wong, A., Yang, C. S.: A Vector Space Model for Automatic Indexing. Communications of the ACM 18, 613–620 (1975)
48. Santos, O. C.: A recommender system to provide adaptive and inclusive standard-based support along the elearning life cycleProc. of the 2008 ACM conference on Recommender systems,. ACM, Lausanne, Switzerland (2008)
49. Scardamalia, M., Bereiter, C.: Knowledge Building. In: Guthrie, J. W. (ed) Encyclopedia of Education. Macmillan Reference (2003)
50. Singhal, A.: Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (2001)
51. Tang, T., McCalla, G.: Smart Recommendation for an evolving E-Learning system: Architecture and experiment. International Journal on E-Learning 4, 105-129 (2005)
52. Tsai, K. H., Chiu, T. K., Lee, M. C., Wang, T. I.: A Learning Objects Recommendation Model based on the Preference and Ontological ApproachesProceedings of the Sixth IEEE international Conference on Advanced Learning Technologies (ICALT'06). IEEE Computer Society (2006)
53. Walker, A., Recker, M., Lawless, K., Wiley, D.: Collaborative information filtering: A review and an educational application. International Journal of Artificial Intelligence in Education 14, 1-26 (2004)
54. Wang, T. I., Tsai, K. H., Lee, M. C., Chiu, T. K.: Personalized Learning Objects Recommendation based on the Semantic-Aware Discovery and the Learner Preference Pattern. Educational Technology & Society 10, 84-105 (2007)
55. Wiley, D. A.: Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. In: Wiley, D. A. (ed) The Instructional Use of Learning Objects (2001)
56. Ye, Y.: Supporting Component-Based Software Development with Active Component Repository SystemsDepartment of Computer Science. University of Colorado, Boulder, Co. (2001)

Weiqin Chen is an Associate Professor at the Department of Computer Science, Oslo and Akershus University College of Applied Sciences, Norway. She is an Adjunct Associate Professor at the Department of Information Science and Media Studies, University of Bergen, Norway. She received her PhD. in Computer Science from the Chinese Academy of Sciences in 1997. Her current research focuses on intelligent support for collaborative learning and work.

Richard Persen received his Master degree in 2008 at the Department of Information Science and Media Studies, University of Bergen, Norway. He is currently working as a system developer at the NorgesGruppen ASA.

Received: November 29, 2011; Accepted: December 29, 2011.

DOI: 10.2298/CSIS111226016S

A Model-based Approach for Assessment and Motivation

J. Michael Spector[1] and ChanMin Kim[2]

[1]University of North Texas
mike.spector@unt.edu
[2]University of Georgia
chanmin@uga.edu

Abstract. Representations support learning and instruction in many ways. Two key aspects of representations are discussed in this paper. First we briefly review the research literature about cognition and processing internal mental models. The emphasis is on the role that mental models play in critical reasoning and problem solving. We then present a theoretically-grounded rationale for taking internal mental representations into account when designing and implementing support for learning. The emphasis is on interaction with meaningful problems. Lastly, we present research that has led to a conceptual framework for integrating models into learning environments that includes technologies for formative assessment and motivation.

Keywords: assessment; complex problem solving; critcal reasoning; mental model; model-facilitated learning; motivation; problem conceptualization; problem representation

1. Introduction[1]

Cognition is both complex and multifaceted. Remembering and misremembering past events, recognizing family and friends, planning vacations, shopping for gifts, and solving puzzles are cognitive processes that are generally taken for granted yet surprisingly complex. It is when we become aware of an error that we are inclined to reflect on the relevant cognitive processes involved. "Why did I mistake that person for someone else? How can I avoid repeating that mistake?" Such thoughts about our

[1] Much of the work represented herein has been inspired by the work of David H. Jonassen. Several of his publications in this area are cited, but his influence extends far beyond his published work cited in this paper. See also the Spector & Park chapter in *The Role of Criticism in Understanding Problem Solving* by Fee & Belland (2012).

cognitive processes are not so uncommon and fall into the domain of metacognition – thinking about thinking, so to speak.

Most persons have a natural desire to be correct (most of the time, anyway). Those with a desire to understand such errors might investigate how knowledge is developed. Epistemologists have investigated such questions for centuries [99]. Psychologists have also investigated the nature of cognition, including critical reasoning and problem solving skill development. Many people engage in metacognitive activities and think about their own reasoning processes from time to time. Such self-reflection should be encouraged and scaffolded during learning. The focus here is on the role that externally created representations or models can play in support of formative assessment and motivation. The approach is naturalistic [98, 99] as we examine how individuals think and develop reasoning skills and expertise. A naturalistic approach takes into consideration both cognitive and non-cognitive aspects that occur in actual cases, including beliefs about the domain being investigated as well as beliefs about one’s own ability to successfully solve problems. It is generally agreed that different kinds of problems require different kinds of reasoning and, as a consequence, different kinds of support for learning [45, 46, 47, 48, 106]. A deep understanding of reasoning and problem solving can provide the basis for personalized and meaningful feedback on specific problem solving activities, including support for motivational aspects that might affect performance [54].

2. Mental Models and Reasoning

To motivate the discussion let us consider a representative problem-solving situation that many have experienced – shopping for a gift. Here is a representative scenario:

Eli and Naomi recently received news from college friends, Sam and Lisa, whom they have not seen for several years, but who happens to be coming to town on their honeymoon. Upon hearing about the upcoming visit, Naomi says to Eli, “well, after more than five years together they finally got married!” and she suggests that they need to find a wedding gift. Eli and Naomi look at their bank account to see what they can afford, and then they start discussing what they might find within their price constraints. This discussion involves recalling many Sam and Lisa stories from their past and those stories begin to lead to gift possibilities.

The point of this scenario is to suggest that (a) there are both cognitive (e.g., planning) and non-cognitive (e.g., emotional) aspects to many problems and decisions, (b) many aspects of the situation are related and likely to be mutually influential (e.g., college memories and gift possibilities), (c) previous experience with similar problems is relevant (e.g., gifts they received when they got married and wedding gifts they have given to others), and, (d) alternative gift possibilities obviously exist.. These four kinds of considerations exist in many different problem solving situations and are likely to influence one's reasoning about a solution. Addressing such considerations often requires reflection on the nature of the problem and some deliberation about alternative means to resolve the situation.

2.1. The nature of reasoning

What is the nature of reasoning? Reasoning is a deliberative, goal-driven activity. A traditional definition of reasoning [25] includes such things as:

- Having an identifiable goal or purpose that one wants to attain; and,
- Being able to identify alternative means of achieving the goal and then analyzing which means might be more feasible or optimal in attaining the goal according to one or more criteria.

The second characteristic is often more challenging than the first. Goals arise frequently in every aspect of life. Some are short-term – "I want fish for dinner"– and some are longer-term – "When I grow up I want to be an educational researcher." Sometimes we decide on the means of attaining the goal without considering alternatives. Analyzing alternatives involves (a) identifying alternative ways of attaining a goal, (b) determining desirable and undesirable aspects among the alternatives, (c) prioritiziong alternatives, and (d) selecting the most desirable alternative using relevant criteria [97]. Reasoning typically involves deliberative thinking, including the consideration and evaluation of reasonable alternatives to reach the goal. Reasoning, being rational and becoming skilled in solving problems requires an ability to confront and embrace complexity, including the assumptions we make when conceptualizing the problem situation and the implicaitons we might associate with alternative solutions. The requirement to embrace complexity runs counter to a natural human tendency to simplify and avoid challenging problem-solving tasks [97].

One might characterize deliberation about assumptions and implications as a reflective process – reflecting on the quality of one's reasoning [18]. Deliberation and reflection are desirable aspects in solving complex problems, but they require time and effort. Devoting time and effort to an activity is an issue that also involves motivation (a desire to succeed) and volition (following through to achieve one's goal) [54, 55, 56]. Suppose that one decides to examine the consequences and implications of one way to reach a particular goal. For example, suppose that one really wants to become an educational researcher (goal), and, further, that understanding advanced statistical methodologies (means) is required to be successful. The volitional aspect involves active follow-through; the implication of accepting the goal is that one must take graduate courses in advanced statistical methodologies that require some mathematical abilility. One may believe that one is weak in mathematics and feel internal resistance to the goal of becoming a serious educational researcher. How might one overcome such internal resistance and succeed?

How might educational support be provided to help one overcome such internal feelings?

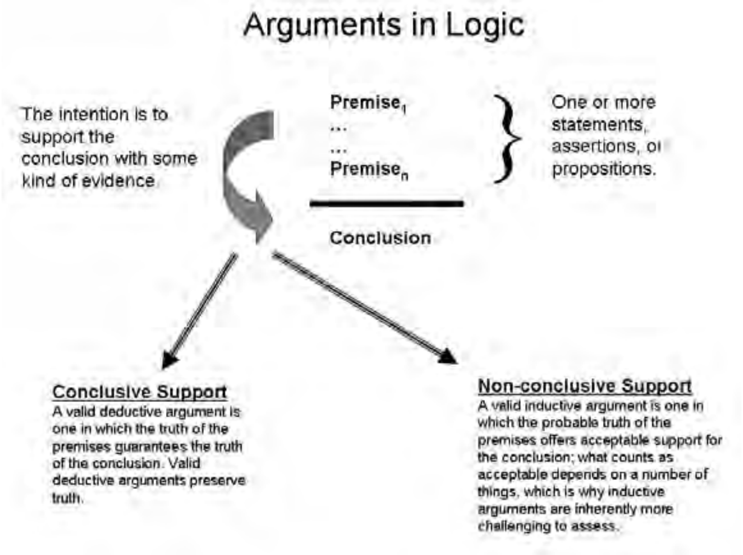
For now we simply want to suggest that the process of deliberation might be extended to include reflections about one's underlying assumptions as well as the consequences and implications of pursuing one course of action over another, including how one feels about those consequences and implications. We ought to recognize that emotions, habits and preferences established over a period of many years influence the processes of examining goals, assumptions, alternative solutions and consequences [59]. Perhaps we are only intermittently rational. Non-cognitive aspects of being who we are play a greater role in reasoning and problem solving than we may be inclined to believe. We proceed on the assumption that non-cognitive aspects influence decision making, problem solving and reasoning in general [54, 55, 56, 58]. Rather than consider these non-cognitive aspects of reasoning as limitations, a naturalistic approach accepts the influence of non-cognitive factors and attempts to take them into account when supporting learning. A virtual change agent to support motivation and volition is one means of implementing this naturalistic approach [55].

Reasoning has been analyzed in a number of ways by both philosophers and psychologists [6, 18, 38, 40, 85, 87, 91]. There are several traditional distinctions to consider. One is between formal and informal reasoning. A second is between deductive and inductive (non-deductive) reasoning, which is associated with logic. In order to create an instructional framework to improve critical reasoning skills, it is necessary to understand general reasoning processes, and argumentation is central to the reasoning process.

An argument can be defined as a collection of statements some of which are offered in support of another. The supporting statements are called premises and the statement being supported is called the conclusion. The landscape of logic is comprised of arguments, each of which consists of statements offered in support of another statement. Logic involves determining which kinds of arguments are good and in identifying deficiencies in specific arguments, which is needed in order to help improve reasoning skills [18, 23, 24, 25].

Like other enterprises involving quality determinations, standards or criteria are needed. In order to develop appropriate standards, one needs to consider the different kinds of arguments that one wishes to sort. If one were sorting apples, for example, it would be relevant to know if both red delicious and golden delicious apples were in the batch to be sorted – knowing this allows appropriate color and other criteria to be used.

Since the nature of an argument is to offer statements intended to support another, one place to start is to determine the *kinds of support* that might be offered. Two kinds come to mind immediately: conclusive support and non-conclusive (suggestive) support. The nice thing about this distinction is that it is all inclusive and the two categories are mutually exclusive. Moreover, such a distinction allows for correction of the type of support offered – that is to say that the person presenting the argument may believe that he or she is offering conclusive support when only suggestive support has been provided. Before evaluating the adequacy of the support actually offered, one ought to be sure the argument is located in the proper category since the evaluation standards are different. An example of this problem can be found at the end of a correlation study in which the researcher claims that the study *proves* a certain point. Proof is a strong notion that is best reserved for the strongest types of argument, which are generally considered to be of the conclusive variety (deductive arguments). Correlation studies may s*uggest* or *show* something, but they do not establish conclusions with certainty. Even rigorous controlled experimental studies aimed at establishing causality do not establish conclusions with certainty, although they may establish conclusions with very high degrees of probability.



The intention is to support the conclusion with some kind of evidence.

One or more statements, assertions, or propositions.

A valid deductive argument is one in which the truth of the premises guarantees the truth of the conclusion. Valid deductive arguments preserve truth.

A valid inductive argument is one in which the probable truth of the premises offers acceptable support for the conclusion; what counts as acceptable depends on a number of things, which is why inductive arguments are inherently more challenging to assess.

Fig. 1. The logical form of arguments.

Mathematical proofs are familiar examples of the deductive variety. Statistical studies are familiar examples of the inductive or non-deductive variety. What is worth carrying forward is that, when reflecting on one's reasoning, understanding the kind of argumentation involved (deductive or inductive) is important. Moreover, when deliberating on the quality of an argument, a critical factor is in determining whether or not adequate evidence has been developed to support the conclusion or decision to be taken.

Finally, we should not forget that non-cognitive aspects influence our reasoning, decision making and problem solving. In a model-based discussion of reasoning, there is a place to include these non-cognitive aspects of reasoning, and they can be given prominence in a formal deductive or inductive argument as well. For example, in a deductive argument, one can provide a statement representing an affective aspect of a problem, such as "If we want to celebrate when Sam and Lisa first met at that Bob Dylan concert, then we should find a gift related to that first date." Likewise, one can introduce perceptions and preferences and affective factors into an inductive argument based on survey and interview results, as frequently happens in educational research. This by no means is a claim that all aspects of non-cognitive factors can be given an appropriate representation in language; in fact, the process of representing affective aspects may well alter the underlying emotions and feelings and is a deviation from a completely naturalistic approach. However, some affective factors can and should be represented in language as they often have relevance for assessing how learners think about a problem; they may also be useful in providing learners with feedback to improve problem solving performance, enhance motivation, promote meaningful follow-through, or perhaps change epistemic beliefs relevant to learning progress.

There is another distinction to introduce – namely, the distinction between formal and informal reasoning. Both formal deductive and inductive logics have informal counterparts. Language influences thought. Language facilitates thinking. We make use of language to express and represent our deliberations, reflections and to explain our actions. While there are several kinds of deductive logic – multi-valued logic, first-order predicate calculus, modal logic, and so on – the simplicity of a two-valued propositional logic can also facilitate thinking and serve to improve the quality of deliberations and reflections. The purpose here is to clarify the relationships between reasoning and mental models, so we next turn to internal representations that are involved in our reasoning. These internal representations can be grouped together using the term ‘mental models’.

2.2. Mental models and schemas

Our interest in mental models is driven by a concern to contribute to the improvement of human reasoning. It is clear that people do not generally reason in terms of formal logic [23, 24, 26, 27, 28]. We have also said that we believe that people are intermittently rational. We occasionally engage in deliberative and reflective processes – cognitive processes – in order to achieve a goal or fulfill a purpose. Those internal cognitive processes are hidden; they are not viewed directly and immediately by anyone; mental models are hypothetical entities. We make inferences about these hidden models based on external representations and observable entities and events, such as things that people say or write, diagrams that people create, actions that people take, and solutions to problems that people solve. It is generally believed that the quality of one’s internal mental models influences the quality of one’s ability to solve problems and reason critically. In the *Tracatus Logico-Philosophicus*, Wittgenstein said that *we picture facts to ourselves* (Remark 2.1) [110]. We naturally construct mental representations of the things we experience. When the things we experience are new or bewildering, we seem to struggle with those internal representations – we may begin to externalize them in our efforts to make sense of what is not immediately clear or easily understood. We often use language for this purpose – the ability to create internal representations of things we experience is truly amazing. Equally amazing is the ability to externalize those representations and talk about what we are thinking, occasionally even sensibly (see Wittgenstein's discussion of language games in *Philosophical Investigations*) [111].

We create internal representations and we interpret those internal representations by talking (or writing or drawing diagrams), creating new re-representations [16, 17, 20, 32, 36, 37, 42, 61, 63, 64, 65, 66, 68, 69, 72, 80]. Claims about internal mental processes and entities are inherently inferential – we do not observe any internal mental objects directly, not even our own. Conclusions about mental models and associated cognitive processes are, as a consequence, at best more or less probable. As a consequence, we ought to be modest in what we claim to know about mental models and cognitive processes (recall the discussion of inductive reasoning).

This word of caution does not mean that mental model research is not worthwhile or that one cannot investigate mental processes in a rigorous scientific manner. There can be and is a science of mental models, just as there can be and is a science called psychology. Indeed, one might argue that mental models are at the core of modern cognitive psychology. Just because that which is being investigated is a hypothetical entity or hypothetical process does not mean that those objects cannot be investigated. Indeed, mental models and cognitive processes are of interest precisely because of their potential to explain a great many human phenomena and behavior patterns of interest to psychologists, educators, instructional designers, parents, and people in general. We want to understand who we are, why we think and act the way we do, why we make the mistakes the way we do, why we are intermittently rational, and how we might become better decision makers and problem solvers.

Cognitive psychologists are in general agreement that people have the ability to process a variety of different information and act appropriately in many different situations. These abilities were mentioned earlier and include perception, pattern recognition, storing and retrieving different kinds of information, and acting on previously gained experience [2, 81, 85, 86, 87, 88, 89, 91, 92, 108, 109, 112]. Pattern matching is an ability at which humans excel [89]. Humans quickly and effortlessly recognize objects, as shown by the ability of most people to recognize a familiar face in a group of people. Pattern recognition is a critical cognitive process that involves perceiving, remembering, and interpreting, which are all essential in many decision-making and problem-solving situations [8, 11, 19, 29, 30, 34, 35, 36, 41, 52, 68, 76, 77, 78, 79, 82, 83, 85, 86].

Pattern matching and recognition are generally believed to be based on schema, which are well-established cognitive artifacts stored in long-term memory based on past experience [4, 88, 89]. Moreover, it is not simply patterns of familiar objects that people can recognize and match with prior experience to help select appropriate actions. People also recognize general situations that call for particular kinds of responses, such as the way that a restaurant is organized with a host who seats customers, with a different person taking orders, and possibly with others bringing drinks and clearing the table. Because such a situation has been experienced many times, one generally knows what to do in a restaurant even though it may be the first visit to that particular restaurant. Schank and Abelson [90] and others [1, 2, 3] refer to the ability to recognize such general situations and respond appropriately as involving *scripts* (a specific kind of internal mental strucrure). In what follows, the more general term 'schema' will be used to refer to specific cases involving previous experience in that situation as well as general cases involving unfamiliar objects set in a familiar kind of situation. Schemas involve somewhat complex but well-established internal representations that enable a person to respond imemdiately in a particular situation [1, 2, 3, 4, 5, 6, 7, 8, 11, 14]. It is often the case that a person may have difficulty in explaining why he or she acted automatically in a particular situation, which is evidence that the internal representation is so well established and so automated that little or no conscious thought is devoted to retrieving and activating the schema [9, 10]. With regard to reflecting on one's assumptions in a deliberative reasoning process, when schemas are involved it is sometimes a challenge to bring all relevant assumptions into consideration. It should be obvious that a schema necessarily involves a simplification of the actual situation in which it is invoked [13, 14, 16, 17]. Only a few key aspects of the situation are needed to activate a schema, such as a host or hostess at the front of the restaurant for seating or a receipt left inside a holder on a tray for payment.

Many problems do not lend themselves to resolution based solely on schema and well-established prior experience. According to cognitive psychologists [33, 43, 44], humans have another way to resolve unfamiliar and puzzling problems – namely by creating internal models of the situation and using those internal representations to think through to a solution. These models are created just when needed and are typically called mental models [43, 44, 94]. Mental models are generated to represent the perceived structure of a puzzling or new phenomenon. These mental models are not and could not be replicas of the world. Like schema, a mental model is necessarily a simplification. Such simplifications are useful in helping a person understand an unfamiliar situation or puzzling phenomenon. Recognizing relevant aspects of the problematic situation are critical for the development of useful mental models; this will be taken up again in a subsequent section involving the implications of internal representations for the design of effective learning and problem-solving support.

Mental models and schemas are internal constructions that enable a person to confront a problem or situation and act in a reasonably appropriate manner [3, 93, 94, 95]. Neither mental models nor schema are directly observed. However, one can elicit a representation of a mental model, and with perhaps more effort, a representation of a schema. These re-representations are often useful in diagnosing miscues and misunderstandings and can be useful in helping a person develop expertise. Such re-representations form the basis for formative feedback and motivational/volitional support in a framework we call model-faclitated learning [75].

To conclude this section on mental models and schemas, it is worth noting that these two categories of internal representations are related and interact in many problem solving situations. A person confronts a problem or situation and creates one or more internal representations to respond to the situation. Domain-specific prior knowledge, established schemas and newly created mental models may all be brought to bear in this process of understanding the problem at hand. The main purpose of this process is to create a causal explanation of the puzzling problem or phenomenon. As Kant [51] noted *in The Critique of Pure Reason*, it is natural and unavoidable for people to think in terms of cause and effect, just as space and time are added to our experience of the world. The reasoning process of invoking schemas and integrating newly constructed mental models fits well with Piaget's [83] epistemology. When a schema is invoked and a newly constructed mental model can be fit easily within that schema, one might say that assimilation has occurred. When no existing schema is found to help resolve the situation, a newly constructed mental model might be said to lead to a process of accommodation, which involves a refinement of an existing schema. One might also introduce the notion of an internal cognitive structure which is akin to a repository of schemas. Humans are generally very good at modeling their experiences. We can anticipate new states of affairs and predict likely outcomes of existing states of affairs with relative ease, thanks to our ability to create and manipulate internal representations (mental models and schemas). Figure 2 is a representation of how mental models and schemas might interact over time as a person builds competence and expertise.

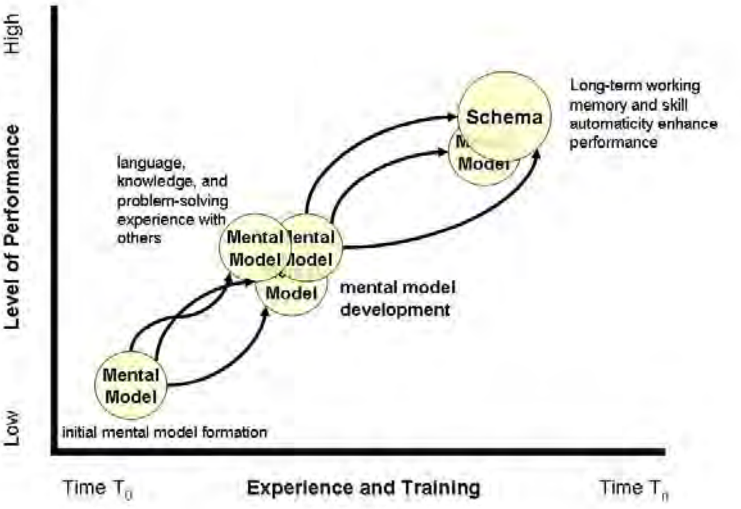

High
Level of Performance
Low
Schema
Model
Long-term working memory and skill automaticity enhance performance
language, knowledge, and problem-solving experience with others
Mental Model
Mental Model
Model
mental model development
Mental Model
initial mental model formation
Time T_0
Experience and Training
Time T_n

Fig. 2. Mental models, schemas, and expertise development

Figure 2 suggests that over time, mental models are constructed and become associated with other mental models and eventually with schemas, which may be modified in order to take into account what has been learned from activation of the mental models and schemas. Based on this view of human reasoning processes it is possible to discuss implicaitons for learning and instruction as the basis for a general model-based framework to integrate external models into learning, performance support, formative feedback, motivational support, and instruction. The general strategy is to elicit representations of internal models and then provide appropriate support, based in part on a comparison with models elicted from highly experienced problem solvers [84].

It is possible to distinguish two different kinds of mental models: *perceptual models* and *thought models* [102]. Glaser, Lesgold, and Lajoie [31] and Johnson-Laird [43] consider perceptual models to be *appearance* or *structural models* that are used to represent an external reality. This kind of model serves to mediate between internal visual images and external representations in the form of statements, whereas *thought models* also include qualitative processes and inductions which support the construction of artifacts to represent complexity and causal relationships. In any case, the point here is that there is an interaction between the construction of internal models and external representations of those models; that interaction will be taken up in the framework to be presented subsequently [29, 30, 33, 74, 75].

3. Implications for Learning and Instruction

Researchers are typically familiar with and concerned about ecological validity, which is the degree to which an experimental situation reflects a naturally occurring real-world situation. Instructional designers have a related principle that suggests that primary instruction and practice should involve problems and situations that closely resemble problems and situations that are likely to be encountered subsequent to instruction. Instruction itself as well as research designs should have a high degree of ecological validity. With regard to instruction, understanding how people process information and reason about problems is a critical aspect of instructional ecological validity. Instruction that ignores how people process information and reason about problems is less likely to be effective [2, 31, 48, 74]. Many examples can be cited to support this basic instructional design principle. The limitations of working memory are well established. When an instructional system or learning environment violates those limitations by presenting too much information all at once with too little learning support, cognitive overload results, learning outcomes become suboptimal, and many learners become frustrated [104] The major implication from the previous discussion for learning and instruction, therefore, is that mental models and schemas should be taken into account when designing instruction and implementing support for learning and performance.

While this implication for designing learning support may seem obvious to many, there is the subsequent challenge to be more specific about all that is implied in such a principle. How can learning designs take into account the hidden processes of constructing mental models, activating schemas, and developing ever more useful and productive internal representations? If one is only concerned about performance and learning outcomes, then perhaps such questions are not a pressing concern. However, if one believes that the development of robust and flexible internal representations is critical for the development of competence and expertise, then such questions become a foundation concern for instructional design, which is what is being proposed in this paper.

When learners are in an instructional situation, they will naturally be engaged in activating schemas and constructing mental models. Another way to think about the implications for the design of effective instruction is through the lens of cognitive efficiency, which can be defined as the optimal effort required to solve a problem correctly or perform a task in a satisfactory manner [39]. The notion of cognitive efficiency involves a tradeoff between time, resources and desired outcomes. There are three primary measures of cognitive efficiency: instructional efficiency, processing efficiency, and outcomes efficiency [39]. The first and third are already familiar to most instructional designers who are concerned that learners achieve acceptable performance in a reasonable amount of time. Processign efficiency is most directly related to the reasoning processes discussed previously and clearly influences both instructional and outcomes efficiency. Measures of cognitive efficiency and the means to enhance it are well known to instructional designers. There are two factors that have yet to be fully explored in the research literature on cognitive efficiency and instructional design research: (1) individual differences that impact internal cognitive processes, and (2) how cognitive efficiency improvements and other instructional design principles can be effectively applied to situations involving complex and ill-structured problem-solving tasks. These considerations are likely to be imporant challenges for future research (see the last section of this paper).

While there is much that is unknown and difficult to determine with regard to how individuals process information, solve problems, and develop competence in reasoning about challenging situations, some initial steps have been taken. First, instruction should be centered around meaningful and realistic problems [48, 67, 74]. Second, the analysis of problems and tasks reveal that there are different kinds, which require different kinds of learning support [45, 46, 47, 48]. Third, representations of internal mental models and schemas can be used explicitly to support learning [84, 98]. Fourth, when learners are struggling, there are often wrong-head beliefs (e.g., “I did not inherit the math gene”), motivational issues (e.g., “this lesson is terribly boring”), and volitional challenges (e.g., “while I would like to stay and understand this material, I would also like to take a break and go see a movie” [57]. Models play a central role in representing all of these aspects involved in a learning situation.

Before presenting a general framework to support model-based reasoning, a discussion of the different kinds of models that might be used in alignment with internal models is relevant. Just as there are different kinds of internal representations, there are a variety of different kinds of external representations or models. External models come in many forms. Some are mathematical in nature, such as a regression model. Some are graphical, such as a schematic drawing. Some are in the form of arguments with premises and a conclusion. Some models involve video or animation depicting how a device works. Some involve a combination of different forms of representation. Which are effective when, for whom, and why?

One temptation is to think at the level of learning styles and preferences based on a belief that those individual differences are key factors in learning effectiveness. Certainly learning styles and preferences are often relevant, but they do not reach a level that aligns easily with how a person processes information. Some learners may say that they are visual learners, for example. They may even have some evidence for such a claim. However, this does not resolve the issue of which visual models will be easily aligned with a particular learner's internal representations and, as a consequence, be likely to help that learner with particular content to be learned. On the other hand, visual learners are likely to respond well to visually oriented motivational and volitional messages [55].

An established approach to support learning in complex domains is to have a general sense of common problems encountered for different tasks. A cognitive task analysis can be a start along these lines, and cognitive task analysis has proven useful for the design of instruction. Those who have implemented intelligent tutoring systems have developed libraries of commonly encountered mistakes and misunderstandings that can be used to inform an instructional decision such as selecting an appropriate learning activity or providing targeted feedback [84, 100]. Dörner [21, 22] reports that those confronting complex and ill-structured problems experience a number of challenges: (a) a failure to grasp the full breadth of a problem situation with a tendency to focus on just one component or familiar aspect of the problem, (b) the inability to reason effectively about non-linear relationships among different aspects of the problem, and (c) difficulty in understanding and predicting delayed effects and the accumulation of effects over time.

In addition, one can elicit a learner's representation or conceptualization of the problem space (a re-representation of mental models and schemas) and use that to assess of progress of learning (e.g., indicate how well it matches an expert's representation), provide formative feedback (e.g., suggest missing factors or relationships among problem components), and support motivation and volition (e.g., display successful application in a situation likely to be meaningful to the learner) [57,75, 84, 100].

The question of whether and how designers and instructors can influence model-building activities in learners is a core educational concern [52]. According to Johnson-Laird [44] and other authors there are several sources of mental models: (1) the learner's ability to construct models in an inductive manner; (2) everyday observations of the outside world combined with the adaptation of familiar and generally recognized models; and (3) the explanations, representations and examples provided by others. All of these sources are relevant for model-based instruction. In the framework we propose, the learner should be encouraged to construct an initial model and reflect on that model based on his/her prior experience and with feedback from peers (#1). In addition, a reference model should be constructed based on an instructor's or an expert's knowledge and, along with models of other students, the reference model can be used to encourage targeted reflective analysis (#2 and #3).

According to Carlson [12], it is possible to design instruction to involve the learner in a process of inquiry in which facts are gathered from data sources, similarities and differences among facts are noted, and concepts developed. In this process, the instructional program serves as a facilitator of learning for students who are working to develop their own answers to questions. On the other hand, instructional programs can present clearly defined concepts followed by clear examples, including a generalized reference model. A designed conceptual model may be presented ahead of the learning tasks in order to direct the learner's comprehension of the learning material. More generally, we can distinguish between different paradigms of model-oriented instruction depending on whether they aim at (a) self-organized discovery and exploratory learning, (b) guided discovery learning, or (c) learning oriented toward the imitation of an expert's behavior or the adaptation of teachers' explanations.

In the next section, a framework that integrates aspects of all three of these approaches into a general framework will be presented. This framework should be considered provisional. Research to explore this and other possibilities will conclude the discussion.

4. A Framework for Integrating Models in Learning and Instruction

A well-established instructional design framework that has wide applicability is cognitive apprenticeship [15]. The cognitive apprenticeship model involves six different methods and the notion that learners new to a domain require more support than more experienced learners. The six methods are: (a) modeling (e.g., show how an experience person solves the problem), (b) coaching (e.g., observe performance and provide timely and constructive feedback), (c) scaffolding (e.g., implementing explicit support to facilitate learners' problem solving), (d) articulation (e.g., getting learners to talk about how they are thinking about solving a problem), (e) reflection (e.g., encouraging learners to compare their solution with that of others), and (f) exploration (e.g., allowing learners to investigate new problems and problem approaches on their own with little or no guidance).

The cognitive apprenticeship model is consistent with Gagné's nine events of instruction (a claim with which some will disagree) as well as with Merrill's

[74] first principles of instruction, just as the relatively new notion of cognitive efficiency is consistent with a great deal of traditional instructional design. Essentially, cognitive apprenticehship can be characterized as a significant and explicit cognitive extensive of earlier instructional design models. What is really new and only recently emerging is the recognition that it is important to take into account how individuals think and feel about complex problems and process information in the course of solving problems. In any case, there is a great deal of evidence that cognitive apprenticeship is a useful instructional design model [15]. The next step is to refine that model to explicitly account for internal reasoning processes.

Model-facilitated learning (MFL) is an instructional design approach aimed explicitly at promoting model-based reasoning. MFL builds on cognitive apprenticeship [15] and Merrill's [74] first principles. MFL is centered around and facilitated by models in the form of expert and student representations of a problem or problem space, a solution approach, and/or a solution. The models may or may not be created by learners, but learner interaction with models is generally an integral aspect of learning activities.

The particular area for which model facilitated learning was designed involves complex and challenging learning tasks and problem-solving situations. Complex learning tasks tend to have many interacting components, some of which may be incompletely defined, and with some non-linear relationships and delayed interactions among the various components [21, 22, 103]. Such problems occur in economic forecasting, engineering design, environmental planning, management decision making and in many other every day problem-solving situations. Using models of complex phenomena to help learners gain a holistic and meaningful sense of the problem is one aspect of model facilitated learning. Having learners engage in modeling activities to gain insight into the complexity of a problem situation is a second aspect of MFL. MFL assumes three stages of learning development and has associated instructional guidelines for each stage [75]. The first stage is problem orientation in which problems or related sets of problems are presented to learners and learners are asked to solve relatively simple versions. The second stage of learner development involves inquiry exploration in which learners are challenged to explore a complex task domain and asked to identify and elaborate the relationships among the various components of the problem. The third stage of learner development involves policy development in which learners are asked to reason in a more global and holistic with regard to rules and heuristics to guide decision making with regard to various problem situations that may arise in that task domain. Principles to guide the elaboration of learning activities and instructional sequences within these stages include such notions as (a) situating the learning experience in the context of meaningful and realistic problems [74], (b) presenting problems of increasing complexity, involving learners in a sequence of related tasks involving the initial problem scenario [106], (c) involving learning in an increasingly set of complex inquiries and explorations with regard to the problem situation, and (d) challenging learners to develop rules and guidelines to guide decision making in anticipated problematic situations.

The foundations for model facilitated learning are derived from system dynamics [103], educational and learning psychology [67, 101], and from instructional design [74]. In addition, MFL adopts the principle of graduated complexity [75] in the form of guidance for the elaboration of instructional sequences. Graduated complexity in MFL is implemented consistent with cognitive apprenticeship methods [15]. According to the principle of graduated complexity, instructional sequences should progressively challenge learners to:

1. Characterize the representative behavior of a complex system, indicating how it behaves over time – guided discovery learning.
2. Identify a desired outcome and key variables and points of leverage with respect attaining that outcome – exploratory learning.
3. Identify and explain alternative causes for observed phenomena – deliberation in the context of problem solving.
4. Reflect on how the system and associated variables seem to change over time and through interventions; this challenge requires perceptual processing but the critical aspect is the ability to focus on key problem components and how they change over time and with intervention; this is also consistent with the reflection method in cognitive apprenticeship – reflective reasoning.
5. Develop a rationale to explain complex phenomena in terms of an underlying system stucture, including decision-making and policy formulation guidelines; this challenge is aligned with both articulation and reflection in the cognitive apprenticeship model – a high level of reflective reasoning.
6. Broaden understanding through diverse and new problem situations; this challenge addresses near and far transfer of learning, is consistent with Gagné's (1985) ninth event of instruction as well as with the exploration method in cognitive apprenticeship – more experiential learning.

In addition to making use of elicited representations of problems to support graduated complexity with an MFL context, models are also useful in designing and implementing virtual change agents [57]. Specifically, elicited models, combined with an analysis of the problem as perceived by experienced and successfuol problem solvers, can indicate particular meanings and causes introduced into the problem situation by the learner that may not be productive in finding a solution and that may well lead a learner to lose interest or quit working on the problem. For example, an elicited student's model may reveal a feeling of confusion due to having encountered a problem far beyond that student's perceived capacity (e.g., "I am embarrased to say that do not see how to perform the steps to reach the goal"), or a student may reveal in the course of creating an external model frustration or other distractors (e.g., "I wish the system would just tell me what to do next"), or perhaps a student's model may include emotional descriptors for some factors revealing feelings that could impede progress (e.g., "the stupid bureaucracy makes it difficult to see how to get from point A to point B in this problem"). In summary, the ability to identify relevant motivational and volitional factors is just as important as identifying problem missteps and miscues in the course of providing timely, informative, and meaningful feedback to students.

5. Further Research

The ability to solve complex problems depends in large part on the ability of individuals to overcome motivational and emotional resistance to explore a challenging problem situation and then construct productive mental models and activate relevant schemas. Mental models are useful in identifying individual resistance to problems and explaining puzzling phenomena and complex systems, especially in terms of cause and effect [35, 93]. Those who are confronting challenging problems are likely to avoid the problem (which can be detected and addressed through the use of external representations) or else construct mental models in order to provide explanations for the new or unusual phenomena. Greeno [33] suggested that productive mental models should be encouraged along with the significant properties of external situations and appropriate interactions; this principle has strong implicaitons for the design of learning and instruction. Moreover, such a principle is consistent with a constructivist approach to learning that suggests that the learning environment should serve as an information resource which can be used by a learner to focus on relevant aspects of a problem and activate relevant schemas and prior knowledge. Learning activities are, consequently, required that enable learners to explore and interact with the learning environment in a meaningful, problem-centered context. Consistent with cognitive apprenticeship and model-facilitated learning, support and scaffolding should be provided to help problem solvers to be successful and develop both competence and confidence [53, 59]. This general approach can also be found in what is currently being called *learning by design* [50, 62]. Additionally, Kirschner and colleagues [60] and Mayer [73] have argued that minimal guidance during instruction, especially with learners new to a problem-solving domain, is not effective. Consistent with cognitive apprenticeship and model-facilitated learning, explicit coaching and overt learning support should be provided to those inexperienced in the domain.

What is not known or well established is how best to support the development of expertise and insight with regard to complex, problem solving activities in specific problem domains. How well instruction created in accordance with the principles of MFL works in terms of developing competence and expdertise, especially in comparison with other instructional methodologies, has not been established. Which kinds of models (student-created, expert-created, partially complete, etc.) are effective with different learners and learning tasks is also not well known, nor is it well known how external models align with internal models in specific problem solving situations. In addition, the efficacy of using model representations to identify and address motivational, volitional and emotional issues has not been sufficiently explored in the research literature.

While versions of MFL have been implemented and evaluated in the first two stages indicated above (problem orientation and inquiry exploration), very few MFL environments exist to promote learning at the last stage of learner development (policy development). As a result, research on effective MFL techniques to promote policy development knowledge remains very open for further research and development, and additional research is needed in the first two stages as well. Additionally, effective MFL instructional sequences for complex problem task domains is not very well established. A central underlying problem concerns the need for well-developed means to assess the progressive development of student understanding in complex task domains. This requires validated means to elicit and evaluate student generated models in response to a wide variety of problem types and scenarios, yet those means are still in the early stages of development. Finally, integrating external models of various kinds and aligned those with individually generated internal models remains an important area open for further exploration.

References

1. Al-Diban, S., & Seel, N. M. (1999). Evaluation als Forschungsaufgabe von Instruktionsdesign – Dargestellt am Beispiel eines multimedialen Lehrprogramms. *Unterrichtswissenschaft*, *27* (1), 29-60.
2. Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
3. Anzai, Y., & Yokoyama, T. (1984). Internal models in physics problem solving. *Cognition and Instruction*, *1*, 397-450.
4. Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, UK: Cambridge University Press.
5. Bhatta, S. R., & Goel, A. (1997). Learning generic mechanisms for innovative strategies in adaptive design. *The Journal of the Learning Sciences*, *6* (4), 367-396.
6. Braine, M. D. S. (1990). The natural logic approach to reasoning. In W.Overton (Ed.), Reasoning, necessity, and logic: Developmental perspectives (pp. 133-157). Hillsdale, NJ. Lawrence Erlbaum Assoc.
7. Braine, M. D. S., & O'Brien, D. P. (Eds.) (1998). *Mental logic*. Mahwah, NJ: Lawrence Erlbaum Assoc.
8. Briggs, P. (1990). The role of the user model in learning as an internally and externally directed activity. In D. Ackermann, & M.J. Tauber (Eds.), *Mental models and human-computer interaction 1* (pp. 195-208). Amsterdam: Elsevier Publ.
9. Brown, D. E., & Clement, J. (1989). Overcoming misconceptions via analogical reasoning: abstract transfer versus explanatory model construction. *Instructional Science*, *18*, 237-261.
10. Bruner, J. S. (1966). *Toward a theory of instruction*. Cambridge, MA: Harvard University Press.
11. Buckley, B. C., & Boulter, C. J. (1999). Analysis of representation in model-based teaching and learning in science. In R. Paton & I. Neilsen (Eds.), *Visual representations and interpretations*. (pp. 289-294). London: Springer.
12. Carlson, H. L. (1991). Learning style and program design in interactive multimedia. *Educational Technology Research and Development*, *39* (3), 41-48.
13. Casey, C. (1996). Incorporating cognitive apprenticeship in multi-media. *Educational Technology Research and Development*, *44* (1), 71-84.
14. Cheng, P. W., & Holyoak, K.J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391-426.

15.Collins, A., Brown, J.S. & Newman, S.E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L.B. Resnick (Ed.) (1989), *Knowing, learning, and instruction* (pp. 453-494). Hillsdale, NJ: Erlbaum.

*16.*Cornelissen, J. P. (2006). Making Sense of Theory Construction: Metaphor and Disciplined Imagination *Organization Studies, 27, 1579 - 1597.*

17.Darabi, A.A., Nelson, D.W., & Seel, N.M. (i.pr.). A dynamic mental model approach to examine schema development of a troubleshooting skill: Delayed retention of mental models as a test of schema development. *Technology, Instruction, Cognition, and Learning*.

18.Dewey, J. (1910). How we think. New York: D. C. Heath.

19.Doerr, H.M. (1996). Integrating the study of trigonometry, vectors, and force through modeling. *School Science and Mathematics*, *96* (8), 407-418.

20.Dole, J.A., & Sinatra, G.M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist*, *33* (2/3), 109-128.

21.Dörner, D. (1976). *Problemlösen als Informationsverarbeitung*. Stuttgart: Kohlhammer.

22.Dörner, D., & Wearing, A. (1995). Complex problem solving: Toward a (computer-simulated) theory. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European Perspective* (pp. 65-99). Hillsdale, NJ: Lawrence Erlbaum Associates.

23.Evans, J.St.B.T. (1982). *The psychology of deductive reasoning*. London: Routledge.

24.Farnham-Diggory, S. (1972). Cognitive processes in education: A psychological preparation for teaching and curriculum development. New York: Harper & Row.

25.Friedman, M. (2002). Kuhn, and the rationality of scienc. *Philosophy of Science*, *69*, 171-190.

26.Funke, J. & Frensch, P. A. (1995). Complex problem solving research in North America and Europe: An integrative review. *Foreign Psychology, 5*, 42-47.

27.Funke, J. (1991). Solving complex problems: Exploration and control of complex systems. In R.J. Sternberg & P.A. Frensch (Eds.), *Complex problem solving: Principles and mechanisms* (pp.185-222). Hillsdale, NJ: Lawrence Erlbaum.

28.Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.

29.Gibbons, A.S. (2001). Model-centered instruction. *Journal of Structural Learning and Intelligent Systems*, *14* (4), 511-540.

30.Gibbons, A.S. (2003). Model-centered learning and instruction: Comments on Seel (2003). *Technology, Instruction, Cognition and Learning*, *1* (3), 291-299.

31.Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R.R. Ronning, J. Glover, J.C. Conoley, & J.C. Witt (Eds.), *The influence of cognitive psychology on testing and measurement* (pp. 41-85). Hillsdale, NJ: Lawrence Erlbaum.

32.Gravemeijer, K., Cobb, P., Bowers, J., & Whitenack, J. (2000). Symbolizing, modeling, and instructional design. In P. Cobb, E. Yackel & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms. Perspectices on discourse, tools, and instructional design* (pp. 225-273). Mahwah, NJ: Erlbaum.

33.Greeno, J. G. (1989). Situations, mental models, and generative knowledge. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing* (pp. 285-318). Hillsdale, NJ: Lawrence Erlbaum.

34.Haberlant, K. (1999), *Human memory.,* Boston: Allyn & Bacon.

35.Hale, C.R., & Barsalou, L.W. (1995). Explanation content and construction during system learning and troubleshooting. *The Journal of the Learning Sciences*, *4* (4), 385-436.

36.Hambrick, D., & Engle, R. (2003). The role of working memory in problem solving. In J. Davidson & R. Sternberg (Eds.), *The psychology of problem solving* (pp. 176-206). NY: Cambridge University Press.

37.Hammer, D. (1997). Discovery learning and discovery teaching. *Cognition and Instruction*, *15* (4), 485-529.

38.Heidegger, M. (1968). *What is called thinking?* [Tr. J. G. Gray]. New York: Harper & Row.

39.Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem solving. *Educational Psychologist*, *45*, 1-10.

40.Hofstadter, D., & Sander, E. (2010). The essence of thought, New York: Basic Books.

41.Ifenthaler, D., & Seel, N.M. (2005). The measurement of change: Learning-dependent progression of mental models. *Technology, Instruction, Cognition, and Learning*, *2* (4), 321-340.

42.Jacobson, M.J. (2000). Problem solving about complex Systems: differences between experts and novices. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Fourth International Conference of the Learning Sciences* (pp. 14-21). Mahwah, NJ: Erlbaum.

43.Johnson-Laird, P.N. (1983). Mental models. Towards a cognitive science of language, inference, and consciousness. Cambridge: Cambridge University Press.

44.Johnson-Laird, P.N. (1989). Mental models. In M.I. Posner (Ed.), *Foundations of cognitive science* (pp. 469-499). Cambridge, MA:The MIT Press.

45.Jonassen, D. H. (2004). *Learning to solve problems: An instructional design guide*. San Francisco, CA: Jossey-Bass.

46.Jonassen, D. H. (2006). *Modeling with technology: Mindtools for conceptual change*. Columbus, OH: Merrill/Prentice-Hall.

47.Jonassen, D. H. (2007). *Learning to solve complex, scientific problems*. Mahway, NJ: Erlbaum.

48.Jonassen, D. H. (2011). Learning to solve problems: A handbook for designing problem-solving learning environments. New York: Routledge.

49.Jonassen, D. H., & Kim, B. (2010). Arguing to learn and learning to argue: Design justifications and guidelines. *Educational Technology Research & Development, 58*(4), 439-457.

50.Kafai, Y.B., & Ching, C.C. (2004). Children as instructional designers: Apprenticing, questioning, and evaluating ine the Learning Science by Design project. In N.M. Seel & S. Dijkstra, S. (Eds.). *Curriculum, Plans and Processes of Instructional Design: International Perspectives* (pp. 115-130). Mahwah, NJ: Lawrence Erlbaum.

51.Kant, I. (1781, 2009). *The critique of pure reason* [Tr. J. M. D. Meiklejohn]. Adelaide, Australia: The University of Adelaide eBooks.

52.Karplus, R. (1969). *Introductory physics: A model approach*. New York: Benjamins.

53.Keller, J. M. (2010). Motivational design for learning and performance: The ARCS model approach. New York: Springer.

54.Kim, C. (2012a). Beliefs about learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp.450-452). Dordrecht: Springer.

55.Kim, C. (2012b. Virtual change agents. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 3405-3407). Dordrecht: Springer.

56.Kim, C. (2012c). Motivational variables in learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (2347-2348). Dordrecht: Springer.

57.Kim C. (in press). The role of affective and motivational factors in designing personalized learning environments. *Eduational Technology Research & Development*.

58.Kim, C., & Balaam, M. (2011). Monitoring affective and motivational aspects of learning experience with the Subtle Stone. *Proceedings of the 11th IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp.640-641). Athens, GA: IEEE Computer Society Publications.
59.Kim, C., Keller, J. M., & Baylor, A. L. (2007, December). Effects of motivational and volitional messages on attitudes toward engineering: Comparing text messages with animated messages delivered by a pedagogical agent. Proceedings of the IADIS International Conference CELDA, Algarve, Portugal.
60.Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, and problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*, 75-86.
61.Kluwe, R.H., & Haider, H. (1990). Modelle zur internen Repräsentation komplexer technischer Systeme. *Sprache & Kognition*, *9* (4), 173-192.
62.Kolodner, J.L., Camp, P.J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., & Ryan, M. (2004). Promoting deep science learning through case-based reasoning: Rituals and practices in Learning by Design classrooms. In N.M. Seel & S. Dijkstra, S. (Eds.). *Curriculum, Plans and Processes of Instructional Design: International Perspectives* (pp. 89-114). Mahwah, NJ: Lawrence Erlbaum.
63.Krems, J. F. (1995). Cognitive flexibility and complex problem solving. In P.A. Frensch & J. Funke (Eds.), *Complex problem solving. The European perspective* (pp. 201-218). Hillsdale, NJ: Lawrence Erlbaum.
64.Kurtz, K. J., Miao, C. H., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences, 10*, 417-446.
65.Lakoff, G., & M. Johnson (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
66.Lakoff, G., a& M. Johnson (1999). Philosophy in the flesh: The embodied mind and its challenge to western thought. New York: Basic Books.
67.Lave, J., & Wenger, E. (1990). *Situated learning: Legitimate peripheral participation*, Cambridge, UK: Cambridge University Press.
68.Lesh, R., & Doerr, H.M. (Eds.) (2003). Beyond constructivism. Models and modelling perspectives on mathematics problem solving, learning, and teaching. Mahwah, NJ: Lawrence Erlbaum.
69.Markman, A.B. (1998). *Knowledge representation*. Mahwah, NJ: Erlbaum.
70.Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. *In* P. A. Alexander & P. Winne (Eds.), Handbook of educational psychology (2nd ed., *pp. 287-304).* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
71.Mayer, R. E., Mathias, A., & Wetzell, K. (2002). Fostering understanding of multimedia messages through pre-training: Evidence for a two-stage theory of mental model construction. *Journal of Experimental Psychology: Applied*, *8*, 147-154.
72.Mayer, R.E. (1989). Models for understanding. *Review of Educational Research*, *59* (1), 43-64.
73.Mayer, R.E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist, 59* (1), 14-19.
74.Merrill, M. D, (2002). First principles of instructional design. *Educational Technology Research & Development, 50*(3), 43-59.
75.Milrad, M., Spector, J. M., & Davidsen, P. I. (2003). Model Facilitated Learning. In S. Naidu (Ed.), *Learning and teaching with technology: Principles and practices* (pp. 13-27). London: Kogan Page.

76.Motah , M. (2006), *The Ontogeny of Memory and* Learning. In Proceedings of 2006 IS Conference. p.p 221-231.

77.Motah, M. (2007), *Learning, Types of Learning, Traditional Learning Styles.* In Proceedings of the 2007 CSITEd Conference.

78.Nair, K.U., & Ramnarayan, S. (2000). Individual differences in need for cognition and complex problem solving. *Journal of Research in Personality, 34* (3), 305-328.

79.Neubauer, A. C. & Fink, A. (2009). Intelligence and neural efficiency. *Neuroscience and Biobehavioral Reviews*, *33*, 1004-1023.

80.Norman, D.A. (1983). Some observations on mental models. In D. Gentner & A.L. Stevens (Eds.), *Mental models* (pp. 7-14). Hillsdale, NJ: Erlbaum.

81.Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review, 103* (2), 381-391.

82.Penner, D. E. (2001). Cognition, computers, and synthetic science: Building knowledge and meaning through modeling. *Review of Research in Education*, *25*, 1-35.

83.Piaget, J. (1951). *The psychology of intelligence*. London: Routledge and Kegan Paul.

84.Pirnay-Dummer, P., Ifenthaler, D., & Spector, J. M. (2010). Highly integrated model assessment technology and tools. *Educational Technology Research & Development, 58*(1), 3-18.

85.Quine, W. V. O., & Ullian, J. S. (1978). *The web of belief* (2nd ed.). New York, NY: Random House.

86.Resnick, M., & Wilensky, U. (1998). Diving into complexity: Developing probabilistic decentralized thinking through role-playing activities. *The Journal of Learning Sciences*, *7*(2), 153-172.

87.Rips, L. J. (1984). Reasoning as a central ability. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 2, pp. 105-147). Hillsdale, NJ. Erlbaum.

88.Rumelhart, D. D. (1980). Schemata: The building bloack of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). Hillsdale, NJ: Erlbaum.

89.Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart, & The PDP research group (Eds.), *Parallel distributed processing. Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (pp. 7-57). Cambridge, MA: MIT Press.

90.Schank, R. C., & Abelson, R. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Earlbaum Assoc.

91.Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32* (1), 102-119.

92.Schraw, G. (2006). Knowledge: Structures and processes. In P. A. Alexander & P. Winne (Eds.), *Handbook of Educational Psychology* (2nd ed.), (pp. 245-264). Mahwah, NJ: Erlbaum.

93.Seel, N. M. (1999). Educational diagnosis of mental models: Assessment problems and technology-based solutions. *Journal of Structural Learning and Intelligent Systems*, *14* (2), 153-185.

94.Seel, N. M. (2003). Model-centered Learning and Instruction. *Technology, Instruction, Cognition, and Learning*, *1* (1), 59-85.

95.Seel, N. M. (2005). Designing Model-Centered Learning Environments: Hocus-pocus – Or the Focus Must Strictly Be on Locus. In J.M. Spector, C. Ohrazda, A. Van Schaak & D.A. Wiley (Eds.), *Innovations in Instructional Technology: Essays in Honor of M. David Merrill* (pp. 65-90). Mahwah, NJ: Lawrence Erlbaum.

96.Seel, N. M., Al-Diban, S., & Blumschein, P. (2000). Mental models and instructional planning. In M. Spector & T.M. Anderson (Eds.), *Integrated and holistic perspectives on learning, instruction and technology: Understanding complexity* (pp. 129-158). Dordrecht, NL: Kluwer Academic Publishers.
97.Spector, J. M. (2001). A philosophy of instructional design for the 21st century? *Journal of Structural learning and Intelligent Systems, 14*(4), 307-318.
98.Spector, J. M. (2010). Mental representations and their analysis: An epistemological perspective. In D. Ifenthaler, P. Pirnay-dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (17-40). New York, NY: Springer.
99.Spector, J. M. (2012). Naturalistic epistemology. In N. M. Seel (Ed.) *The encyclopedia of the sciences of learning* (pp. 2433-2435). New York. Springer.
100.Spector, J. M., & Koszalka, T. A. (2004). *The DEEP methodology for assessing learning in complex domains* (Final report to the National Science Foundation Evaluative Research and Evaluation Capacity Building). Syracuse, NY: Syracuse University.
101.Spiro, R. J., Coulson, R. L., Feltovich, P. J., & Anderson, D. (1988). Cognitive flexibility theory: Advanced knowledge acquisition in ill-structured domains. In V. Patel (Ed.), *Proceedings of the 10th Annual Conference of the Cognitive Science Society* (pp. 375-383). Mahwah, NJ: Erlbaum.
102.Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Wien: Springer.
103.Sterman, J. D. (1994). Learning in and about complex systems. *Systems Dynamics Review 10*, (2-3), 291-330.
104.Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257-285.
105.Turner, R.M. (1994). *Adaptive reasoning for real-world problems: A schema-based approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.
106.van Merriënboer, J. J. G., & Kirschner P. (2007). Ten steps to complex learning: A systematic approach to four-component instructional design theory. New York: Routledge.
107.Volet, S. E. (1991). Modelling and coaching of relevant, metacognitive strategies for enhancing university students´ learning. *Learning and Instruction*, *1* (4), 319-336.
108.Wertheimer, M. (1959). *Productive thinking* (Enlarged ed.). New York: Harper & Row.
109.White, B. (1993). ThinkerTools: Causal models, conceptual change, and science education. *Cognition and Instruction*, *10* (1), 1-100.
110.Wittgenstein, L. (1922). *Tractatus logico-philosophicus* (translated by C. K. Ogden). London, UK: Routledge & Kegan Paul.
111.Wittgenstein, L. (1953). *Philosophical investigations* (translated by G. E. M. Anscombe). Oxford, UK: Blackwell.
112.Young, R. (1983). Surrogates and mappings: Two kinds of conceptual models for interactive devices. In D. Gentner & A.L. Stevens (Eds.), *Mental models* (pp. 35-52). Hillsdale, NJ: Erlbaum.

J. Michael Spector is Professor and Chair of Learning Technologies in the College of Information at the University of North Texas. He was previously Professor of Educational Psychology and Instructional Technology, Doctoral Program Coordinator for the Learning, Design, and Technology Program, and a Research Scientist at the Learning and Performance Support Laboratory at the University of Georgia (2009-2011). Previously, he was Associate Director of the Learning Systems Institute, Professor of Instructional Systems, and Principal Investigator for the International Center for Learning, Education and Performance Systems at Florida State University (2004-2008). He served as Chair of Instructional Design, Development and Evaluation at Syracuse University (2000-2004) and Director of the Educational Information Science and Technology Research Program at the University of Bergen (1996-1999). He is a distinguished graduate of the United States Air Force Academy and earned a Ph.D. in Philosophy from The University of Texas at Austin (1978). His recent research is in the areas of intelligent support for instructional design, system dynamics based learning environments, assessing learning in complex domains, distance learning, and technology integration in education. Dr. Spector served on the International Board of Standards for Training, Performance and Instruction (ibstpi) as Executive Vice President; he is on the Executive Committee of the IEEE Learning Technology Technical Committee, is a Past President of the Association for Educational and Communications Technology (AECT) as well as a Past Chair of the Technology, Instruction, Cognition and Learning Special Interest Group of the American Educational Research Association. He is the editor (10th year as editor) of the Development Section of Educational Technology Research & Development, and he serves on numerous other editorial boards. He co-edited the third edition of the Handbook of Research on Educational Communications and Technology, is again lead editor on the fourth edition, and has more than 100 journal articles, book chapters and books to his credit. Dr. Spector is active in international educational technology developments and was the Conference Co-Chair and Local Organizer for the 2011 IEEE International Conference on Advanced Learning Technologies held at UGA 6-8 July.

Dr. ChanMin Kim is an Assistant Professor of Educational Psychology and Instructional Technology at the University of Georgia, in the United States. Her research agenda involves promoting learners’ motivation and volition throughout the implementation of a virtual change agent in online learning environments.

Received: December 26, 2011; Accepted: January 23, 2012.

DOI: 10.2298/CSIS111223020T

Analysis of processes of cooperation and knowledge sharing in a community of practice with a diversity of actors

Diane-Gabrielle Tremblay and Valéry Psyché

Abstract. According to some literature, communities of practice should normally stem from a voluntary initiative within an organization, whose members share some knowledge or expertise they wish to improve. However, over time, we have seen that communities tend to be created within organizations, in order to attain objectives of learning and knowledge development. This represents a challenge in the context of a community of practice taking the form of a research network in partnership that brings together members with common interests certainly, but spread out in different organizations and even several countries in which they perform different types of work. Also, the community does not exist in a vacuum and the explanation for what happens within it does not lie solely within the way the group interacts; indeed the individuals are part of different organizations and thus have different priorities, in relation with these affiliations. In this context, our research objective was to determine the factors that facilitate or hinder cooperation within a community of practice composed by two groups of actors, community and university actors. We thus found that individuals' different work affiliations might not facilitate the work within the CoP and that ICT/web 2.0 tools are not always a solution to increase participation in a CoP. Although participants are somewhat familiar with the tools, they mostly seem content with receiving and accessing information, not searching for a more active participation. Some explications and solutions will be proposed.

Keywords: community of practice, collaborative research, cooperation, knowledge sharing, time, trust.

1. Introduction

Sustained participation among members is a major challenge for many research projects in partnership, since many of them have the goal not only to perform research projects, but to achieve them in collaboration and support exchanges among the organizations represented (academic, community, or other as appropriate).

We studied a Community of practice (CoP) based on a research in partnership. In our study, the practice we are interested in is related to the learning of the "collaborative research partnership" in a community of practitioners from diverse backgrounds (unions, community groups, etc.).

This paper presents our observations on the evolution of a community of practice (CoP) created to share knowledge within a research program in partnership, through the use of web 2.0 tools (website, blog, newsletter, RSS feed, etc.) developed with and for members of this network to promote the collaboration between them. The research aimed to identify the challenges and difficulties in knowledge-sharing. If communities of practice are more often described and considered as spontaneous phenomena, more recently, many firms and organizations have wanted to create Communities of practice (CoPs) altogether, from scratch, in order to foster learning and knowledge exchanges in various contexts. We thus decided to observe a network of collaborative research which was created with a view to develop research exchanges and partnerships, from the perspective of a community of practice.

We will show that it is not as easy as the literature seems to indicate to create a CoP in a research partnership context. The fact that members come from different organizations, are not as strongly bound by the same objectives, and need time to build trust between them explains the difficulties in creating a CoP. Also, while many participants say they feel committed to the CoP, they do not participate as actively as could be expected, in information and knowledge sharing, given this commitment to the CoP. This time element and the distinction between commitment and active participation in the CoP do not appear to have been put forward in previous literature and this would be important to take into account in future endeavors to create active, engaged CoPs in various work environments. Some explications and solutions will be proposed.

To start, we present the context of this research; then we expose our research questions. In a second part, we will report on theoretical aspects. We then present our methodology, our results and, finally, we discuss the elements that we can learn from this analysis in view of the literature on communities of practice. We will highlight a few elements which, in our view, have been previously neglected, especially in a context of analysis of more spontaneous CoPs, and which need to be taken into account today, when we study CoPs in various contexts, but especially when one wants to create a CoP.

2. Research Context and questions

2.1. The domain

The domain of the CoP that we studied is a network based on a research partnership between universities and social actors. More specifically, this research partnership is a group of academics and social partners interested in some social issues and who have received a grant from the Canadian Social Science Research Council to develop knowledge exchanges and research on a few social issues (work and family reconciliation, ageing and life course issues). The network of partners brings together researchers, business people, community associations, trade unions, sectoral committees essentially, and is targeted to conduct joint research, but also to share research findings and other information or knowledge developed by its members. The research partnership also aims to create learning resources, tools, as well as to foster knowledge transfer and information exchange between community partners (as they are called in the project) and university researchers.

2.2. The Community

The governance structure of the research network consists of several clusters (5) of thematic research partnerships co-chaired by a member of each, a community investigator and a university researcher, to ensure equal representation of both communities in the governance bodies. The community researchers are usually persons who are responsible for research or socio-economic issues within their organization (association, union, etc.). They have worked on research projects, and are thus familiar with research issues, but do not work in the same context and do not have exactly the same objectives in research (knowledge and publication for university researchers, more knowledge and action for community actors, even if all are interested in some form of publication and action. The clusters meet at will to discuss research issues and to develop research projects and propose activities to the advisory committee. The advisory committee brings together the chairs of the sub-groups and the director of the research partnership, that is 11 persons for the decisions and discussions on research projects, but also on the roles and responsibilities of each in projects, on knowledge tools and all elements of the project.

Management and collaboration are supposed to flow freely, based on the clusters' activity and ideas. Over two years, there were over fifteen meetings between the members – often within specific clusters or sub-groups. There were also 3 large conferences, and some 8 symposia and meetings open to all since the beginning, but at the time of the research, only 2 conferences had been held, one was in planning, and 4 symposia or meetings had been held.

The equal treatment of partners from the community and from university in the conduct of research is, in principle at least, ensured by the joint management of projects, and meetings involving at least one community partner and a university researcher, in addition to at least one student to conduct the research. This pattern is designed to foster a sense of belonging among members and develop direct cooperation between them on specific issues and research projects.

New partners (e.g., businesses, places of observation or research) are invited to join through contacts that are made during a research or through community partners themselves. There have been discussions and proposals for new partners (2 community groups and 2 researchers) they have been submitted to the Executive for approval.

A person is responsible for coordinating exchanges between all members, keeping track of activities and emails. Another person has devoted at least the equivalent of a part-time to the implementation of a website, which includes a blog, to facilitate direct knowledge exchanges between members. Also, members are encouraged to provide information or documents they wish to upload on this site to share with others, and websites of the members are also present on the site, so that members can get to know each other's activities better.

2.3. The practice

The objective of this CoP is for community and academic researchers to exchange information, to learn from each other and to develop skills and practices of collaborative research as a result of their participation in this research network. It is also that they develop collaborations even outside of the projects and research discussed in the network for other activities (conferences, research) or on other topics that they may find they have a common interest in.

Collaborative working and knowledge sharing[1] (as well as see [1]) already exists among several community and academic researchers, since a number have worked together for many years, but this is more the case between university researchers from different universities, or bilateral relations between one professor and one community group in previous research. Some researchers also collaborated in teams of two or three for about four years. These researchers collaborate essentially by peer-to-peer emails, and were not very familiar with blogs or other formats. However, the research network has developed new tools for exchange and sharing on the web, to promote exchange and dissemination of knowledge on a wider scale, and give access to exchanges to all members. Notably, there is a website which can receive information and documents from the members (news section, PowerPoint and videos from conferences, documents, research notes, etc.) but also a blog which should favors interactive collaboration between partners and researchers. This was planned to enable members to exchange information and just allow anyone (even non-members) to have access to the web and send in information. Joint development of knowledge, as well as knowledge

[1] Which can be defined as is an activity through which knowledge (that is information, skills, or expertise) is exchanged among people, such as members of a department, an organization or another type of group.

See Wikipedia: http://en.wikipedia.org/wiki/Knowledge_sharing

sharing between partners and researchers was thought to happen in the way it is presented in CoP literature, through spontaneous exchanges and learning, with the collaborative tools developed for the project.

Most academic researchers and the community actors present here have had relatively extensive experience with collaborative research and many knew at least one tool for sharing knowledge other than email. The challenge for everyone, especially those of the community sector, associations and unions, appears to be to find time to attend meetings and to participate in exchanges, as we will see further on. The network has sometimes had to split the meetings to discuss issues with everyone concerned, and thus the interest in developing new forms and tools for collaboration is high in this case, as in many other partnerships where learning and exchanging knowledge are the objective.

3. Research Problem

While communities of practice are generally viewed as being spontaneous groupings of people, many recent initiatives have been derived from an intentional initiative; indeed, over the years, more and more studies in management and education deal with the creation of communities of practice that can be described as intentional [2], but still, the literature tends to indicate that knowledge sharing is quite spontaneous, which is something we wanted to look into in more detail. How and why does this happen or not?

A “classical” or spontaneous community of practice [3] should actually come from a voluntary initiative within an organization, whose members share some knowledge or expertise they want to improve. Although the CoP studied here expected to see relatively spontaneous knowledge exchanges, once the tools were in place (website and blog) this was not the case.. We also find relevant interest in the conditions under which there is or there is no participation leading to the reification of activities for collaborative research partnerships and learning in this context [4, 5].

This objective of co-construction of knowledge[2] through exchanges with peers is obviously a challenge. While bringing together members with common interests in a given social issue was expected to spontaneously

[2] Which can be defined as the development knowledge within a group, collectively, ie. university researchers with community people, rather than university researchers defining the research questions, mode of analysis, results. Kersey et al. (no date) consider that co-construction appears to « closely resemble two types of initiative that the dialogue research community distinguishes between; dialogue initiative and task initiative. Dialogue initiative tracks who is leading the conversation and determining the current conversational focus while task initiative tracks the leader in the development of a plan to achieve a problem solving goal. »

foster knowledge exchanges, this did not happen as expected. We therefore set out to understand why the common interest in the issue did not suffice to foster active exchanges on the blog and important contributions in terms of information and documents to be put on the website. As we will see further on, people coming from different organizations (unions, companies, associations, public administrations, etc.) have other commitments, they assume different functions and are often busy elsewhere. This represents a great challenge for research in partnership and collaborative exchanges to be developed, while at the same time it is often said that the diversity of actors participating in a community make it more creative and innovative [6].

In this particular case, we had thought that the fact that people were spread out in several countries on two continents (Canada, France, Belgium, Sweden, etc.) would to a certain extent force them to use the web and blog to exchange information easily, while CoPs within an organization often find that people don’t use the digital tools put at their disposal and continue meeting in the corridors to exchange information [7], thus limiting the exchanges to a more limited number and also making it difficult to trace the knowledge exchanges and learning that do occur. While the international distribution of members (although the majority are in Montreal and second, in Ottawa, Canada) had been expected to foster more activity on the blog, this was not the case. In principle, strategies, tools, innovative technologies of the social web can promote and facilitate exchanges of information and documents, as well as collaboration among members. Given that this was not the case, we decided to try to understand the reasons that made it difficult to foster knowledge sharing within this community.

The research was thus designed to determine the factors that facilitate or hinder cooperation within a community of practice, provided that the bulk of publications on communities of practice focus on successful communities and very few try to explain why things work or don’t work in terms of the knowledge exchanges and learning. It therefore appeared necessary to identify the obstacles to collaboration in the context of this emerging community of practice in order to identify factors that may be a source of mobilization and commitment of the actors, or otherwise encourage more active participation. This research should help better understand the actual dynamics of communities of practice, and the factors that favor or hinder knowledge exchanges and learning. This is a form of action-research [8-10] and one that involves actors from different environments who have much to learn from each other. Although this does present some challenges, it is not uncommon since it is in such a context that new information can be acquired and learning can occur [8].

Indeed, communities do not exist in a vacuum and it is important to try to understand the real challenges that can appear when members come together from a variety of organizations. To what extent can the trajectory of the development of the CoP, the learning and knowledge exchanges, be explained by the traditional factors that explain CoP development, that is common enterprise and mutual engagement, and to what extent can they be explained by other factors? This is the main question we will address here, since it seems this issue is rather new on the agenda, now that we are looking more and more at intentional communities, and not at CoPs that appeared spontaneously, as was the center of interest for many years (and still is) in research on CoPs.

4. Theoretical framework

The basic definition of a community of practice is as follows: Communities of practice are groups of people who share a passion for something they already do or practice and who interact regularly to learn how to improve this practice [11].

4.1. Communities of Practice: Definitions

Coined by [12], the term community of practice refers to a group of people with a common area of expertise or professional practice, and who meet to exchange, share and learn from each other, face to face or virtually [3]. The group usually evolves naturally due to the common interest of members in a particular area, but there is more and more interest recently in CoPs that can be created specifically for the purpose of acquiring and exchanging knowledge related to a given field. What has been called the intentional [7] creation of communities is becoming more common in various organizations and more open environments as well.

The members of the CoP are usually bound by a common interest in a field of knowledge (...) a desire and a need to share problems, experiences, models, tools and best practices [13]. It is through the process of sharing information and experiences with the community that members learn from each other, and have the opportunity to grow personally and professionally [12].

Wenger et its colleagues [3] have also developed a model of developmental stages of communities of practice. In this model, the level of maturity refers to the stages in the evolution of a community. They present the five stages of the life of a community. This is of course an ideal -type and reality can diverge from the theoretical model. In principle, however, from a more or less formal network of people, the community is at the stage of development potential. Subsequently, the community moves to the stage of unification, when it bonds together, and then to maturity. It then reaches a momentum, despite possible ups and downs and normally an external event would then trigger the need for change. The model is of course theoretical and the duration of the various stages is different depending on the community. Anyway, most research suggests that it takes several months before a community reaches the stage of maturity and produce concrete results [14].

Note that communities of practice are different from teamwork on several points. Thus, in principle, teams are generally defined by the results they must deliver, while communities rarely have a specific outcome to provide to the organization. Similarly, in principle, the team members are bound by a given objective; while those of a community are united by the knowledge they share and develop together. In operational terms, communities, unlike teams, rarely have a defined work plan [15]. After reaching their goals, the teams would normally disintegrate, whereas communities of practice are created to last, continuing to develop skills and knowledge over time.

In practice, however, boundaries are sometimes blurred between these two organizational forms that are the teams and communities of practice [15-18].

Also, the development of any community is obviously influenced by its environment, and the past of the sponsoring organization (a form of path dependency), but it can also be influenced by the cultural, economic and political environment in which it is embedded, the environment being more or less favorable to its development [11]. The degree of recognition of the work done within the CoP by each organization may also influence its development, as well as financial, material and human resources available to it, especially as regards the animation resource, more often put forward now, in the context of intentional CoPs.

4.2. Communities of practice and learning

According to [19], learning is a function of the activity, context and culture in which it occurs. It is "located" as opposed to learning based on presential activities, in classrooms, where knowledge is often abstract and out of context.

For [12], learning is acquiring knowledge in a social setting, in situations of co-participation or joint venturing. Social interaction and collaboration are two key components of situated learning, engaging learners in a community of practice, which embodies certain beliefs and behaviors to be acquired [20]. Literature indicates that as the beginner or new participants move from the periphery of the community to its center, they become more active and engaged in community culture, and indeed more likely to assume the role of expert or “knowledgeable person” in the CoP. These ideas are what [12] call the process of “legitimate peripheral participation”.

Learning is recognized by several authors as the major objective of the communities. Thus, [21] present a typology of communities within companies and distinguish between forms of learning observed in communities and teamwork. They believe that communities are designed to enable learning in action at work (learning in working), while the work team provides learning through interaction and the project or functional group enables learning by achieving given tasks (learning by doing). In reality, the frontiers between these types of learning are sometimes blurred, especially since not all CoPs are designed for work projects, some being oriented towards learning on a given subject, without any specific work objective.

4.3. Factors of success of communities of practice

It is difficult to find in the literature a clear definition of a successful community of practice (CoP), let alone a virtual community of practice (VCoP), operating remotely, using technologies. In most cases, the CoP seeks to share knowledge within an organization or network of organizations [7], and success is usually evaluated based on the achievement of this knowledge-sharing objective. Some authors also consider a community of practice is successful when it achieves the objectives it sets itself, whatever the nature of these objectives [22].

In various studies, we found a few indicators of success: the achievement of various goals, member satisfaction, interest in continued participation in a CoP, as well as various forms of knowledge sharing [6, 21].

Among the factors that contribute to success, we can also mention individual attitudes towards other members of the community (their social presence, motivation, collaborative culture ...), and their common interest (common goals, shared practice in the community ...), [21, 23-28].

Communicating with others in a CoP is to create a social presence. [27] defines social presence as “*the degree of salience of another person in an interaction and the consequent salience of an interpersonal relationship*" (p. 38). Social presence would affect the likelihood that an individual will participate in a CoP, especially if it is virtual [27].

The motivation to share knowledge is essential to success in communities of practice. They are even more active if they have a return on investment. This "return" can be tangible (promotion, bonus ...), intangible (reputation enhancement, increased self-esteem ...) or of interest to the community (exchange of best practices related to knowledge, interaction and common goal). Members are encouraged to become active participants in a CoP when they consider that the information has meaning for the public good and represents a moral obligation or a collective interest [23]. Collaboration is essential to ensure that communities of practice develop. This is a key success factor identified in the literature, especially in places where the level of education is high and among the more experienced members [26].

Among the factors limiting the success of CoPs, obstacles such as ego, personal attacks, or time constraints have been identified as possibly preventing participants from engaging in knowledge sharing [28].

4.4. Factors for developing successful communities of practice?

The success factors of CoPs are related to actions to implement them. Wenger et its colleagues [29] identified seven actions that may contribute to cultivate a community of practice: 1. Designing for change; 2. Open dialogue between perspectives from inside and outside; 3. Encourage different levels of participation; 4. Develop community spaces both public and private; 5. Focusing on the value; 6. Combine familiarity and excitement; 7. Create a rhythm for the community. We are particularly interested in the last five points, which inspired us in our research project, including an attempt to develop partnership and sense of belonging to the CoP.

These elements presented by [29] and the literature presented above lead us to think that it is actually quite easy to get people engaged in a community and in its exchanges and learning activities. However, it seems the literature on CoPs presents a somewhat idealized view of the way such groups function. It seems to often ignore or underplay the role of exogenous factors such as external power or work relationships, the fact that people belong to organizations who impose constraints on their time and activities and cannot necessarily be as active as they might like in the CoP. In this article, we want to take into account this dimension, which has been highlighted in recent work [30], but is not often put forward. Indeed, communities do not exist in a vacuum and the explanation for what happens within them does not lie solely within the way the group interacts[3]. The community we studied is composed of many members coming from a variety of organizations, which should theoretically lead to richer knowledge sharingfor the participants, although it may present challenges. Which factors play a role? This is the main question we will address here.

5. Research Methodology: a participatory design approach

Our research project therefore aims to identify the factors that can explain the degree of knowledge sharing within a CoP. We selected a methodology that would: (1) take into account the particularity of the research partnership (based on participation of researchers and practitioners), (2) contribute to the emergence of conditions conducive to learning from the perspective of a community of practice, (3) through the design of tools adapted to the practice of research partnerships.

It is of course simplistic to equate the process of emergence or creation of a community of practice to a process of design tools, be they technological, and social. However, collaboration and knowledge sharing cannot be done without support of technological tools, especially when community members are in different places and organizations [31].

The participatory design is an approach derived from the participatory action research [32]. This approach is defined as an iterative process of negotiation between a diversity of actors - different from each other in terms of their disciplines, their concerns and interests - in order to effectively influence the design process [31]. This flexible approach does apply to our configuration of actors that includes a multidisciplinary group composed of researchers

[3] Thanks to a colleague for directing us to the article by Kimble et al. (2010), which was an important part of our thinking on this case study.

involved in the study (from universities and enterprises), designers and developers. In this configuration, the role of 'participatory' designers is a role of facilitators whose goal is to give control to users (researchers) so they can make their own decisions [33], while facilitating effective participation in the design process. This role was held by a post-doc student hired specifically to work on the development of the CoP and to analyze its functioning and results.

The Participatory Action Research (PAR) can be used to define a methodology that promotes the effective participation of diverse actors in the design process. Effective participation, the one that really influences the design process must be organized around participatory activities (debate of ideas, prototype demonstration, role play, etc.) leading to the production of boundary objects [34, 35]. These objects "to-think-with" (model, scenario, cases, etc.) facilitate mutual understanding and trust among participants from various backgrounds. The idea of boundary object is closely linked to that of reification which, in shaping the experience of participants, produces objects and thereby tends to create points of interest around which the negotiation of meaning is organized” [11]. Thus, reification is an indispensable collective anchor to sharing and capitalization of knowledge [36].

While respecting the cycle of PAR methodology, we relied on three phases of participatory design: (1) Exploration and discovery (requirements definition, storyboarding), (2) Prototyping (modeling, testing and experimentation), and (3) Evaluation (dissemination of results in a format understandable by the participants). This methodological approach has resulted in several participatory discussions, demonstrations, and documentation. In what follows, we present some of the techniques used in all three phases.

5.1. Exploration and discovery

This first phase of the methodology aimed to capture needs. It has allowed us to identify quickly the services of potential interest for the participants. Among the techniques and tools used for exploration, we have: observation, online questionnaires, paper questionnaires, debates, interviews and scenarios of collaborative practices, research in partnership.

Observations in the scenario of knowledge sharing and collaboration

At the launch of this initiative, a first meeting to get the partnership and CoP going took place in June 2009. The participants then had discussions in each thematic group to determine topics of interest to research partners. Subsequently, there were meetings within each group to go further and undertake research and seminars. Since the research partnership aims to develop exchanges and networking among members, from the beginning, they were encouraged to cooperate, not only within the project, but also outside, on other issues or activities. Several students (four doctoral students, three postdoctoral researchers) were brought into contact with social partners to develop research projects.

*Storyboard*s (of collaboration in research partnership) have been made. They come from the data collected and were used later as inputs for the phase 2.

In addition, the participants discussed their needs for socio-technical tools for research collaboration. This debate was guided by facilitators with a questionnaire.

Questionnaires and scenarios

We distributed three questionnaires during the first year of the network.

> The first data collection was carried out during the launch of the research network in June 2009. It focused on community needs in terms of tools and technologies. We distributed paper versions of the questionnaire, called Q1, in the conference room, at the first meeting of the network where most members were present. There was tracking and follow-up throughout the two days to recover a maximum of responses.
>
> The second data collection was performed immediately after a meeting of the advisory committee in February 2010. It focused on how members situate themselves towards the others. To perform this, we sent an email and referred people to an online questionnaire (GoogleDocs technology) called Q2. Not having received many responses (5 responses), we then returned the questionnaire by mail as a word attachment. We only got two new answers.
>
> The third data collection was done two months later, in April 2010. It focused on patterns of collaboration among members within the research network and also in their organization. Based on the experience of Q2, we sent an electronic version, called Q3, directly to members via email. The response rate was the lowest of the three cases because of the subject, perhaps, or lack of time or the fatigue of responding to questionnaires.

Scenarios of practices of collaborative research partnership were then designed. They were based on the data collected and were used as inputs for the next phase.

5.2. Prototyping and design tools

After this phase of exploration and discovery, the tools were tested by participants in the network. We do not dwell on other steps in this phase which is essentially a design and development of tools specific to software engineering. This second phase led to the development of the first version of a newsletter, a blog and a website to include Web 2.0 technologies such as son feeds, tag clouds, etc.

5.3. Assessment

This last phase was of great importance in our study, as it was here that we took action to document and analyze the practice of collaboration. These steps are: the follow-up by validation and the assessment of participants' thinking about their collaborative activities.

Follow-up by semi-structured interviews on the practice of collaboration

In making this assessment and the qualitative part of the research, we targeted a small group of members (among the most active in the network), and we contacted them by telephone for interviews. Out of twenty-three members contacted, we had ten responses, eight positive responses, that is to say people who agreed to answer our questions. Finally, we conducted semi-structured interviews, most of which with "community partners", which is good since these are the most difficult to get actively involved in the CoP. The interviews were conducted in July-August 2010.

The objective was to monitor the main modes of participation in the CoP, by gathering information on participation and motivations for participation or for less active involvement.

The interviews were semi-structured. The questions were open and based on some previous work done on this theme by the Cefrio research center, which has developed expertise in this field [2, 7]. We discussed the themes of participation, communication, dissemination and knowledge sharing, achievement of goals within the CoP, members’ satisfaction and interest to continue

Questions related again to mode and frequency of participation, use of technologies and feeling of comfort, but also developed more extensively on motives for participating actively or not, achievement of goals within the CoP, members’ satisfaction and interest to continue.

Data processing for the analysis of interviews

Our method of data processing for the interviews is qualitative, based on a thematic analysis, and without community involvement in the analysis. To help us with the analysis, we used a statistical tool, Google Analytics, which is connected to the website and blog. Google Analytics tells us, among other things, the extent of use of the site, the origin of visitors, pages viewed, etc. This was used as secondary material to complete our questionnaires and interviews.

After this analysis, we provided a first ontology of the CoP showing its process of collaboration and knowledge sharing, using ontological engineering [37, 38] in order to give the members some feedback on their activities.

6. Results

6.1. Results from the process of exploration and discovery

6.1.1. Most respondents rather passive

The responses to questionnaires (Q1, Q2 and Q3) gave a low response rate, despite the fact that they were sent to members through various modes of communication.

Our observations indicate that members lack the time to give their views on issues relevant to the network. This suggests that although it is a partnership program, planned to function as a Community of Practice on a given theme, in order to "co-construct knowledge", participants are not very active in the activities or knowledge exchanges.. The vast majority are happy to receive information, participate in organized activities, some partners are glad to be participating actively in research with students and academics. Those who participate in specific research projects are more active, but usually focus on their business, their research, developing little exchanges or transmitting little other information on their activities or other publications.

As mentioned above in the section on the Research question, our results prompted us to identify the reasons for non-participation in online activities. It appears to be mainly lack of time, and work overload that are at the origin of the problem, but also may be a lack of appropriation or trust in the knowledge sharing tools and the CoP.

Some communication tools gave us access to relevant data on the interest of the participants in the network and on habits of participation through a statistical tool (Google Analytics[4]).

This confirms that members have a relatively passive attitude vis-à-vis the website, which includes tools (news section, space for their resources, RSS, statistics), allowing them to gain visibility within the CoP and on the web. Indeed, there is a space which identifies the activities related to members and a space dedicated to resources and links they wish to share with other members. In both cases, members were regularly invited to send any information or document concerning and of interest to the community. They can also receive the latest news from the network directly into their mailboxes or on the homepage of their browser by default, thanks to RSS technology to which they have access from the website.

In terms of access to the webpage, statistics (below) confirm the trend. Data (from March 15th to November 15th) indicate that the website was

[4] Google analytics is a tool that gives data on website traffic, according to various variables: timeline, country of origin, number of new hits vs returns, etc. See http://www.google.com/analytics/

explored 1708 times and by 1086 unique visitors There have been approximately 55.68% new visits. The average time spent each time on the website is 2 minutes and 23 seconds. Some 4,627 pages were explored or 2.71 pages per visit.

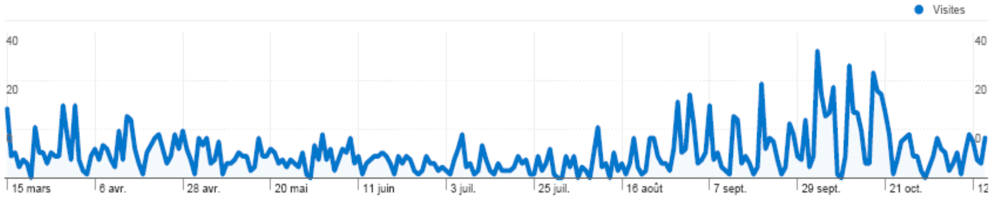


Fig. 1. Site activity

Over the whole period, the number of visits declined during the summer months (June, July and August): it went from 239 visits in April to 111 visits in July (see Fig. 1). Then, the curve shows an increase in August and continues into September and October. This increase seems to correspond to a period when a particular activity was planned (in this case a major symposium for the network in the beginning of October). It is important to note that June, July and August are usually the months when people take holidays (academics usually take one month vacation in July or August, the partners of other groups 3 to 4 weeks in the same period).

The newsletter is another way that we targeted to send information to members and to help maintain contact with them and it is also received rather passively. Again, despite repeated calls for input, the network did not receive information to communicate to other members, except after very direct contact with one person, or when things came up during meetings. The webmaster surveys the web (including electronic academic journals) to find relevant information about members to share with the others; information on events and publications in which they participated is thus brought to the knowledge of the others.

6.2. Analysis of the Community of Practice and evaluation of the practice

6.2.1 Need for more time to participate in a CoP:

The interviews gave us a better picture on various themes of participation, communication, dissemination and sharing of knowledge, achievement of goals within the CoP, members' satisfaction and interest to continue.

As concerns achievement of goals, most participants did not have a very precise objective, except to gain information and knowledge, and they consider they did get access to information, although they contributed rather little themselves. Nevertheless, they were generally satisfied with the way things were going and were interested in continuing.

On the theme of participation, most participants interviewed said they felt engaged or highly engaged in the search network, but described their participation as infrequent. Indeed, some say that their participation is "sporadic, according to the issues raised, according to our needs or the needs of the research network " or "it depends on the projects that are ongoing or planned activities", or: " my organization has never released more time for me to do more than that". Moreover, half of the participants think that the participation rate in the network is sufficient. Thus, some say: "I don't think I can get more involved ... There is no way I could".

The other half admits not having enough time; for example, one says: "It will not change if there are greater incentives because I'm the only person responsible for social issues in my organization; we used to be three, but now I'm alone ... I could not do more, given my job". They give very specific examples of time constraints; "... We want to participate, but don't have time" or "I am so busy that I do not run after information. I have a business, it is a small social economy business; sometimes there are big organizations, where there are people who are specifically responsible for research. But I have a business, so I am making an effort to support research. But I'm too busy to add or seek more information. This level of information suits me perfectly."

Another element is that people don't spontaneously participate in a CoP to share knowledge. Some want more leadership from the others, or specific projects proposed to enhance participation. «There's probably a lack of leadership, or projects put forward in a more structured way... because I think it does not come naturally to partners to propose information, or research or knowledge they have". Most partners clearly wait to be directly solicited for participation.

Some don't feel confident to go forward and propose activities or knowledge; they don't feel there is enough trust to test ideas or projects: "There are some interesting pairings to be made between our organization but the initiative will not come from us, partners; this needs to come from the network's management, to propose topics for our exchanges" .Another says: "In my view, I have a 75% participation level. I missed a meeting; when I am contacted, I participate, but when I am not solicited, I'm not proactive."

Overall, all participants are active in the community as they participate in at least one research project or attend seminars, meetings or other activities on site or by phone. They feel they are active in exchanging ideas when solicited, but are less active with regard to online activities (putting things on the blog or asking for them to be put on the website), do not tend to really exchange knowledge, unless they are actively collaborating in a specific research project. In general, they do not share other knowledge that they have in relation with the theme of the network, nor will they seek to find knowledge or information via the communication tools / website. Finally, that they are reluctant to be proactive in advocating for knowledge exchanges, research and activities, although in the context of this project, which is aimed precisely at knowledge building through cooperation between academics and community, they are encouraged to be as active as the project management on these issues.

On the issue of communication tools, the majority of participants surveyed show a certain ignorance of the main tools of web communication, mainly the blogs, and some confusion about their use. Thus, some argue: “The blog is not a tool that I spontaneously feel attracted to. The website, I do not go every day, I went there a few times”, or “I don’t have the reflex to go there once a week, the consultation takes time. If it is updated every day, I know there's new material, I’ll maybe go”. Another person says: “There are many means of communication, they cover a broad spectrum of needs but they are not sufficiently clear to me. ... I do not think it is clear how to proceed, who we need to contact to put such information in the newsletter, I think there is a lack of information” (even if they have received regular monthly emails from the coordinator asking them for information to be shared. As a result, until recently when one person decided to write some short messages for the blog, there has been little participation. Some post-docs say they will write things, “When they have finished their research” but there is no spontaneous thrust to put information out there to share with the others. The network coordinator will put things on the web and in the newsletter, to give others the idea that they can also share similar things, and she has even put a video found on the web, from one of the network members giving a conference on a theme of interest to the network, but the majority of participants still remain rather passive. They don’t feel comfortable with the tools (blog mainly) and they tend not to use these tools to search for information.

On the issue of knowledge sharing, the majority of participants surveyed said they expect to find extensive information (activities on the members, on students, research reports, technological, text search or practice, recent news, etc.) related to the themes of the network, which is pretty much what they do find, but they don’t feel compelled to contribute in any way. Moreover, the majority of participants surveyed would be willing to share relevant information on their topics of interest with other members, but indicate they need to be solicited. While they were solicited in the months following the end of this part of the research, their habits did not change much. They still apparently don’t feel confident enough to go onto the blog and write things in there, and they still wait to receive information from others, not contributing much themselves, except when specifically solicited to participate in a meeting, conference or research project.

6.2.2 Trust, time and leadership towards a common goal

Our interview results appear very relevant to identify the causes of low participation in knowledge sharing in the CoP as well as possible actions to revitalize the community of practice in the future and generate knowledge-sharing. Among the actions which could be put forward on the basis of the interviews, let us mention the following: developing trust, releasing time for participants and developing leadership towards a common goal.

6.2.2.1 University vs. organizational representatives: need to develop trust and a common goal

Depending on the member's profile (academic vs. community), expectations and modes of operation in the CoP are somewhat different. For example, rates of participation and work are different, the level of expectation in relation to the solicitation is different, and the degree of commitment is also variable, especially in light of the fact that a person can be actively involved in ongoing research or merely participates in conferences and activities. In any case, participation in online tools is low, except for email exchanges; indeed, there seem to be a good number of emails between a certain number of individuals in the network, but we could not keep track of emails, many of which are bilateral ones (peer to peer). Moreover, non-academics seem to find the pace of research projects to be slow, much slower than what they are used to, especially when they are working with students, who also may have courses and other activities. While academics may complain on the time it takes to get access to research places, non-academics want quick results. So the time issue is crucial, both worlds not functioning at the same speed.

One social partner says: "Take the example of M., a PhD student, I met her last summer, I invited her to meet with management, we created a project but I did not see her for one year. Mrs. M. came back last week for the same project. So, things have not moved over almost a year. I had invited her to our Hi Tech event, I introduced her to business leaders in our industry, the stage was set so that she could go to meet them, advance the project, then it came back just this year. So as I say, one participates at the rate at which it is sought." This may be a particular case, where a student was overwhelmed with her courses, but in the same vein, another said: "We had a meeting with Mr E. (University professor), who wanted to initiate a project on subject X. I met him, I sent various tools that we have produced on this subject, but it took time before I heard back from him". Partners from the community environment start to develop distrust when they have such experiences. It thus appears that the issue of trust and time are quite central to engaging in active participation in a CoP, or other research project designed as such.

6.2.2.2 Rotation of organization representatives: need for stable partners to develop trust

In community organizations however, another issue is that representatives of the organization often change, and thus, participants do not always have the time to become what is called "experts" in the CoP language, that is to feel comfortable in knowledge exchanges, or even to master the subject of research. There is frequent turnover among the personnel of such organizations for various reasons (retirement, change of position in the organization, organizational change ...) and thus among representatives in the network. Given these frequent changes, it takes more time to develop trust in the other individuals, to know who you are talking to, what is their position, etc. This is all the more important when there are people from various backgrounds: unions, community groups, business people, government, etc. As some interviewees mentioned, this rotation of participants does not facilitate proper cooperation, because you have to learn to know people, develop trust, etc. The community must be somewhat reorganized each time someone changes. Meanwhile, university researchers do not usually move from their positions, so there is a difference here between the two types of partners. The academics become better known and identified, while the community reconstructs itself, to a certain extent, regularly.

6.2.2.3 Active Involvement vs Commitment

Among the members who admit they do not participate very actively in the network, most say they are committed, even very committed in some cases. Again, these are probably those who are more actively engaged in research or specific projects that are more involved in exchanges. Nevertheless, for most participants, the perception is that the sense of belonging is separate and not necessarily proportional to the rate of active participation in the community. This is something which appears new since in most of the literature on CoPs, it seems “natural” or automatic that participants will be committed and actively involved in the community.

6.2.2.4 Lack of a core of leaders

As noted by [11] there are three levels of participation in a community of practice: (1) the *core* is heavily involved and takes leadership in meetings and projects, (2) another group, considered active, consists of members attending and participating regularly but who are not leaders, and (3) the “peripheral” group consisting of rather passive participants who are satisfied with their low level of involvement. According to Wenger, the third group usually represents the majority of the community. This is also the case in our network, so in terms of what is new and what has changed in knowledge sharing and CoP issues, we can say that this proposition on the three levels of participation still seems to hold true, even if people should have become more accustomed to ideas of knowledge sharing and CoP, as well as the tools such as blogs and others used to support the knowledge sharing. However, the group of leaders seems much less present in our community. In fact, this group, which considers itself an active and committed group, takes little initiative and awaits requests from the coordination. Some participants that we identified as part of this (potential) group of leaders and recognized as being very active feel peripheral or even go so far as to ask for more leadership in the network. That is to say they are looking for leadership without thinking they could be leaders themselves or participate more actively and even participate in this leadership, which should be the case in a research partnership designed to co-construct knowledge, and which should not be based on hierarchical (top down) relations. Thus, they do not seek to share knowledge, but rather wait to be asked for contributions. There seem to exist only two levels of participation in our community, the active level and the peripheral level, a problem that was also observed in other cases of CoP [7].

We found that the director of project, administrative coordinator, web coordinator as well as the community and university representatives at the executive board for the core participant group. They search for information and distribute it through the web, add elements on the blog, etc. On the other hand, the peripheral group, which is also composed of university and community people, is satisfied with receiving the newsletter and going on the website to get information when they are informed of its publication.

7. Discussion

This is an emerging community, and exchanges could develop over the years, as people get to know each other, as trust builds up and also possibly as they become more familiar with the ICT tools. The first questionnaire on the social networking tools showed some ignorance of the main tools, including blogs. Few members knew more than three social networking tools (if we include Facebook!). Accordingly, we obtained mixed results in terms of participation in the CoP, at least with regard to communication tools. Indeed, members feel involved in the network, participation in seminars and conferences is good, and there are ongoing discussions on the projects under the three main research themes. Some community partners even admitted: "Through the network, I am in contact with other groups, for example organization X, so the network has made it possible for me to create new links, to interact with Organization X and Y ". There are many email exchanges, although we do not monitor all emails sent from one member to another; however, it seems difficult to develop trust throughout the network and activate online exchanges, through the web site / blog and other tools.

However, ignorance of web communication tools and their use is not the only explanation for the low participation using these tools. Indeed, the participation rate to questionnaires and telephone interviews is low. If only to get answers to questionnaires or confirmation of appointment for an interview via email, it took many emails (often unanswered). Yet the majority of participants surveyed said they want to be contacted by email.

Despite poor results in terms of questionnaire or participation with the communication tools on the website (blog, etc.), the community seem to be becoming more and more active over time. Thus, time and trust appear to be the essential elements for a CoP to actually emerge.

As mentioned above, in terms of what is new and what has changed in knowledge sharing and CoP issues, Wenger’s proposition on the three levels of participation still holds true, even if people should have become more accustomed to ideas of knowledge sharing, as well as the tools such as blogs.

8. Conclusion

What makes communities of practice successful over time is their ability to generate enough excitement, relevance and value to attract and engage members [29]. Although many factors such as management support or an urgent problem to solve can inspire a community, nothing can replace the sense of vitality, according to these authors. There is a real challenge to create this vitality (or drive) in a community of practice. We note that although in theory, to set up a community of practice and keep momentum may seem a simple thing, in practice it is somewhat more challenging.

Given our results, where people say they feel part of the community, feel engaged as well, we tend to think that a few more elements may be needed beyond the sense of vitality. We observed that some do not participate because of time, others do not take leadership, but they expect things to be organized for them. They look at the website, read the newsletter, but don't find the time to contribute actively, thus staying in the status of "peripheral participation". So if Wenger's three levels of participation still hold true, it may be necessary to revise the conditions of successful CoPs and add the elements of trust, time and something like "togetherness" or a real feeling of "community", the latter being necessary when people come from different organizations, with different priorities and different time spans.

The evidence gathered indicates that we must take into account the practical distinction between the various types of participants in the CoP, the fact that they need to build trust relationships, to get to know each other, to also try to function in the same timeframe, the university and research time frame obviously being different from that of business, or union people, but even of community groups as we found here.

The analysis of this research partnership in terms of knowledge sharing within a community of practice has made it possible to identify some difficulties associated with the active participation in knowledge sharing, which are supposed to be the norm in CoPs. The fact that this is a rather large research project in partnership, covering three main topics and including several studies, means that all partners do not feel concerned at all times and it is true that partners tend to be more active in specific research projects than in the general knowledge sharing which is one of the objectives of the CoP/research project. However, as the objective is to involve all in knowledge sharing, it is important to identify sources of difficulty in this regard and the time and trust issue come out as the main elements. Of course, these cannot be confirmed with a single case study, and future research will be needed to confirm the importance of these two dimensions, but we argue that these elements have been neglected in the literature and should be put forward.

Indeed, the literature on CoPs often ignores or underplays the role of exogenous factors such as time constraints and lack of trust or previous knowledge of CoP participants and thus presents an idealized view of the way such groups function. Our research highlighted the fact that communities do not exist in a vacuum and the explanation for what happens within them does not lie solely within the way the group interacts. This means that while there is much research concentrated on what happens “within the CoP”, research should probably look more towards the environment and what happens “around” the group, including time commitments of the participants and the time needed to develop trust within the group. The community we studied is composed of many members coming from different organizations and who therefore have different priorities. Obviously, if we want to get such communities actively engaged in knowledge sharing, a few prerequisites need to be taken into account.

In conclusion, we set out to determine the factors that facilitate or hinder cooperation within a community of practice composed by two groups of actors, community and university actors. We found that the different affiliations might impact on the CoP and lead to less active participation. This is the case, not because of opposition from the employer or power issues, but mainly for lack of time. We showed that it is not as easy as the literature seems to indicate to create a CoP and ensure knowledge sharing within it.

Indeed, the fact that members come from different organizations, are not bound by the same objectives, and need time to build trust between them explain the difficulties in creating a CoP,. While some may consider that the deliberate creation may affect the success in achieving participation, it must be remembered that no individual was forced to enter the research CoP. The organizational context is probably more important, and the presence of a university actor, alongside actors from community and union groups could be analyzed in more detail in future research, although it did not come out as a problem or issue in the interviews. Also, while many said they were committed to the CoP, they did not engage as actively as could be expected, given this commitment to the project. This time element and the distinction between commitment and active engagement in the CoP do not appear to have been highlighted in previous literature and this is surely something to put forward and be attentive to if organizations want to build active learning CoPs.

This is something which appears new to us since in most of the literature on CoPs, it seems “natural” or automatic that participants will be committed and actively involved in the community. Our research shows that this is not necessarily the case that they can have time constraints and clearly need time for trust to develop within the group in order to foster active participation. This is interesting since it confirms what [39] found, that is the importance of trust, limited to blogging activities in their case. Our results go a little further in indicating that it is not only in blogging that trust is essential, but in any type of knowledge-sharing activity which is designed in the form of a CoP.

9. References

1. Cabrera A. and Cabrera E. F., *Knowledge-sharing Dilemma.* Organization Studies, 2002. 23(5): p. 687-710.
2. Bourhis A. and Tremblay D.-G., Rapport de recherche du projet Télétravail: concilier performance et qualité de vie. 2001, CEFRIO: Québec.

3. Wenger E., McDermott R., and Snyder W. M., *Cultivating Communities of Practice: A guide to Managing Knowledge*. 2002, Boston, MA: Harvard Business School Press.
4. Hildreth P., Wright P., and Kimble C. *Knowledge management: are we missing something?* in *4th UKAIS Conference*. 1999. London: York, UK: McGraw Hill.
5. Wenger E. Communities of Practice: Learning as a social system. 1998 [cited.
6. McDermott R., Knowing in community : 10 critical success factors in building communities of practice. IHRIM Journal, 2000. March 2000.
7. Bourhis A. and Tremblay D.-G., Les facteurs organisationnels de succès des communautés de pratique virtuelles. 2004, Cefrio: Québec. p. 140.
8. Guillaume O., Recherches partenariales : coordination et coopération entre chercheurs d'entreprise et chercheurs universitaires. Interventions économiques 2011. 43.
9. Hatchuel A., *The two pillars of new management research.* British Journal of Management, 2001. 12: p. 33-39
10. Starkey K. and Madan P., *Bridging the relevance gap: aligning the stakeholders int the future of management research.* British Journal of Management, 2001. 12(Supplement 1): p. 3-26.
11. Wenger E., Cultivating Communities of Practice: A quick start-up guide. 2002.
12. Lave J. and Wenger E., *Situated Learning. Legitimate peripheral participation*. 1991, Cambridge: University of Cambridge Press.
13. APQC, *Building and Sustaining Communities of Practice*. 2001, American Productivity and Quality Center: Houston, TX. p. 205.
14. Mitchell J., The potential for communities of practice to underpin the national training framework. 2002, Australian National Training Authority: Melbourne. p. 104.
15. McDermott R., Learning across teams: How to build communities of practice in teams organizations. Knowledge Management Review, 1999. **8**(may/june): p. 32 - 36.
16. Gherardi S. and Nicolini D., *The organizational learning of safety in communities of practice.* Journal of Management Inquiry, 2000. **9**(1): p. 7-18.
17. Gherardi S. and Nicolini D., *To transfer is to transform : The circulation of safety knowledge.* Organization, 2000. **7**(2): p. 329-348.
18. Tremblay D.-G., Concertation éducation travail - Politiques et expériences, in Éducation - Recherche, Hardy M., Editor. 2003, PUQ.
19. Lave J., Cognition in Practice: Mind, mathematics, and culture in everyday life. 1988, Cambridge, UK: Cambridge University Press.
20. Kearsley G. *Explorations in Learning & Instruction: The Theory Into Practice Database*. Situated Learning (J. Lave) 1994-2010 [cited 27 août 2010]; Available from: http://tip.psychology.org/.
21. Cohendet P., Créplet F., and Dupouët O., *Innovation organisationnelle, communautés de pratique et communautés épistémiques : le cas de Linux.* Revue française de gestion, 2003. 147(Nov.- déc. 2003): p. 99-121.
22. Cothrel J. and Williams R. L., *On line communities: Helping them form and grow.* Journal of Knowldege Management, 1999. **3**(1): p. 54.
23. Ardichvilli A., Page V., and Wentling T., *Motivation and barriers to participation in virtual knowledge sharing in communities of practice.* Journal of knowledge management, 2003. **7**: p. 64-77.
24. Créplet F., Pour une approche des PME : leur évolution et leur développement dans une perspective cognitive. Entre communautés d'action et communautés de savoir. 2001: Strasbourg.

25. McDermott R. and O'Dell C., *Overcoming cultural barriers to sharing knowledge.* Journal of Knowldege Management, 2001. **5**(1): p. 76-85.
26. Sveiby K.-E. and Simon R., *Collaborative climate and effectiveness of knowledge work - an empirical study.* Journal of Knowledge Management, 2002. **6**(5): p. 420–433.
27. Tu C.-H., *The management of social presence in an online learning environment.* International Journal on E-learning, 2002. April-June: p. 34–45.
28. Wasko M. and Faraj S., "It is what one does": why people participate and help others in electronic communities of practice. Journal of Strategic Information Systems, 2000. **9**: p. 155-173.
29. Wenger E., McDermott R., and Snyder W. M. *Seven Principles for Cultivating Communities of Practice*. Harvard Business School Working Knowledge e-mail newsletter (March 25, 2002) 2002 [cited from http://hbswk.hbs.edu/ archive/2855.html].
30. Kimble, C., C. Grenier, and Goglio-Primard K., *Innovation and knowledge sharing across professional boundaries: Political interplay between boundary objects and brokers.* International Journal of Information Management, 2010. **30**(5): p. 437-444.
31. Esnault L., Zeiliger R., and Vermeulin F. On the Use of Actor-Network Theory for Developing Web Services Dedicated to Communities of Practice. in EC-TEL 2006 Workshop: Innovative Approaches for Learning and Knowledge Sharing. 2006.
32. Spinuzzi C., *The methodology of participatory design.* Technical Communication, 2005. 52(2): p. 163-174.
33. Clement A., Computing at work: empowering action by low-level users. Commun. ACM, 1994. **37**(1): p. 52-ff.
34. Bowker G.C. and Star S.L., *Sorting Things Out, Classification and its consequences*. 1999, Cambridge, MA: MIT press.
35. Gasson S., *A genealogical study of boundary-spanning IS design.* European Journal of Information Systems, 2006. 15(1): p. 26-41.
36. Hildreth P and Kimble C., *The Duality of Knowledge.* Information Research, 2002. **8**(1).
37. Psyché V., Rôle des ontologies en ingénierie des Environnement Informatique pour l'Apprentissage Humain (EIAH) : Cas d'un système d'assistance au design pédagogique, in Informatique. 2007, Université du Québec à Montréal: Montréal. p. 527.
38. Psyché V., Mendes O., and Bourdeau J., Apport de l'ingénierie ontologique aux environnements de formation à distance, in STICEF - Technologies et Formation à distance, Hotte R. and Leroux P., Editors. 2003, INRP. p. 89-126.
39. Sangmi Chai and Kim Minkyun, *What makes bloggers share knowledge? An investigation on the role of trust.* International journal of information management, 2010. **30**(5): p. 408-415.

Diane-Gabrielle Tremblay is professor of labour economics at the Télé-université of the University of Québec, Canada; she has been appointed Canada Research Chair on the socio-economic challenges of the Knowledge Economy in 2002 (http://www.teluq.uqam.ca/chaireecosavoir/) , and renewed in 2009 , and is director of the research center CURA on work-life balance over the lifecourse (www.teluq.da/aruc-gats). She has been invited professor at the Sorbonne-Paris I, and Universities of Lille 3, Angers, Toulouse, Lyon 3, Louvain-la-Neuve, in Belgium, University of social sciences of Hanoi

(Vietnam) and the European School of Management. She has published many articles in various journals such as the *Applied Research on Quality of Life, Social Indic*ators *Research*, the *Journal of E-working, the Canadian Journal of Urban Research, International Journal of Entrepreneurship and Innovation Management, Canadian Journal of Communication, Canadian Journal of Regional Science, Leisure and Society, Women in Management, Géographie, économie et société, Carriérologie, Revue de gestion des resources humaines*, and others.

Valéry Psyché holds a Ph.D. in Cognitive Informatics from the University of Quebec at Montreal, (UQAM). She began her research career at the Research Centre LICEF TÉLUQ where she is associate researcher. She has been involved for 10 years in numerous projects, mainly in the field of educational technology and cognitive informatics. As a researcher, she developed a postdoctoral research project aimed at creating a community of practice for CURA (Community University Research Alliances). She keeps research collaboration with a world-class center in ontological engineering at Osaka University (Japan). She also teached at the University of Montpellier (France). She has published articles in the proceedings of many international conferences such as Artificiel Intelligence in Education (AIED), Intelligent Tutoring Systems (ITS), and LORNET Research Network on Intelligent, Interactive, Learning Object Repository Networks (I2LOR), and in various journals such as Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (STICEF), Sociologies ; and others.

Received: December 23, 2011; Accepted: February 12, 2012.

DOI: 10.2298/CSIS111214018D

Web Service Support for Collaboration between Demographers

Mirjana Devedžić, Vladan Devedžić, and Sonja D. Radenković

University of Belgrade, Serbia
devedzic@fon.rs, {mdevedzic, sonjafon}@gmail.com

Abstract. The paper promotes the use of novel Web services in the daily work and research of social scientists and other professionals. The case presented in the paper pertains to demographers and their research, but the technology used is generic and can be easily instantiated for use by other social science researchers. Specifically, the case covers facilitating collaboration between a university research group in the field of demography and professionals in the field of demographic statistics. The technology used is a set of new Web services developed as parts of an EU research project. The paper explains the case itself and the motivation for using the services, describes the services themselves, and discusses the experience acquired and the benefits and lessons learned by using the services so far.

keywords: Demography, Web services, collaboration, research, content & knowledge provision.

1. Introduction

The term "Web service" is somewhat ambiguous; different people often mean different things when they say "Web service" [10]. In this paper, the term "Web service" is understood as an autonomous, self-contained, and self-describing software component accessible through a Web server that provides functionality through a standardized set of interfaces. Web services can be combined in and used by other applications, can communicate using open protocols, and can be orchestrated to implement a more complex functionality.

To many common end users, Web services are things like instant messaging services (e.g., Yahoo! Messenger, Windows Live Messenger, Skype, and AIM) and social networking and microblogging services (such as Facebook and Twitter). Very often, people also use various booking services (such as accommodation booking at Booking.com), payment services (e.g., PayPal), and buying/selling and auction services (such as eBay).

What kinds of Web services are of interest to researchers? To what extent researchers have adopted and do use Web services in their daily work?

To help answer these questions, the Research Information Network from UK has recently conducted a comprehensive study [6]. Although focusing on Web 2.0 services in the first place, the study has also indicated typical uses of other Web services by researchers of different profiles and backgrounds. The results have shown that the majority of researchers predominantly use digital library services and services related to submission and publishing of their research in scholarly journals. Many of them also use well-known generic tools and services such as Google Scholar (73%) and Wikipedia (69%). On the other hand, while most researchers have a positive attitude towards Web 2.0 services (blogging, social networking, social bookmarking and tagging, wikis, RSS feeds and the like), only a few have made them a routine part of their working life. These few researchers typically use SlideShare (http://www.slideshare.net/) as a presentation sharing service, social bookmarking and tagging services such as Delicious (http://delicious.com/), and discipline-specific services such as http://arts-humanities.net, a 'hub' for teaching and research in the digital humanities. Very few use blogging to publish work in progress or to comment other researchers' work, although their number is slowly increasing. Not surprisingly, the majority of those researchers who have adopted Web services as tools for their daily use come from computer science, math, and engineering; very few come from social sciences. The study concludes with the recommendation to raise awareness (among the researchers) of tools and Web services, and the uses to which they can be put, as well as to provide guidance and training.

However, a notable exception in the study is related to collaborative research: the use of services is positively influenced by researchers' involvement in collaborative research activities. Here collaborative research can be work as part of a local team, work with collaborators in different institutions, participation in an informal, local research network, and participation in wider, discipline-based research networks.

This last fact has been one of the starting points for the work presented in this paper. Another one was the recent development of useful Web services that support collaborative learning, work, and research; the services have been developed within an EU-funded, 3-year research project called IntelLEO (http://www.intelleo.eu). A suitable case of collaboration between researchers and other professionals working in the field of demography has triggered the work.

2. Problem statement

The objective of the research presented here is to promote the use of novel Web services in the daily work and research of social scientists and other professionals. The paper describes a case related to the field of demography, but the approach can be easily extended to other social sciences as well. Moreover, extension to fields outside of social sciences is quite feasible as well.

Specifically, the paper focuses on the following questions:

- What specific kinds of Web services can be of interest to demographers, given the intrinsically collaborative nature of their daily work and research?
- What exactly are the benefits demographers get from using such Web services?
- What does it take to start using specific Web services in collaborative work and research in demography, and how to start?

The obvious underlying assumption here is collaborative work. However, the approach to collaboration covered in this research is not restricted to collaboration between individuals; the most suitable cases are actually those of collaboration between different teams and/or organizations.

3. Motivation

Much of the work done by demographers is intrinsically collaborative. This collaboration often includes researchers and professionals from more than one institution. For example, researchers from universities typically rely on various databases and statistics related to the population in a certain region, or in a country, or in a city; in many cases, these include various census data and data about events related to vital statistics, such as births, marriages, deaths, and so on. Such databases are usually maintained by institutions like national and local bureaus of statistics, as well as by various governmental and provincial offices. Hence university research groups often collaborate with professionals from such institutions. Another typical example is a whole category of interdisciplinary projects, usually initiated by certain ministries and other governmental institutions, where demographers from universities and demographic research centers are asked to participate to spatial planning, social care, educational policy development, and similar population-related programmes.

In all such cases, a common problem faced by both the researchers and the other professionals participating to the collaborative activities is that not all data, statistics, documents, facts, and knowledge are readily available. In spite of the fact that censuses and vital statistics provide a number of useful data to the researchers, there are always new demographic insights, new social processes and circumstances, and new externalities that the researchers are aware of, but are not tracked and updated regularly in the existing databases. It also happens that data useful to researchers are difficult to obtain due to the other institutions' organizational restrictions and policies. Likewise, federal, state, and provincial governments sometimes contract demographers and other researchers to conduct studies and surveys that diverge from established research methodologies. These often assume studying international statistics, documents and regulations, as well as additional data collection and processing to fit the contractual requirements.

Table 1 shows examples that illustrate this common problem and the points made here.

Table 1. Topics and phenomena of interest to demographers, but not well covered in information sources related to census data and vital statistics (The topics/phenomena indicated in the table illustrate the case of Serbia, but are also observed in many other countries as well)

Topic / Phenomenon	Explanation	Missing data / Further work needed	Target information of interest
Cohabitations	Unmarried couples living together	Number of cohabitations, their durations, resulting fertility, number of cohabitations ending in marriages	More complete picture about the size of married population, celibate, and attitudes towards traditional life; more accurate classification of population according to their marital status
Population ageing	The number of elderly citizens vs. the number of young ones	Social status of elderly citizens, their habitation types, health conditions, sources of income, level of poverty	Estimates of population ageing trends, related actions of local governments, regional and international contexts of ageing
Brain drain	Emigration of highly educated labor force	Demographic data before and after emigration, such as marital status and fertility, professional activities, type of college degree, level of further education	Migration trends to help the government create appropriate population policies
Social mobility	Horizontal and vertical mobility and "generation shift"	Mobility related to ethnical recognition (especially for minorities) and types of habitats	Quantitative insights into how social, political, financial, and other crises affect social mobility
Population activity	Citizens' primary and secondary (additional) activities	Additional activities of a part of the population struck by poverty and turbulent political events	More complete insight in the share of additional activities in the total population activities

It is often advocated by both parties that a major step towards alleviating this problem would be to develop and maintain a repository of various documents, research material, online tools and resources, and other information that all interested individuals and organizations could use in their daily work and activities. Uploading and annotating new relevant resources in this repository, as well as removing or repurposing existing ones, can help discover new trends and interests of researchers and institutions, indicate specific services of interest to organizations and individuals that should be developed to support and facilitate their activities, and point out emerging topics and profiles of professionals required to cover them in future projects.

Such a repository should be supported by appropriate, intuitive, and easy-to-use Web services to allow for easy access to the resources stored therein. In addition, using these Web services should enable interested individuals and organizations not only to collaborate over specific resources in their projects, but also to contribute to the repository with newly created and/or discovered resources. And that is where the IntelLEO Web services come into the play.

4. IntelLEO services

IntelLEO project aims at developing Web services to support collaboration between two or more organizations. An important underlying assumption here is *temporal integration* of the organizations – they are supposed to work collaboratively, over a certain period of time, and enjoy the benefits of collaborative learning and knowledge building (LKB) activities. This temporal integration can be organized around a collaborative project, or can stem from other activities of mutual interest for the organizations involved (e.g., organizing a seminar, or establishing a research communication over a certain topic).

In addition, such collaboration can origin from objectives and interests of the organizations as business entities, but is most fruitful when simultaneously accommodates personal objectives and interests of the individuals (i.e., employees) involved. For example, an employee of one organization may need to extend her knowledge of a certain topic or certain practices in order to get her work done. What IntelLEO services enable her to do is to find appropriate online resources most efficiently, discover people from the organizations involved who are most competent on the topic, create an online working group with them, share her knowledge and experience with the other members of the group, learn from the experience of the other group members, and the like, and yet stay within the business policies and interests of her organization and contribute to fulfilling them.

Specifically, IntelLEO services include:

- Human Resource Discovery – searching for human expertise in the organizations involved, to support LKB for individuals and groups, according to predefined objectives; checking for experts' availability;
- Working Group Composition – Proposing a suitable working group based on identified available expertise, individual competences and experience, and organizational objectives;
- Collaboration Tracing – Tracing of LKB collaboration contexts, events and activities, interaction (type, frequency etc.), feedback provided, team results, and resource usage;

- Content/Knowledge Provision – Context-dependent and proactive discovery/delivery of LKB resources within the organizations involved, including discovery/delivery of information and knowledge on individuals and working groups;
- LKB planning – Selection of the most appropriate sequence of LKB activities and criteria to provide resources in a specific context (e.g., for a specific individual and/or group)
- Organization policy – Identification of rules and/or objectives relevant for a specific content/context and filtering (from a set of provided resources) those that fit with the identified organizational rules, objectives, and strategies.

In order to use IntelLEO services efficiently, collaborating organizations are supposed to maintain a common repository of resources of interest, with specific and controlled access rights of the repository users. IntelLEO project provides technical and administrative guidelines on how to set up such a repository.

5. Application case

Through private connections, IntelLEO dissemination activities, and another research project, a group of demographers from the Institute of Demography (ID) at the Faculty of Geography of the University of Belgrade has got introduced to the IntelLEO Web services. Although the IntelLEO project is not completed yet, and its services are still getting improved, ID wanted to experiment with the current versions of the services and find out if they can help them address some of the issues described in the section on motivation. One way or another, ID found all IntelLEO services to be of interest to them, but was mostly interested in the Content/Knowledge Provision service (CKP).

Together with researchers involved in both IntelLEO and the other research project mentioned above, ID has set up a case, a repository, and a testbed for working with CKP. They have also come up with several typical scenarios for use of CKP by demographers. The scenarios considered here are related to the issue of population ageing.

Population ageing is the increase in the number and proportion of older people in society [9]. It has three possible causes: migration, longer life expectancy (decreased death rate), and decreased birth rate. Being a global phenomenon, population ageing has a significant impact on society.

5.1. Practical problems that hinder research on population ageing

Data on elderly population of interest to demographers are only partially available straightforwardly from census sources. Other data require additional census data processing and cross-referencing (e.g., do elderly people live alone, do they live in multi-generation families, are they married, what are their education levels and nationalities, and so on).

Still other relevant demographic data on elderly people are not available from census data and in many countries are not even tracked systematically (e.g., what are their health conditions, their typical diseases, who helps them, do they have children who visit them and take care of them, what are the activities they can conduct on their own, what is their degree of invalidity, how they spend their time, what are their incomes, and the like).

Some institutions may even keep track on some of these data (e.g., healthcare institutions and NGOs), but when it comes to using the data in demographic research it takes considerable time and effort to get the right information and to get through various administrative and data access procedures. Worse still, when professionals from national and local bureaus of statistics, public health, and other institutions call for meetings with demographers in order to get suggestions and advice on how to extend and modify relevant forms for the next census, it usually turns out to be too costly to do it or that the time remaining is too short.

5.2. Application scenarios

As scenario 1, assume that Joan, a professional from the National Bureau of Statistics (NBS), is working on a publication that would cross-reference selected data from the latest census to focus on (and cover in more detail) elderly population only. She is interested to learn what specific data should be cross-referenced, what composite indicators should be calculated, and how to present them in the publication to be of most use to the readers. Using IntelLEO services and especially CKP, Joan can be in a regular (online) touch with demographers from ID and have access (through CKP) to the repository of relevant resources exemplifying the issues of interest to population ageing studies. Moreover, she can upload (to the repository) resources available at NBS, at another institution, at an NGO, or elsewhere, and ask researchers to check them out and comment them in order to help the new publication include more relevant data and exclude irrelevant ones.

Scenario 2 is another hypothetical scenario that ID has come up with. It might involve Helen, a young PhD student and researcher from ID who currently works on her thesis related to population ageing in a selected region of the country. She collaborates with the regional government institutions and NBS alike, and has already contributed a number of useful documents to the common repository. She is always eager to learn what data relevant for her research can be obtained from these institutions, and at what cost. She is also interested in finding out how easy it is for these institutions to start tracking specific data they did not track in the past, end is ready to illustrate how it is done in other regions, or in other countries. She can do it by using CKP to upload specific resources to the repository and by asking people from the other institutions to use CKP to take a look at them.

Two issues are important here, in both scenarios. First, all resources uploaded to the repository using CKP can be appropriately annotated to facilitate subsequent search and use. Part of the annotation happens automatically and the end users do not have to take care about it. The other part is done by end users in the form of tagging. Annotations help users to get to the relevant material easily and to recognize other users interested in the same resources and topics.

Second, remember that all IntelLEO services, including CKP, facilitate online *collaboration* between different individuals, groups, and organizations. As the next subsection illustrates, it is likely that in the above two scenarios Joan and Helen will "run onto each other" through the repository resources and CKP. Joan can always see in CKP that a document in the repository relevant for her publication is annotated by Helen (and possibly some other users interested in the same document and/or topics). The opposite is also true. Thus Joan may want to use other IntelLEO services to initiate more intensive collaboration with Helen.

5.3. Using CKP

To use CKP and other IntelLEO services, one needs an account. Collaborating organizations may want to restrict access to their shared repository, but getting an account is free. Installing IntelLEO services and a shared repository of resources on a server and running them from it is a straightforward task for a server administrator.

On the client side, all one needs to do is install a Web browser add-on for interacting with IntelLEO services; it is done in a couple of mouse clicks. The add-on is represented in the browser as a toolbar (called the IntelLEO toolbar) with three buttons – *IntelLEO*, *Upload*, and *Resources*, Fig.1. (The toolbar can be easily hidden/shown at any time, using the browser's Tools/Options menu.)

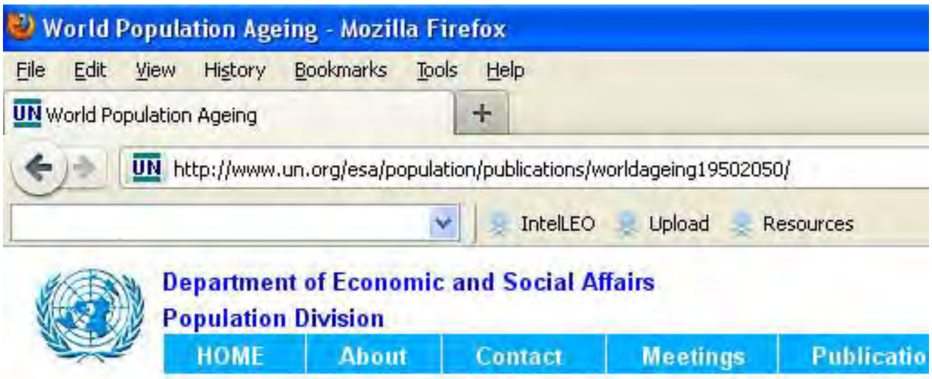
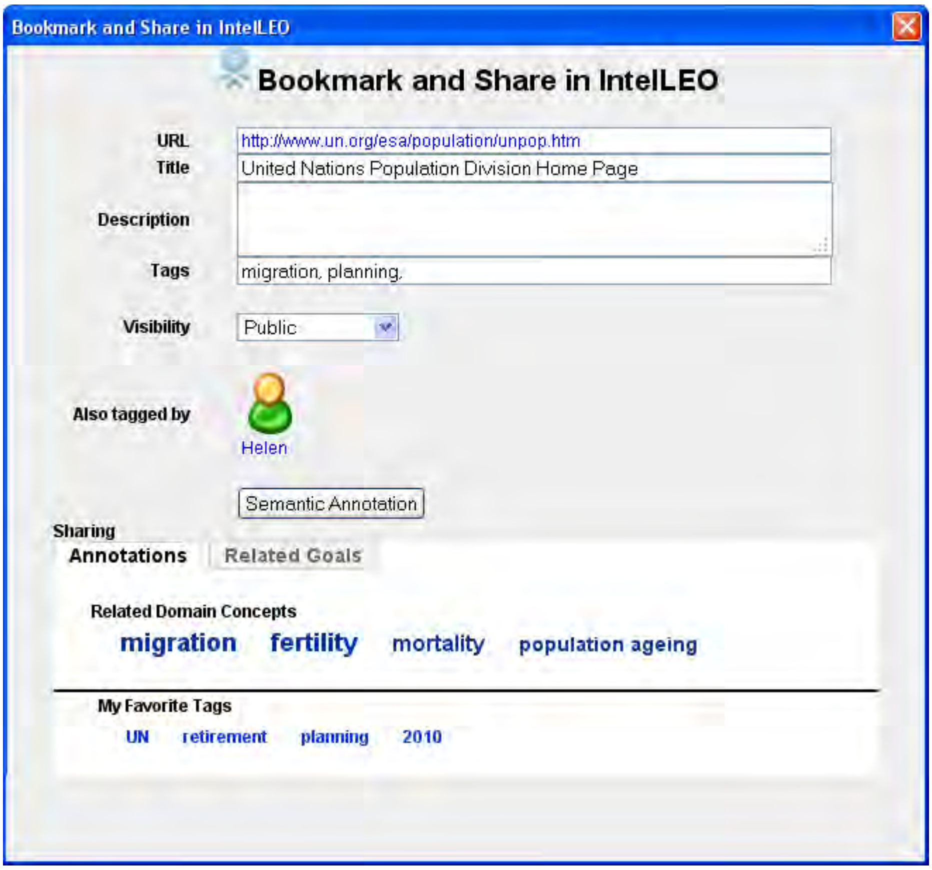

World Population Ageing - Mozilla Firefox
File Edit View History Bookmarks Tools Help
UN World Population Ageing
+
UN http://www.un.org/esa/population/publications/worldageing19502050/
IntelLEO
Upload
Resources
Department of Economic and Social Affairs
Population Division
HOME
About
Contact
Meetings
Publicatio

Fig.1. IntelLEO services toolbar

Now assume that Joan (or Helen) from the scenario(s) described above wants to bookmark a resource in (and possibly upload it physically to) the common repository. The resource can be an online one (a Web site, a Web tool, an online document, and the like), or it can be a local document stored on her computer. She opens/accesses the resource in her browser and clicks the *IntelLEO* button in the IntelLEO toolbar. A popup window like the one shown in Fig. 2. opens.

In the upper part of the popup window, a few standard terms from Dublin Core Terms vocabulary (http://dublincore.org/documents/dcmi-terms/) instantiate the usual metadata for the resource. If any of them is not retrieved automatically from the resource, it can be filled manually if the user wants to do it. The *Tags* field is not mandatory, but it is highly recommended that the user tags each resource appropriately, for it may benefit all other users of the repository. *Visibility* is related to the access restrictions that can be set for the resource being bookmarked in the repository. The possible values that can be set include *public*, *organization*, *group*, and *private*.



Bookmark and Share in IntelLEO

Bookmark and Share in IntelLEO

URL: http://www.un.org/esa/population/unpop.htm

Title: United Nations Population Division Home Page

Description:

Tags: migration, planning,

Visibility: Public

Also tagged by: Helen

Semantic Annotation

Sharing

Annotations | Related Goals

Related Domain Concepts

migration fertility mortality population ageing

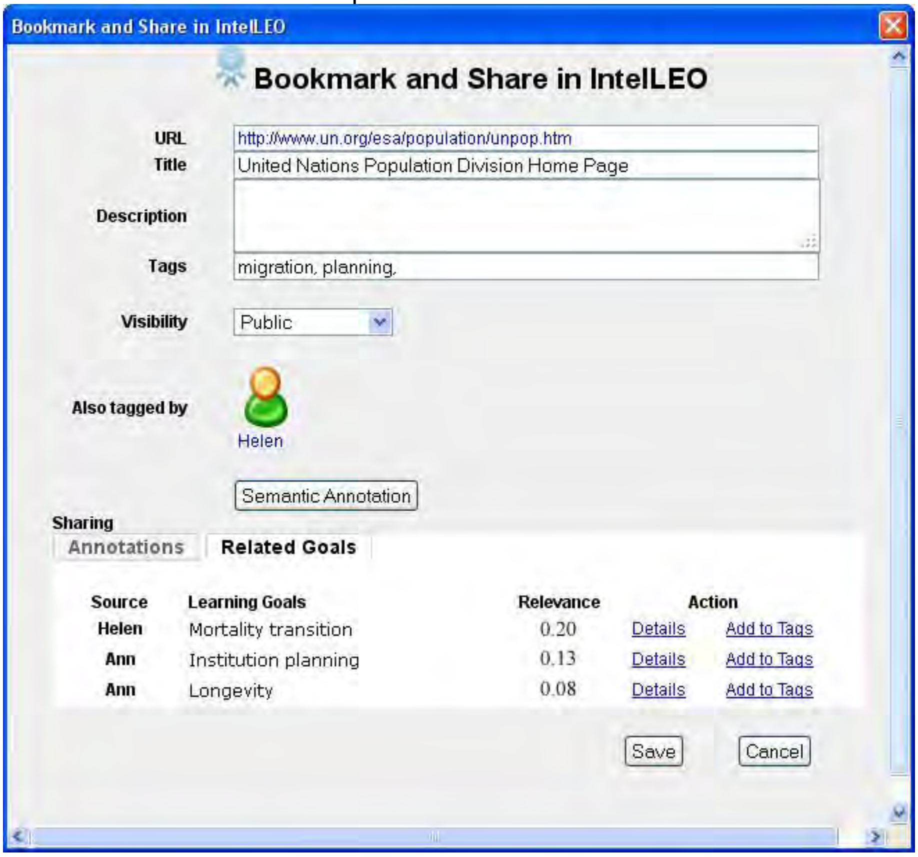
My Favorite Tags

UN retirement planning 2010

Fig. 2. CKP service: annotation and tagging

The lower part of Fig. 2. is more interesting from the end user's point of view. In collaborative work users share common resources, so *Also tagged by* indicates who else has tagged (and hence presumably is interested in) the same resource. The first time a resource is uploaded to the repository, this field is empty; when *another* user accesses the resource through CKP, this field will indicate all the other users who have tagged the resource. The user's own frequently used tags are indicated in *My Favorite Tags*, and clicking any of them adds the tag to the *Tags* field.

When the user clicks the *Semantic Annotation* button, *Related Domain Concepts* are inserted automatically by CKP, provided that CKP is pre-fed by an appropriate domain ontology. Domain ontology contains the domain vocabulary and describes domain concepts and their relationships. Experts from ID and from the IntelLEO project have developed the ontology of population ageing. It is because of that ontology, used by CKP under the surface, that *Related Domain Concepts* in Fig. 2. look as they do – CKP has automatically recognized what the resource is about. This can be a useful addition to the end-user's tags, because it allows for an automatic resource annotation with domain concepts.


Bookmark and Share in IntelLEO
Bookmark and Share in IntelLEO

URL: http://www.un.org/esa/population/unpop.htm
Title: United Nations Population Division Home Page
Description:
Tags: migration, planning,
Visibility: Public
Also tagged by: Helen

Semantic Annotation
Sharing
Annotations
Related Goals

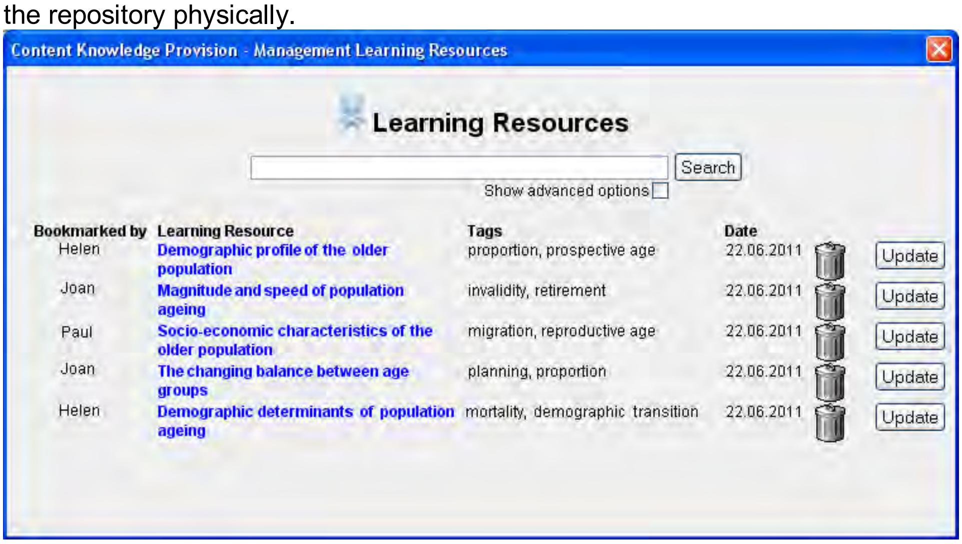
Source	Learning Goals	Relevance	Action	
Helen	Mortality transition	0.20	Details	Add to Tags
Ann	Institution planning	0.13	Details	Add to Tags
Ann	Longevity	0.08	Details	Add to Tags

Fig. 3. CKP service: the Related Goals tab

Joan (or Helen) can also use the *Related Goals* tab in the CKP service, Figure 3. It allows her to interact with the IntelLEO LKB planning and, indirectly, with other IntelLEO services as well. Users of the IntelLEO LKB planning service can specify what they use a shared resource for. In the IntelLEO terminology, it is called a *learning goal*. For example, Helen may have used the resource described in Figure 3 to find out more (i.e., "learn about") *mortality transition* as an important issue in population ageing. Alternatively, she may have just thought that the resource is good for people from ID to be aware of it. In either case, she may have also indicated this intended use of the resource through the IntelLEO LKB planning service. In such a case, the *IntelLEO* tab will show it in the *Learning Goals* column.

Relevance indicates on the 0-1 scale how relevant the resource is for that purpose (i.e. for that learning goal). Clicking *Details* takes the user to the IntelLEO LKB planning service and a more in-depth description of the corresponding learning goal; however, describing all the idiosyncrasies of the IntelLEO LKB planning service is beyond the scope of this paper. Clicking *Add* is another useful CKP feature – it automatically adds the corresponding learning goal to the *Tags* field.

Clicking *Save* saves the bookmark related to the resource in the shared repository. It means that a description of the resource (essentially, most of the things shown in Fig. 2 and Fig. 3) will be uploaded to the repository, and the resource itself will be physically kept elsewhere on the Web. Clicking the *Upload* button in the IntelLEO toolbar, Fig. 1, brings up dialogs similar to those shown in Fig. 2 and Fig. 3, but the effect of clicking *Save* in these dialogs is that both the resource description and the resource itself will be physically uploaded to the repository. This is useful in cases when the user wants to store a copy of an original resource in the shared repository, or when resources originally stored on the user's local computer should be uploaded to the repository physically.



Content Knowledge Provision - Management Learning Resources

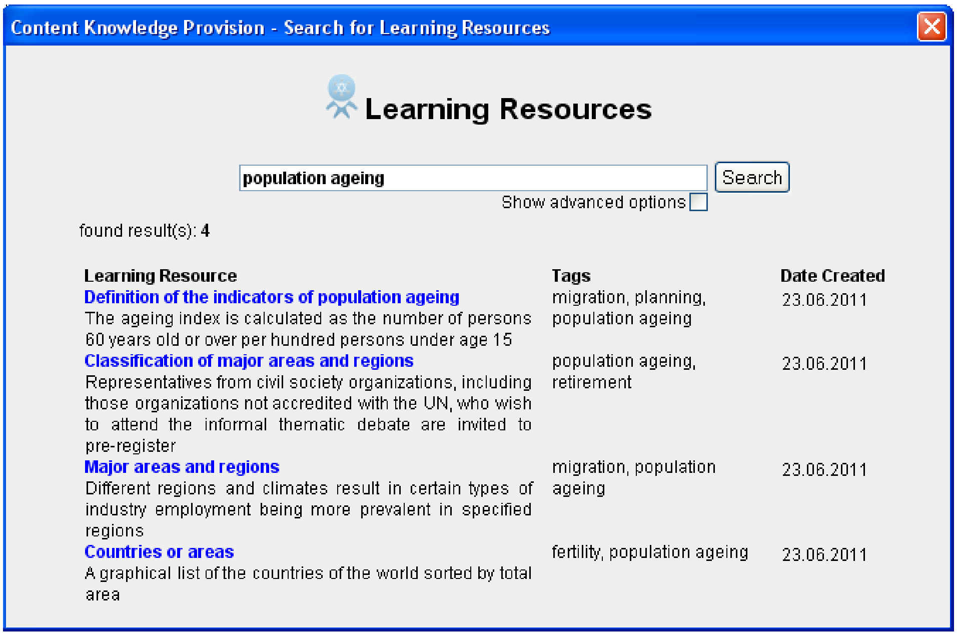
Learning Resources

Bookmarked by	Learning Resource	Tags	Date		
Helen	Demographic profile of the older population	proportion, prospective age	22.06.2011		Update
Joan	Magnitude and speed of population ageing	invalidity, retirement	22.06.2011		Update
Paul	Socio-economic characteristics of the older population	migration, reproductive age	22.06.2011		Update
Joan	The changing balance between age groups	planning, proportion	22.06.2011		Update
Helen	Demographic determinants of population ageing	mortality, demographic transition	22.06.2011		Update

Fig. 4. CKP service: browsing the repository

If Joan wants to browse the repository, she clicks the *Resources* button on the IntelLEO toolbar. It brings up the window like the one shown in Fig. 4. If she clicks a (learning) resource there, a window like the one shown in Fig. 2 opens for that resource. She can also delete a resource by clicking its garbage bin icon, and additionally annotate and update it by clicking the corresponding *Update* button.

Finally, if Joan wants to query the repository for specific resources, she can enter the search term(s) and click the *Search* button. If she also checks the *Show advanced options* box, the search and search results can be further refined. For example, the CKP service can be instructed to query the repository looking for the resources that have the same annotations like the Joan's tags and learning goals. When the *Search* button is pressed, CKP returns a list of the ranked search results. The process of ranking the search results is based on calculating semantic similarity between the resources and Joan's tags and learning goals (as well as her competences, activities, and other details coming from her user profile and from the IntelLEO LKB planning service), depending on the refinements selected when the *Show advanced options* box is checked.



Content Knowledge Provision - Search for Learning Resources

Learning Resources

population ageing

found result(s): 4

Learning Resource	Tags	Date Created
Definition of the indicators of population ageing The ageing index is calculated as the number of persons 60 years old or over per hundred persons under age 15	migration, planning, population ageing	23.06.2011
Classification of major areas and regions Representatives from civil society organizations, including those organizations not accredited with the UN, who wish to attend the informal thematic debate are invited to pre-register	population ageing, retirement	23.06.2011
Major areas and regions Different regions and climates result in certain types of industry employment being more prevalent in specified regions	migration, population ageing	23.06.2011
Countries or areas A graphical list of the countries of the world sorted by total area	fertility, population ageing	23.06.2011

Fig. 5. CKP service: searching the repository

6. Evaluation and discussion

In order to get a feeling of how demography researchers and professionals perceive CKP and other IntelLEO services, a formative evaluation has been conducted with 22 people (4 teachers and 10 PhD and MSc candidates associated with ID, and 8 professionals from NBS and partner institutions). The objective was to find how relevant, how useful, and how interesting would it be for such users to make CKP and other IntelLEO services part of their regular working environment. This section focuses on the evaluation of CKP only.

The participants were explained the functionality of CKP. The moderators have then demonstrated the use of CKP to the participants and asked them to reflect briefly on how (if) they would use it in their work. After they did, the moderators asked them to run CKP themselves, completing tasks that roughly corresponded to the scenarios described in section 4.2. In the end, the participants were asked to fill in specifically designed questionnaires.

In designing the questionnaires, three important sets of guidelines have been used. The first one was borrowed from evaluations of IntelLEO services conducted as part of the IntelLEO project, where one of the objectives has been to collect information about motivational issues related to learning in the workplace [5]. The second one is published at the *Usability and user experience surveys* Web site (http://edutechwiki.unige.ch/en/Usability_and_user_experience_surveys) and covers a number of guidelines and practical examples of organizing evaluations related to perceived usefulness and perceived ease of use of software systems and applications and their user interfaces. The third one was published by [8].

The questionnaires and the questions were related to the following criteria:

- relevance (5 questions) – issues like benefits for individuals and teams/organizations, support when looking for potential resources and partners, support for collaborative work/research, and organizational support
- organization of information (4 questions) – issues like simple and natural dialogue, sequence of screens (interactions), and logic and terminology related to task
- perceived usefulness (6 questions) – issues like accomplishing tasks more efficiently, improve job performance, increasing productivity, increasing inter-organizational collaboration, enhancing effectiveness on the job
- perceived ease of use (6 questions) – issues like learning how to use the service, straightforwardness, using it without written instructions, clear and understandable interaction with the service

The questions were designed as 5-point Likert scale ones (the possible answers being from 1 to 5, corresponding to highly unlikely to highly likely, or strongly disagree to strongly agree, or the like), but free-form qualitative judgments and comments were also encouraged.

Table 2 summarizes the evaluation results. Due to space limitations, the figures shown are rounded averages of the ones obtained from the figures obtained for specific questions related to the same criterion.

Table 2. Results of formative evaluation of CKP by a group of demographers

Criterion	**Percent of Likert-scale answers (1 – lowest, 5 – highest)**					**Notes**
	1	2	3	4	5	
Relevance	7	3	24	34	32	See discussion on free-form comments
Organization of information	21	11	26	37	5	Indicates possible need for changes in the user interface
Perceived usefulness	12	9	12	40	27	-
Perceived ease of use	0	4	12	51	33	-

Two observations follow from Table 2 immediately. First, most of the participants thought that CKP offers functionalities relevant for their work (the Relevance criterion), that it would be useful in their everyday work (Perceived usefulness), and that it is easy to use (Perceived ease of use) – over 50% of all participants answered the corresponding questions with '4' or '5' on the Likert scale. Second, a lot of reluctance is observed in answers to questions related to Organization of information. Many participants did not like some of the labels in the user interface (e.g., Related Goals in Fig. 3), and some thought that certain labels were even redundant (the Annotations section in Fig. 2). This may indicate a need for possible changes in the user interface of CKP.

In addition, the analysis of free-form comments and suggestions revealed that a considerable number of participants thought that organizational support (or restrictions) can be of high importance for adoption of Web services such as CKP. Free distribution and share of certain resources may be difficult to achieve in practice, which may hamper a wider adoption of CKP among professionals.

Some participants also did not seem to understand the use of the term 'learning' in the CKP screens. While the IntelLEO services are originally developed to support learning activities at workplaces, many participants thought that it sounded more like 'studying at school' and that using some other term(s) would make the user interface clearer. In fact, they agreed that the CKP functionality is actually more related to knowledge and resource management, although it can be understood as learning as well. Others commented that it was difficult to grasp immediately (i.e., after initial explanations and use of CKP) the actual difference between CKP and widely used social bookmarking services like Delicious (http://www.delicious.com/) and Diigo (http://www.diigo.com/).

Still, the overall impression from the participants' free-form comments was that their perception of the IntelLEO services, especially CKP, was very positive. Some of them even put in their comments that they would like to be advised on other similar Web services. Over 80% of their comments indicated that using CKP regularly would help them, in their view, do their job more efficiently and more effectively, and would also help their institutions a better insight into the needs of their collaborating organizations. An important suggestion emerging from their comments was that promoting the use of Web services in their institutions would be very beneficial.

7. Related and future work

A recent study has indicated that researchers in the field of demography are not completely aware of the benefits they can have from using Web services and other new Web technologies and tools [1]. Moreover, creating online communities of demographers is not that widespread yet, compared to the communities in other disciplines. More effort is needed in order to raise the awareness of demographers for new Web technologies and online communities, and especially for using them in everyday work in their organizations.

A possible path to follow in this direction is to learn how professionals from industry and business areas have started to use Web services and Web 2.0 technologies at workplaces, a movement known as Enterprise 2.0 [4]. Likewise, getting familiar with some of the existing Web services for demographers may be helpful. For example, DOTS Demographics (http://www.serviceobjects.com/products/address/demographics) provides ZIP code-level, small-segment demographics for in-depth local-area profiles for USA. It’s built with sources such as U.S. Census data, housing and urban development (HUD) data, current-year demographics and data culled from millions of consumer purchase records.

Of course, as in many other situations, motivational aspects are the key to starting and maintaining successful and collaboration and resource sharing. To this end, [5] argue that a continuous dialogue between interested parties at workplaces and organizations may contribute to creating a sense of shared ownership over resources and tools, and become a motivational driver for use of related knowledge sharing technologies.

As IntelLEO includes several other services in addition to CKP, experience is growing in terms of evaluating and using them. In a recent survey, [7] have found users to be willing to contribute to organizational learning and resource sharing provided that they is receive some recognition of and feedback about their contribution. The recognition should come from their peers and organizations, and should state organizational expectations explicitly.

CKP and the other IntelLEO services are developed as Web services, and it is necessary to understand the aforementioned ambiguity of the term "Web service". It comes from the fact that developers and end users view it differently. The way the term is used in this paper so far is more suitable for end users. However, technically, a *Web service* is a software system designed to support interoperable machine-to-machine interaction over a network [2]. To enable this interaction, a Web service has an interface described in a machine-processable format and can receive and send messages typically encoded using a specific protocol based on HTTP and XML.

More recently, the emphasis in the Web service technology is moving towards *RESTful Web services*, and as a part of future development of CKP its transformation to a RESTful Web service is planned. RESTful Web services use HTTP and the principles of *Representational State Transfer*, or *REST* [3]. These mean that a client application (e.g., a Web browser) initiates requests to Web servers, the servers process the requests and return appropriate responses, and both the requests and the responses are related to the transfer of representations (states) of resources. A resource is generally anything on the Web that can be addressed (by a Uniform Resource Identifier, or URI, such as http://example.com/resource/). Its representation is typically a document that captures its current or intended state (e.g., an HTML page) when the client application accesses its address. The returned representation of the resource places the client application in a new state; it enables the client to, e.g., traverse a hyperlink in the returned representation to access another resource. This, in turn, will typically return the current representation of the other resource, which will place the client application into yet another state. RESTful services are simple Web services; their implementation typically involves using just basic HTTP methods, such as GET, PUT, POST, and DELETE. Thus transforming CKP into a RESTful service is expected to increase its performance and usability.

A part of transformation of CKP to a RESTful service should include modifications in its user interface (or making it configurable) to match the mindsets of its users (demographers, in this case). The same holds for the other IntelLEO services. For example, as the evaluation has shown, replacing 'learning' with 'collaboration' in the user interface would make some users feel more comfortable with CKP.

Further work is also needed to enrich CKP with more demography-related ontologies/vocabularies (or, again, to make it easily configurable for using other ontologies). As discussed in [1], good-quality domain ontologies are not easy to develop. Thus reusing existing relevant ontologies (e.g., the GeoNames ontology http://www.geonames.org/ontology/documentation.html) with CKP can not only increase the service's usefulness, but also save a lot of work.

8. Conclusions

Web services for demographers are not plentiful, although some do exist. Demographers are often not aware of such services, or sometimes it looks as if it would take too much effort to start using them. Still, given the fact that much of the work done by demographers requires collaboration of people from more than one institution, there is an evident need for using services that can facilitate such collaboration. These services should pertain to resource sharing, common repositories, easier navigation, and increase of on-the-job effectiveness.

The recent trend of opening government databases to research, businesses, and general public – and many of these databases are related to population of the corresponding countries – new opportunities arise for researchers and other professionals in demography to do more creative and more useful work. In such a situation, the benefits they can have from using new Web services are related to a better insight into available resources, to learning from other colleagues and professionals about new resources, and to making the things that researchers and professionals want to accomplish easier to get done. To get started with using Web services, awareness of existing services is necessary. Using them is not a technical problem, and some of them are already out there on the Web. It also helps if professionals and researchers in demography and other social sciences get aware of many Web services that facilitate collaborative work, such as IntelLEO Web services, and that can be adapted for use by social scientists.

Acknowledgements. This publication was partially supported/co-funded by the European Community/European Union/EU under the Information and Communication Technologies (ICT) theme of the 7th Framework Program for R&D (FP7). This document does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of its content. Research presented in this paper is also partially conducted within the national research projects "Infrastructure for Technology Enhanced Learning in Serbia" (Project No. 47003), and "Studying Demographic Phenomena as a Support to Public Policies in Serbia" (Project No. 47006) funded by the Ministry of Education and Science of the Republic of Serbia.

References

1. M. Devedžić, V. Devedžić. (2010). "Imagine: Using New Web Technologies in Demography". *Social Science Computer Review*, Vol.28, No.2, pp. 206-231.
2. H. Haas, A. Brown (Eds.). (2004). "Web Services Glossary." http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/ (last visited: May 22, 2011)
3. R. Fielding. (2000). Architectural Styles and the Design of Network-based Software Architectures. PhD Dissertation. University of California, Irvine. http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm (last visited: May 22, 2011)

4. D. Hinchcliffe. (2007). "The state of Enterprise 2.0." [Online]. Available: http://blogs.zdnet.com/Hinchcliffe/?p=143
5. T. Holocher, B. Kieslinger, C.M. Fabian. (2011). "Applying Participatory Methods to Address Motivational Aspects in Informal Workplace." *International Journal of Advanced Corporate Learning* (*iJAC*). Vol 4, No 1, pp. 18-24.
6. Research Information Network. (2010). "If you build it, will they come? How researchers perceive and use Web 2.0." www.rin.ac.uk/system/files/attachments/web_2.0_screen.pdf (last visited: May 22, 2011)
7. M. Siadaty, J. Jovanovic, N. Milikic, Z. Jeremic, D. Gasevic, A. Giljanovic. (2011). "Lear-B: A Semantics-enabled Collaborative Environment for Self-regulated Workplace Learning." *Submitted to ISWC 2011, The 10th International Semantic Web Conference*, Koblenz, Germany, Oct. 23-27, 2011.
8. T. Tullis, B. Albert. (2008). Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. p.317. San Francisco: Morgan Kaufmann Publishers.
9. K. Watkins et al. (2005). "UN Human Development Report 2005". United Nations Development Programme. http://web.archive.org/web/20080527203423/http://hdr.undp.org/en/media/hdr05_complete.pdf (last visited: June 20, 2011)
10. Wiktionary.org. (2011). "Web Service." http://en.wiktionary.org/wiki/web_service (last visited: May 22, 2011)

Received: December 14, 2011; Accepted: February 13, 2012.

DOI: 10.2298/CSIS111128019X

Exploring the Use of Contextual Modules for Understanding and Supporting Collaborative Learning Activities: An Empirical Study

Lu Xiao

Faculty of Information & Media Studies
University of Western Ontario
London, Ontario, Canada
lxiao24@uwo.ca

Abstract. We report three student groups' collaboration experiences in a semester-long classroom project. The project included both tasks that required completion in virtual group workspace and activities that could be carried out in the physical world environment. We observed different collaboration patterns among the groups with respect to building and maintaining social relationships, submitting individual work to the group, and scheduling group meetings. We use Bereiter's two contextual modules, intentional learning and schoolwork, to help us understand the observed patterns and suggest that the group leader's contextual module plays a significant role in all members' group learning experiences and outcomes. We propose design implications that are intended for encouraging learning-based (as opposed to work-based) practices in virtual group environments.

Keywords: computer-supported collaborative learning, classroom study, virtual group learning environment

1. Introduction

Research on collaborative learning has shown that groups working together for a common task can outperform individuals working alone by producing higher achievement and greater productivity. In the collaborative learning model, students work together in small groups to achieve a common academic goal, such as a semester project or a homework assignment. This is fundamentally different from the traditional "direct-transfer" or "one-way knowledge transmission" model in which the instructor is the only source of knowledge or skills. In collaborative learning activities, students are viewed as active participants in the learning process in which they interact with peers and experts, which has the potential to produce greater learning than a student learning on their own.

However, research studies have also shown that the effectiveness of collaborative learning is dependent on many conditions such as the group composition (size, gender, heterogeneity, etc), the group tasks, the group members' background and motivation with respect to the learning subject, and the means of communication, among other variables. In this paper, we report three student groups' collaboration and learning experiences in a semester-long classroom project. The project included both tasks that required completion in a virtual group environment and activities that could be carried out in a physical world environment. We compared and contrasted the group processes in these two environments. Our initial assumption was that groups would have quite different collaboration styles and practices due to the differences of these environments. Contrary to our assumption, we observed that the groups had similar patterns in both virtual and real group environments with respect to building and maintaining social relationships among the group members, submitting an individual's work to the group, and scheduling group meetings.

To help explain the observed phenomenon, we used Bereiter's contextual modules, i.e., the intentional learning module and schoolwork module, which in the past were mainly used in explaining the influence of contextual variables on an individual's learning experiences in a physical world environment. The results suggest that the group leader's contextual module plays a significant role in all members' group learning experiences and outcomes. We propose design implications that are intended for encouraging learning-based (as opposed to work-based) practices in the virtual group environment.

The rest of the paper is organized as follows: first, we describe the research design of the classroom study, including the group tasks in two environments, the tool that provided the virtual environment, and the data collection and analysis techniques for understanding the group processes; next, we present the findings of similar patterns in detail; then, we discuss Bereiter's contextual modules and demonstrate how we used these modules to help analyze the findings; and last, we discuss the design implications of a collaborative tool for encouraging learning-based practices in virtual group environment.

2. Research Design

The classroom study was conducted in a junior-level undergraduate course on project management at a major US university. This is a core course for the major of Information Sciences & Technology (IST). Student learning progression was largely evaluated by the group performance in a semester group project – a project that required students to research best practices for distributed teamwork at each of its five phases. The project included five major activities corresponding to the five phases that a distributed team goes through. During each major activity, the groups identified the top three challenges or risks that a distributed team faces at that particular project phase; compared and contrasted two to three technology tools to be used by a distributed team; and identified at least three best practice recommendations for the use of technology tools by a distributed team. At the end of each major activity, the group submitted a mini-report. After completion of the five major activities, the group produced a final report.

2.1. A Virtual Group Learning Activity

We designed a virtual group activity to offer students experiences of distributed teamwork. In the activity, each student group brainstormed challenges that a distributed team may face during a specific phase, and identified the top and bottom three challenges that were most and least important to address for the distributed teamwork. Students were required to provide a rationale justifying their choices of top and bottom challenges. The proposed challenges, members' choices, and the rationales were all archived and shared within the group.

In each major activity, the students completed this virtual group activity in two days and then conducted the remaining tasks. The students could choose the communication and collaboration means for the remaining tasks. At the end of a major activity, they submitted a mini-report that summarized their work. The students had one and a half weeks to complete the remaining tasks. Figure 1 gives an overview of the group project.

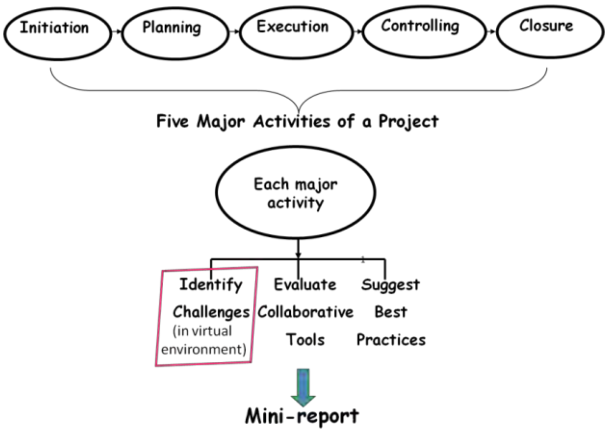
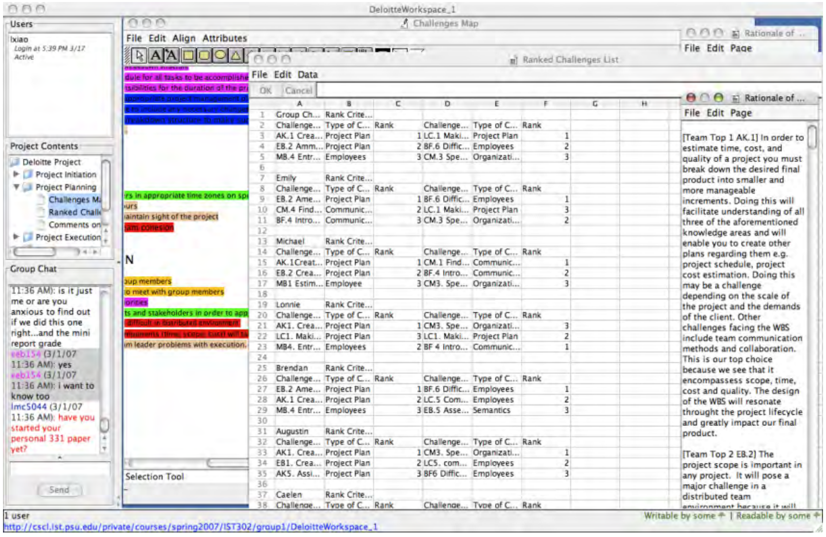

Initiation
Planning
Execution
Controlling
Closure
Five Major Activities of a Project
Each major activity
Identify Challenges (in virtual environment)
Evaluate Collaborative Tools
Suggest Best Practices
Mini-report

Figure 1. Overview of the group project

2.2. A Virtual Group Learning Environment

The student groups carried out the virtual group learning activity in a virtual group workspace provided by the collaborative tool. The workspace had several subspaces: a user list space indicating which students were present in the workspace, a group chat space, a document list space presenting a tree view of the shared group documents, a document space for currently opened documents, and a rationale space for the documents' associated rationales. Given the virtual workspace and its available toolkit, the students posted challenges in a shared whiteboard and their choices of top and bottom three challenges in a shared spreadsheet. The rationale space was essentially a blank space with no structure . It would display the rationales when related whiteboard or spreadsheet documents were opened in the document space. Figure 2 shows a screenshot of one student group's workspace provided by the collaborative tool. This group used a color scheme to distinguish who had posted which challenge on the shared whiteboard. The details about the tool design and architecture can be found in Xiao, TEL, 2011.



lxiao
Login at 5:39 PM 3/17
Active

Deloitte Project
Project Initiation
Project Planning
Challenges M...
Ranked Chall...
Comments on...
Project Execution

11:36 AM): is it just me or are you anxious to find out if we did this one right...and the mini report grade
eeb154 (3/1/07 11:36 AM): yes
eeb154 (3/1/07 11:36 AM): i want to know too
lmc5044 (3/1/07 11:36 AM): have you started your personal 331 paper yet?

dule for all tasks to be accomplishe
sibilities for the duration of the pr
rs in appropriate time zones on sp
urs
aintain sight of the project
am cohesion
N
oup members
o meet with group members
orities
ts and stakeholders in order to app
difficult in distributed environment
m leader problems with execution.

	A	B	C	D	E	F
1	Group Ch...	Rank Crite...				
2	Challenge...	Type of C...	Rank	Challenge...	Type of C...	Rank
3	AK.1 Crea...	Project Plan	1	LC.1 Maki...	Project Plan	1
4	EB.2 Amm...	Project Plan	2	BF.6 Diffic...	Employees	2
5	MB.4 Entr...	Employees	3	CM.3 Spe...	Organizati...	3
6						
7	Emily	Rank Crite...				
8	Challenge...	Type of C...	Rank	Challenge...	Type of C...	Rank
9	EB.2 Ame...	Project Plan	1	BF.6 Diffic...	Employees	1
10	CM.4 Find...	Communic...	2	LC.1 Maki...	Project Plan	3
11	BF.4 Intro...	Communic...	3	CM.3 Spe...	Organizati...	2
12						
13	Michael	Rank Crite...				
14	Challenge...	Type of C...	Rank	Challenge...	Type of C...	Rank
15	AK.1Creat...	Project Plan	1	CM.1 Find...	Communic...	1
16	EB.2 Crea...	Project Plan	2	BF.4 Intro...	Communic...	2
17	MB1 Estim...	Employee	3	CM3. Spe...	Organizati...	3
18						
19	Lonnie	Rank Crite...				
20	Challenge...	Type of C...	Rank	Challenge...	Type of C...	Rank
21	AK1. Crea...	Project Plan	1	CM3. Spe...	Organizati...	3
22	LC1. Maki...	Project Plan	3	LC1. Maki...	Project Plan	2
23	MB4. Entr...	Employees	2	BF 4 Intro...	Communic...	1
24						
25	Brendan	Rank Crite...				
26	Challenge...	Type of C...	Rank	Challenge...	Type of C...	Rank
27	EB.2 Ame...	Project Plan	1	BF.6 Diffic...	Employees	1
28	AK.1 Crea...	Project Plan	2	LC.5 Com...	Employees	2
29	MB.4 Entr...	Employees	3	EB.5 Asse...	Semantics	3
30						
31	Augustin	Rank Crite...				
32	Challenge...	Type of C...	Rank	Challenge...	Type of C...	Rank
33	AK1. Crea...	Project Plan	1	CM3. Spe...	Organizati...	1
34	EB1. Crea...	Project Plan	2	LC5. com...	Employees	2
35	AK5. Assi...	Project Plan	3	BF6 Diffic...	Employees	3
36						
37	Caelen	Rank Crite...				
38	Challenge...	Type of C...	Rank	Challenge...	Type of C...	Rank

[Team Top 1 AK.1] In order to estimate time, cost, and quality of a project you must break down the desired final product into smaller and more manageable increments. Doing this will facilitate understanding of all three of the aforementioned knowledge areas and will enable you to create other plans regarding them e.g. project schedule, project cost estimation. Doing this may be a challenge depending on the scale of the project and the demands of the client. Other challenges facing the WBS include team communication methods and collaboration. This is our top choice because we see that it encompassess scope, time, cost and quality. The design of the WBS will resonate throught the project lifecycle and greatly impact our final product.

[Team Top 2 EB.2] The project scope is important in any project. It will pose a major challenge in a distributed team environment because it will

http://cscl.ist.psu.edu/private/courses/spring2007/IST302/group1/DeloitteWorkspace_1

Figure 2. Screenshot of a student group's workspace

2.3. Data Collection and Analysis Techniques

Following a triangulation approach, we collected and analyzed various kinds of data to understand the students' collaboration and learning experiences in the group project, including surveys, class observation notes, semi-structured interviews, and the grades of the mini-reports. Students who chose to participate in the study were offered extra credits. To address possible ethics concerns, a different task was made available for those who chose not to participate but wanted to earn extra credits. In total, 33 of 38 students chose to participate in the study. These 33 students were from the seven groups formed by the instructor. However, there were only five groups whose members all agreed to participate in the study.

2.4. Quantitative Data: Survey of Likert Scale

30 participants filled out a survey about their experiences at the end of the project. This survey included three sections: collaboration experiences, feedback on the design of the shared rationale space, and individual learning experiences. This paper focuses on the students’ experiences of learning in the groups, so the discussion about survey results will focus on the survey items about collaboration experiences. The design of these survey items were based on existing scales about group work (see Table 1): item 1 was about shared group identity and was from the scale used in Roberson’s study [13], and item 4-12 were customized from the existing scale used in the study of partially distributed teams [12]. Item 2 and 3 were created to examine the participant’s satisfaction on the teamwork.

Table 1. Survey items about collaboration experiences

Value ranging from 1 as “strongly disagree” to 7 as “strongly agree”)
1. I see myself as a member of my group
2. I felt satisfied about our group work in challenges assessment activities
3. I felt satisfied about our group work in mini-report activities
4. My group members were competent in terms of generating a diverse set of explanations
5. My group members were competent in terms of generating good quality of explanations
6. My group members were quite competent in terms of generating good quality of mini-reports
7. Much disagreement on performing the tasks existed in challenges assessment activities
8. Much disagreement on performing the tasks existed in mini-report activities
9. A great deal of disagreement regarding the tasks existed in challenges assessment activities
10. A great deal of disagreement regarding the tasks existed in mini-report activities
11. Little tension existed in my group
12. When I needed help I counted on my group

2.5. Qualitative Data: Semi-structured interview and Observation

Three of the five participating groups were interviewed about their collaboration experiences. These groups were chosen based on the grades of first mini-reports. The selected groups were group 1 (highest grade), group 2 (lowest grade), and group 3 (middle grade). A semi-structured interview guide was used. Specific questions were asked to elicit the members' responses about the group history, social relationships, group dynamics and their collaboration experiences. The interview questions also solicited the interviewees' opinions on the design of the shared rationale space and their history of working in and leading a group project. Following the procedure recommended by Mason [9], the "big" question was decomposed into "small" questions. Because the purpose of these interviews was to gather rich data understanding the group processes and members' feedbacks on the activities, including the software program that supported the virtual group activity, the interviewer was not asked to follow a strict prompt protocol; instead, she was allowed to ask the interviewee questions to clarify or probe details of something the interviewee had said. There were 32 interview questions in total. Listed below are examples:

1. Have you worked with anyone from the group before? Who are they?
2. Do you know if some of your group members have worked with each other before?
3. How do you think of your group, compared to other groups that you have worked with? Examples of the aspects include cooperation, communication, meetings, member involvement, and conflict management?
4. Is there any critical incident or event happened in your group that you want to talk about?
5. Did each member contribute to the project equally? (Follow up: How? Why not?)
6. Which one of your group members do you think contributed to the group project most in this project? Why?
7. Is there any conflict in the group? Working style conflict? Personality conflict? Etc. (follow up: How did the group manage the conflict? Do you think sharing the explanations to each other helped on managing the conflict that is related to the task?)
8. Are you willing to work with your group again?

All interviews were conducted in a face-to-face setting after completion of the third major activity. They lasted from 20 to 90 minutes and were recorded and transcribed.

Besides interviewing the group members, the researcher attended each class lecture, observing the members' interactions. The researcher also attended group meetings of these three groups whenever her schedule allowed.

The interviews and observation notes were coded through an open coding process using ATLAS.ti software. In coding the first document, the researcher created 76 codes. The researcher added two more codes in coding the second document. The researcher then generated the coding scheme which had four coding families and 64 codes. In this process some codes were merged, e.g., the three codes about the interviewee's attitude on the software design – "*attitude on software*", "*dislike about the software*", and "*like about* *the software*" were merged into one code – "*attitude about the software*". The researcher then presented her coding schema and collected interview data to a professor in the Department of Education of a major US university, which served as a peer debriefing process to check on the data analysis process and increase its trustworthiness. The professor offered additional advice on coding process and agreed on the generated coding scheme. The researcher then coded the rest of the documents with the coding scheme. After coding all the documents, the code list was reviewed and the corresponding quotations were reexamined. The final coding scheme had 54 codes, and 712 quotations.

3. Findings

In this section, we first discuss three groups' performances in terms of their group reports' grades and rationale grades. We then present our observed similar patterns of the groups' processes in the virtual group environment and the physical world environment.

3.1. The Group Performance

The group performance is reflected through the group's average grade of four mini-reports and the average grade of four challenge assessment tasks. The instructor graded the groups' mini-reports based on the reports' quality. The researcher graded rationales based on the level of thinking the rationales revealed, that is, a rationale that was well articulated and deeply thought out would be given a higher grade than one that showed less intellectual effort. For example, the challenge of "*keeping track of the group members' work in distributed teamwork*" was proposed by two members of different groups for the project monitoring phase. One's rationale was graded 5/5.5 - "*Throughout the project lifecycle, the project team must be carefully managed by the diligent watchful eye of the project monitoring and controlling phase. The monitoring and controlling phase must ensure that the project team is making progress according to the project plan. This will ensure project quality with regard to scope, budget and time and result in successful projects. However in the distributed team environment this is increasingly difficult due to physical barriers. Also (it's) very difficult due to the grandeur of the objective of this challenge. It is very difficult to manage the entire team effectively enough in a non-distributed team environment. All the more difficult in this environment*", whereas the other's rationale, "*logging the work each team member does is extremely important. It allows the project manager to track accountability for each group member in case of mistakes/error*" was graded 3.8/5.5.

As there were five participating groups and each group had four mini-reports, there were 20 mini-reports' grades in total considered in the study. The average grade was 9.0, with the highest grade being 10 and the lowest grade being 6. The average grade of the mini-report for the highest performance group was 9.6 (i.e., averaging its four mini-reports' grades) and for the lowest performance group was 7.7.

The average grade for rationale statements was 4.6, with a range of 1 to 5.5. The quality of the rationale statements was in general good with 41% of the rationale statements' grades being 5 or above, and 17% of the rationale statements' grades being 3 or below. The grade of the highest performance group on rationale quality was 5.3, and that of the lowest performance group was 3.5. Table 2 shows the five participating groups' grades of mini-reports and rationales. As it shows, group 1, 2, and 3 were comparable in terms of their performance, with group 1 being slightly behind.

Table 2. Group Performance

Group	Average grade of mini-reports	Average grade of rationales	Group Performance (sum of the two grades)
1	9.3	4.4	13.7
2	9.7	4.3	14.0
3	9.6	4.7	14.3
4	9.0	4.3	13.3

3.2. Group Work Experiences: Survey Results

30 participants filled out the survey with 5 of them from non-participating groups, i.e., groups that had non-participating members. These 5 participants' responses were removed from further analysis.

Next, a group's response to a survey item was calculated as the average of its members' responses on the item. As the group size was five or six, it would be inappropriate to perform statistical tests to examine the significance of the average value. Instead these results were used for a coarse comparison of the members' experiences between the three groups. As a seven-point Likert scale was used, the biggest difference between two responses on an item would be 6. The difference of at least 1 full point on the Likert scale was considered the cutoff value for the difference between the response values. The analysis showed that being in different groups made survey responses significantly different except in items 7 and 9-11. The author also used ANOVA to examine whether being in a group had an impact on the responses. In other words, "group" was used as the factor in the analysis and there were three levels of that factor, i.e., three groups.

Table 2 presents the results. Item 1's responses show that all three groups were indeed formed. However, group 2 and 3's higher response values seemed to indicate that the groups were more cohesive than group 1. Compared to group 1, group 2 and 3 had higher value responses for item 2 – 6. These imply that group 2 and 3 were more satisfied than group 1 about the group performance. The members of these two groups were more confident about the quality of their rationale statements than group 1 members.

Table 2 Survey results on group work experience for group 1, 2, and 3

Survey Item	Group’s Average Response			P value of ANOVA Group as the factor (N = 25)
value ranging from 1 as “strongly disagree” to 7 as “strongly agree”)	Group 1 (N = 6)	Group 2 (N = 6)	Group 3 (N = 5)	
1. I see myself as a member of my group	5.5	6.8	6.8	0.005
2. I felt satisfied about our group work in challenges assessment activities	4.3	5.8	6	0.009
3. I felt satisfied about our group work in mini-report activities	4.7	6.4	5.3	0.002
4. My group members were competent in terms of generating a diverse set of explanations	4.5	6	6.2	0.004
5. My group members were competent in terms of generating good quality of explanations	4.7	6.2	5.7	0.004
6. My group members were quite competent in terms of generating good quality of mini-reports	5	6.6	6.3	< 0.001
7. Much disagreement on performing the tasks existed in challenges assessment activities	2.3	2.6	2	0.291 (insignificant)
8. Much disagreement on performing the tasks existed in mini-report activities	2.3	1.8	1.8	0.019
9. A great deal of disagreement regarding the tasks existed in challenges assessment activities	2.2	2.2	1.7	0.240 (insignificant)
10. A great deal of disagreement regarding the tasks existed in mini-report activities	2.8	2.2	1.7	0.109 (insignificant)
11. Little tension existed in my group	4.3	6.2	5.5	0.148 (insignificant)
12. When I needed help I counted on my group	4.7	6.6	6.7	0.001

All three groups had little disagreement during the project regarding the tasks (items 7 – 10), agreed that there was little tension among the group members (item 11), and agreed that one could count on his/her group members when in need of help (item 12). However, groups 2 and 3 had noticeably higher response values than group 1 on items 11 and 12, which suggests that group 2 and 3 perceived a better group relationship and built more trust among the members than group 1 members.

The interview and observation data supported these findings from the survey results. Group 2 and 3 had better social relationships than group 1, and members were more satisfied about each other’s contribution as well.

3.3. The Similar Patterns of Group Processes in Two Learning Environments

The comparison of the group processes between the two group environments shows that each group had its way of dealing with *building and maintaining social relationships among the group members*, *submitting individual's work to the group*, and *scheduling group meetings* independent of the environment.

Building and maintaining social relationships

Group 1

The effort of building group relationships was not observed in group 1 in either environment. Group 1 members Bill and Thomas were friends. They sat next to each other in lectures and turned in their report sections together. They seldom talked to other members during the class except asking the group leader occasionally about the assignments and what they were supposed to do. The remaining four members did not know each other prior to this class. Justin and Jack were often seen standing outside the classroom talking to each other either before or after the class. Tom talked to Justin and Jack during the classes sometimes. Bob stayed quiet most of the time and sat furthest away from the other members.

In the virtual group environment, group 1 used the chat tool when they were first in the group workspace to test if it worked. However, they used it little during the three-month period, compared to the other two groups. The group generated 67 messages in total, most of them about the workspace tool and the task.

Group 2

Members of group 2 had very good relationships in this project, partially because this group had a good history of working together. Kim, Martin, and George had worked together since their freshman year. Rex had never worked with them before but had been friends with them since his freshman year. Kim and Martin were high school friends and came from the same town. Jordan knew George through a mutual friend.

Because of schedule conflicts, the only time that the group could meet was in the evening; however, this did not seem to bother the members. They really enjoyed the meetings. During the meetings, they worked on the project as well as playing music and joking with each other. Joking was also a strategy the group applied when there was a conflict or tension among the members; as Martin said, "*We don't let the tension build up. We joke around*". Because they spent much time on other things besides work, the group often met for several hours in order to get things done. The members acknowledged this issue but they still enjoyed this collaboration style. For example, Martin said, "*one thing is that everyone always looks at a sense of* *humor during the group meetings. We all liked joked around a lot, which definitely helped a lot. It actually made the meetings fun*".

This was the first time that Jordan worked with the other four members. During the interview, group 2 leader George mentioned that Jordan was a very strict person and followed methodology precisely. In spite of the fact that this group's work style is far from being strict, Jordan also enjoyed working in this group – "*I think this has been one of the most interesting groups I have ever worked with, in terms of really just being loose all the time and really being able to just kind of throw out all ideas and we joke around a lot. Unfortunately that also leads to getting off task sometimes, but I feel that we are all very intelligent group members and the quality of our work is very good*".

In the virtual environment, members also found ways to maintain and grow the relationships. Group 2 generated 681 messages in the group chat during the project period. It was evident that the group members used this chat for activities other than work. There were many naive behaviors, such as posting messages in Japanese, using special characters in messages, and discussing the researcher's appearance. Below is one example to illustrate this:

1. sdg5015 (3/21/07 9:58 PM): dvb eats glass
2. sdg5015 (3/21/07 9:58 PM): if that's a turn on
3. ajk5054 (3/21/07 9:59 PM): harmoniously
4. ajk5054 (3/21/07 9:59 PM): there
5. sdg5015 (3/21/07 9:59 PM): pandas eat glass?
6. ajk5054 (3/21/07 9:59 PM): yes
7. sdg5015 (3/21/07 9:59 PM): is that why they're endangered?
8. ajk5054 (3/21/07 9:59 PM): thats why they are extinct
9. sdg5015 (3/21/07 9:59 PM): and sad?

Group 3

Group 3's members sat in the same row in the class. Laura and Jimmy had worked together previously; Stephen and Jonah had worked in a quality team before. The rest of the members did not know each other before taking this class together. During the semester, Stephen and Jeffrey were in the same group in another class, as were Laura and Jeffrey. Overall, everyone got along with each other in the group, and members were all positive about the group and developed trust for each other in completing the shared tasks – "*In the first two meetings, we would only schedule a meeting when everyone is there so we know that everyone will be there. Now if one or two people cannot make a meeting but other people can. We still schedule the meeting. We trust them that they can get the work done even though they are not going to be at the meeting*".

Group 3 generated 1512 messages during the project period. Similar to group 1, most of these messages were about the tasks. But there were occasions that members joked with each other:

1. ctm5017 (2/28/07 8:53 PM): ok what do we want to make the groups
2. lmc5044 (2/28/07 8:53 PM): ill give you a cookie
3. lmc5044 (2/28/07 8:53 PM): :)

Submitting individual’s work to the group

Group 1

Group 1 had issues with members submitting their work. Although Bill and Thomas always sent their individual sections on time, this was not well-recognized by the other group members. For example, Bob blamed Bill and Thomas for causing the grades of the mini-reports to be lower than expected, “*We split it up the parts evenly. And these two people complained about doing their parts even though they won't that big. They turned them in late a couple of times, or hard to get a hold of, or didn't show up in the class. So we end up having to do stuff in the last minute… And we get dark points because of them*”.

For completing the tasks in the virtual environment, Bill was considered to have contributed the least: he proposed 21 challenges in total while the other members proposed 30, and he only participated in the first task of selecting and ranking the challenges. Although the group leader made the group selection on the top and bottom three challenges in all five major activities, the other members were not aware of his work. For example, when asked who made the group selections since there was no group meeting about it, Thomas commented, “*I think someone went through and looked, whether it was Jack or Justin. I was never asked for my opinion except I ranked them in the assessment as individual. That was the only input*”.

In summary, group 1 had several problematic issues compared to groups 2 and 3: the lack of group history and social cohesiveness; disagreement on the group members’ contribution; and the lack of visibility of individual effort in the project. All of these could have contributed to the fact that group 1 had lower scores on survey items related to social relationship and trust in the group.

Scheduling group meetings

Group 1

Group 1 did not establish a regular meeting schedule for the project. During the three-month period, the group met only four times, twice at the beginning for one group assignment and for dividing up the group report sections, and the other two for final presentations.

Group 1 did not meet online for the virtual learning activities either. The group members did not participate in the task of selecting the group's choice of top and bottom three challenges. Instead, the group leader looked at each member's choices and made the selection. Although the group leader did so in all five major activities, the other members were not aware of his work. For example, when asked who made the group selections since there was no group meeting about it, Thomas commented, "*I think someone went through and looked, whether it was Jack or Justin. I was never asked for my opinion except I ranked them in the assessment as individual. That was the only input*".

Group 2

The group 2 members loved having meetings. This group met weekly during the project despite the fact that the only time that they could meet was in the evening. They often joked around during the meetings and the atmosphere was more like a group of friends getting together for fun than working on a business matter. For example, in a meeting starting at 7 p.m. on Wednesday, the members did not finish the meeting until it was close to mid-night.

Group 3

In general, for the tasks in virtual group workspace, the group met online in the workspace to brainstorm the challenges together and then again at another time in the AIM group chat room. The group also had at least one face-to-face meeting for each mini-report.

4. Intentional Learning vs. Schoolwork Module

Bereiter's concepts of intentional learning and schoolwork modules were used to further compare the three groups. According to Bereiter's work, students whose academic activities are all mediated by the schoolwork module do not recognize the learning goals of the activities. On the other hand, students who adopted the intentional learning module in carrying out the activities take the goal of learning into consideration when working on the activities. Also, students of the schoolwork module organize their activities for task performance, whereas those of the intentional learning module organize their activities around goals of personal knowledge construction. The third dimension to distinguish the two modules is the morality of the students. Students who adopt the intentional learning module commit to a truth and depth of understanding, while those who adopt schoolwork module tend to be satisfied with the superficial and pretense of knowledge [14].

We adapted his work and considered the dimensions of the contextual module at the group level: practice of division of labor, leader's contextual module in collaborative learning activities, and moral dimension of the group activities.

4.1. Practice of Division of Labor

According to Bereiter's work, students whose academic activities are all mediated by the schoolwork module do not recognize the learning goals of the activities. On the other hand, students who adopt the intentional learning module in carrying out the activities take the goal of learning into consideration when working on the activities. Additionally, the intentional learning module is organized around goals of personal knowledge construction while the schoolwork module is organized for task performance. In the group learning context, this dimension is about whether the learning goal is taken into account in the group practice. The data from this study show that this is specifically related to the group practice of division of labor. In the group project, a large amount of time in group work is on completing tasks and writing the report. The three groups have very different strategies for division of labor in producing mini-reports as follows:

Group 1

In group 1, each group member picked a section from the report template and produced the content for the section. Group leader Justin chose to be the compiler responsible for integrating individual sections into one coherent piece. He required the group members to send him individual sections two days before the actual due date for compiling. The group members were to stick to the same section and responsibility for all five mini-reports. There were two members who worked on the section discussing two software programs. These two members decided that one would write about a software program first and send it to the other member. Then the other member would write about the second program and send the completed section to the group leader. The group leader did not send out a compiled version to the group prior to the submission, so the other group members did not get to read the report until they received it from the instructor along with the grade.

Group 2

Group 2 wrote the mini-reports during the face-to-face group meetings. For a mini-report, the group members would first write down the ideas on the whiteboard of the meeting room, and discuss which ideas or topics to be included in the report and why. Members then worked in small groups on different sections. When one small group finished the work, they would then switch to a different section or help the others.

Group 3

Group 3 members used a rotation strategy to divide up the work for writing mini-reports: the members rotated to different sections for the five mini-reports, enabling each member to construct different sections over the course of the project. The group created an online group folder for sharing the sections and the integrated mini-reports. Usually, the members uploaded the individual sections in the folder the day before the actual due date. The compiler then integrated them into a coherent report and uploaded the integrated version to the group folder a couple of hours before the submission for the group to review. The compiler was responsible for submitting the final version to the company.

These findings suggest that a group's strategy of division of labor can be very different depending on the focus: task performance oriented or knowledge construction oriented. Group 1's strategy was task performance oriented, which is reflected in the group leader's reasoning of having everyone work on the same section in all five activities. When asked "*How did your group decide on the mini-report content*", He said, "*pretty much we just distributed the work and it comes back, and we just put it together. What helps me get consistent is everybody wrote the same part every time. And like in section D where two people are writing in the same section, one talking about one software, one talking about the other, one writes before the other, so it goes from XX to XX, then to me. So at least there is some consistency there at least worked out the same ideas*". In group 2 and 3, however, there was certainly the intention of having everyone involved in different aspects of the project and be responsible for different sections at different stages. For example, group 3 members all liked the rotation strategy and acknowledged that it helped them to explore different parts of the project. Stephen talked about the tradeoff of the rotation strategy – "*I don't argue with the system we have about rotating it, cause it gets everyone to do one part of it, but if we want the best output, I'll probably write my challenges and I'll go back to read the person's tool assessment and best practices for groove, because I am probably the most familiar with groove right now*".

As it shows, Stephen acknowledged that the purpose of this project was not just about getting the best performance, but also about gaining the experience by working on different sections.

Although group 2 did not have a systematic rotation strategy, the group made sure that the mini-report was constructed together in group meetings and members were not assigned to the same sections all the time. In such a highly collaborative group, there was no clear rule as to who worked on which section, and group members knew how each section was constructed and why. Members liked this working style. The leader said, "*I think cohesively we were better because we don't really like most of the groups (in which) everyone just does individual contribution and pieces together, although we sort of do that, it's more of like in a distributed environment, like the work is delegated consistently, and like we form teams and we switch teams, … it's always usually a group of two, or a group of three when we work, it's just* *constantly moving around, communication with each other when we meet face to face and online*".

These empirical data suggest that one dimension to distinguish an intentional learning group from a schoolwork module group is its practice of division of labor. An intentional learning group acknowledges the individuals' learning goals from the group activity and involves all members in the learning tasks that the project entails. Conversely, the schoolwork module group puts the emphasis on group performance, neglecting the members' learning goals, and develops the group practice that is believed to be best or most convenient to deliver the group product.

4.2. Leader's Contextual Module in Collaborative Learning Activities

The results of this classroom study suggest that group leader's contextual module with respect to the collaborative learning situation has great impact on the group's contextual module. The interview and observation data indicated that in this study group 1's leader adopted the schoolwork module, while group 2 and 3's leaders adopted the intentional learning module. This can be illustrated through the following aspects:

The Group Leader's Recognition of the Project's Learning Goal

The group leaders had different understandings about the group project, which showed their different motivations in completing the project. When asked what the project was about, group 1 leader said that the project was about how to improve project management in the distributed team environment, as it was described in the project description. He did not seem to recognize his individual learning goals from the class either. For example, he did not spend time on the content of the mini-reports when compiling the sections together because "…*considering the grades came back very consistent and decent, there were never an area that I need to consume myself on that*". During the interview, he talked about another class in which the students had several tests and during each test, the students needed to write down their solutions to problems and explain their reasoning. Justin's comments indicate that he did not recognize the learning goal of the problem-solving tasks and merely focused on getting the job done by solving the problem, "*Then, the next question is why? I just don't find that a useful question at that point, cause it's an open ended question at the first place – how would you approach this problem? And then the only logical true explanation to why to me is because that's the knowledge I have to approach the problem. .... I am sure there might be some genetic behind it, but how they are brought up, and how I approach a problem is the same thing. Don't ask me why I solve the problem like that*".

As the group leader, Justin had great impact on group practices and division of labor. For example, he did not encourage group discussions as he saw no need to do so with respect to the group's grades. He did not send out the compiled report to the rest of the group for review. In summary, he did not see the learning benefits to managing the group work in this way. Justin affected the other group members' learning experiences by imposing this module on the group. For example, group member Thomas said, "*The only thing that I learn is when we actually see each other's work and why. That helps to better flush out the phase. The report is no teamwork*". Jack also said, "*All IST group projects are supposed to gain experience in working with teams. We are not gaining that type of experiences*".

George, the group 2 leader, however, acknowledged that learning was also an important goal of the project. He said, *"…the outcome wasn't as important as the process. We are working together on a team and understanding the concept of project management, making sure your deliverables are on time, dealing with the client. It is more a way of preparing the students for professional world. I think this is more of educational usage than an actual corporate usage*".

Although group 3 leader Laura did not specifically mention the learning goals, her answer showed that she considered the importance of offering every member chances of working on different aspects of the project. She said, "… *so everyone will have a chance to do the little bit of work... they will just be like one person will get to do one of each task*."

The three group leaders' understanding of the group project suggests that group 1 leader's perspective was task-oriented and focused on the "work" aspect of the activity, whereas the group 2 and 3 leaders considered the learning aspect of the project and recognized the necessity of offering members learning experiences about the topic.

The Group Leader's Perspective on Group Meetings

Group meetings are crucial to group members' learning experiences from group activities. Effective meetings not only facilitate communication and encourage productivity of the group, but also provide channels to help the students learn from each other and construct knowledge together. However, the group leaders had different attitudes on the importance of group meetings. The group 1 leader Justin believed that it was not crucial for the group to have a meeting because the instructor would not know whether the group had had a meeting or not, and scheduling a meeting was hard to do. He said, *"… we are just pretty much anti-group meetings. The 2 or 3% of doing better on the report is not worth of trying to find an hour time for the group meeting."*

He further elaborated there was no need to get each other's opinion and knowledge when working on the individual sections and that all the individual work could be done by individuals – "*we don't need the support or knowledge from everybody else, we can complete the work and get it done individually and turn it in and not have to struggle for group meeting*."

The group 3 leader valued the importance of group meetings in group projects and organized regular meetings both face-to-face and online. Laura explained the reason of having two kinds of regular meetings in the group, "*…Online is good for summarizing things for just making sure that people are task and understand what they have to do, and maybe clear up a few questions. But general brainstorming and talking about what we want to do and how we want to structure the project that kind of stuff that's better to do in-person. By doing both being in-person and online, it kind of makes things more convenient and allows more meeting times*".

George, the group 2 leader, also considered it very important to have group meetings. George acknowledged that the members liked being together, the cohesiveness of the group made the meeting pleasant, and the members were comfortable brainstorming ideas – "*I think we are more creative face-to-face*".

4.3. Moral dimension in group activities

The results suggest that the moral dimension in distinguishing the schoolwork and intentional learning modules exists at the group level of the modules as well. For example, in selecting the top and bottom three challenges as group choices, all three groups seemed to adopt the same practice by examining individuals' choices and selecting the challenges with the most occurrences as the group choices. However, the process of doing this and generating the group rationales differed. In group 1, leader Justin constructed the group rationales for the group choices and the other members were not involved in the selection and rationale production process. Of the five group choice and group rationale tasks, Justin told Bob and Jack about his work once and that was because they were present in the virtual workspace when he was doing the work. Bob and Jack did not offer to join in the process in later tasks. Bill and Thomas were not aware of this. Bill said, "*I actually never helped with that… I don't know who made the decision, and who wrote down the group explanations*." Thomas also commented, "*I think someone went through and looked, whether it was Jack or Justin. I was never asked for my opinion except I ranked them in the assessment as individual. That was the only input*". The fact that the group decision was not made by the group and the group was not bothered by this indicates that there was at least an inclination towards satisfaction with the superficiality and pretense of knowledge that the group presented to the outside – the group decisions.

Group 2 and 3 members were more serious about choosing the correct challenges for the group that best represented the top and bottom challenges from the brainstormed list. They discussed their choices and had everyone write at least one group rationale. Some members even felt self-conscious about their challenges being selected as bottom ones. For example, a group 3 member told the researcher that he disliked the idea of choosing the bottom challenge, because "*I don't think anyone is losing any sleep over, but it's like* *oh man why that is a bottom challenge, that is a good one, but everyone else thinks it's a bottom challenge. So that's the only thing I don't like about it. It's just that little negative aspect*".

The same member explained in details how the group made the group choices, *"… then we moved to the rank challenges in which we would all take our individuals we have to have everyone's (challenges) out there so we can pick them, read the rationales, to see which ones we liked best and then based on commonality between the group picks, we will pick the team, and if there is a controversy or there isn't anyone's that show a pattern or repetition, then we have to sit down and actually really debate which ones are the top…*".

In fact, there was an incident in group 2 in which some members picked the top challenges which were considered bottom challenges by others for a particular project phase. The group members decided to meet face-to-face and went through everyone's rationales to make the group decisions. In summary, group 2 and 3 committed to a truth and depth of understanding when making decisions on the top and bottom three challenges.

5. Design Implications

The usefulness of Bereiter's contextual modules in describing learner groups suggests new directions for designing intentional learning groups to support and enrich individual learning experiences in collaborative learning activities. For example, the method of using a script for structured interaction (O'Donnell & Dansereau, 1992) can be used to instruct the group's division of labor practices. Recognizing the impact of the leader's contextual module on the whole group's learning experiences, we as designers of virtual group learning environment should consider designing toolkits that are specific for the group leaders for the purpose of reminding them the learning goals of the project and the benefits of group setting in learning processes, helping them organize group meetings by providing templates, and providing an overview of the group's progress and members' division of labor through visualization. For example, the educational tool can provide templates for dividing up the group task so as to scaffold the group practice of division of labor that acknowledges the individuals' learning goals and emphasizes the members' participation in knowledge construction in the project. Along with these templates, the tool will provide a visualization mechanism to indicate the amount of time each member spends on a specific part of the group tasks, reflecting on the members' participation in various aspects of the group activities.

6. Conclusion

Although numerous research studies have demonstrated the benefits of collaborative learning, there have also been reports on the failure of collaborative learning setting. Why do some groups learn well when others don't? Why do some groups' members enjoy the collaborative learning process and report enriched learning experiences while others don't? To answer questions like these, this study explored the use of Bereiter's contextual modules to compare three learner groups. The results show that groups can be interpreted using the intentional learning and schoolwork modules as well with adapted dimensions about the modules. Furthermore, the evaluation of collaborative learning activities should not only be based on the group performance but also the group process. Following this perspective, the paper suggests design implications for supporting intentional learning in the virtual environment.

This is only an initial step towards a systematic understanding of how contextual variables affect individual learning experiences in a collaborative learning setting. Instead of trying to generalize the results, which are based on just one classroom study, the intention of this paper is to shed light on the usefulness and usage of the concept of contextual modules in explaining learning groups and their contextual factors. This paper should be viewed as a starting point in a new research direction, rather than a conclusive summary of the research program. Further investigations are called for to use the modules to interpret the learning groups in different contexts and identify additional dimensions of contextual modules for learning groups.

Acknowledgement. I thank John M. Carroll and Mary Beth Rosson for their advice and support in this project. Craig Ganoe provided technical support of the workspace tool. I also thank Dr. Shawn Clark for offering his course for this research. Finally, I thank the research participants of the study.

References

1. Barros, B. & Verdejo, F. Analysing student interaction processes in order to improve collaboration: The DEGREE approach. Journal of Artificial Intelligence in Education, 11, 211-241, (2000)
2. Baumeister, R.F.A self-presentational view of social phenomena. Psychological Bulletin, 91, 3-26, (1982)
3. Bereiter, C. Aspects of an Educational Learning Theory, Review of Educational Research, Vol. 60, No. 4, Toward a unified approach to learning as a multisource phenomenon. pp. 603 – 624, 1990
4. Constantino-González, M. & Suthers D. Coaching collaboration in computer-mediated learning. In G. Stahl (Ed.), Computer support for collaborative learning: foundations for a CSCL community (pp. 583-584). Mahwah, NJ: Lawrence Erlbaum Associates, (2002)

5. Dillenbourg, P., Ott, D., Wehrle, T., Bourquin, Y., Jermann, P., Corti, D. & Salo, P. The socio-cognitive functions of community mirrors. In F. Flückiger, C. Jutz, P. Schulz and L. Cantoni (Eds). Proc. of the 4th International Conference on New Educational Environments. Lugano, May 8-11, (2002)
6. Houssman, J. Self Monitoring and Learning Proficiency. In Computer Classroom. Hofstra University, EDD, (1991)
7. Inaba, A. & Okamoto, T. Development of the intelligent discussion support system for collaborative learning. Proc. of Ed-Telecom '96. (pp 494-503), Bostoo. (1996)
8. Jermann, P. & Dillenbourg, P. An analysis of learner arguments in a collective learning environment. Proc. of the third CSCL Conference, pp. 265-273, Stanford, (1999)
9. Mason, Jason. “Qualitative Researching”. SAGE Publication Inc.: London. p3. (2002)
10. McGrath, J. Groups: Interaction and Performance, Prentice Hall College Div., (1984)
11. O'Donnell, A. M., & Dansereau, D. F. Scripted cooperation in student dyads: A method for analyzing and enhancing academic learning and performance. In R. Hertz-Lazarowitz and N. Miller (Eds.), Interaction in cooperative groups: The theoretical anatomy of group learning (pp. 120-141). London: Cambridge University Press, (1992)
12. Plotnick, L., Ocker, R., Hiltz, S. R., Rosson, M. B., Leadership Roles and Communication Issues in Partially Distributed Emergency Response Software Development Teams: A Pilot Study, Proceeding of the 41st Hawaii International Conference on System Sciences,(2008)
13. Roberson, Q., Justice in Teams: The Effects of Interdependence and Identification on Referent Choice and Justice Climate Strength, Social Justice Research, Vol.19, No. 3, pp. 323-344(22), (2006)
14. Scheffler, I. In praise of the cognitive emotions. Teachers College Record, 79, 171 – 186, (1977)
15. Xiao, L. Design for articulating and sharing rationales in virtual group learning activities, The IASTED International Conference on Technology for Education and Learning (Oct 24-25, Beijing, China), (2011)

Dr. Lu Xiao is an Assistant Professor in Faculty of Information & Media Studies at the University of Western Ontario. Influenced by activity theory and distributed cognition, her research work is on collaborative and social computing. Her main research interests include design to foster reflection in computer-mediated group activities, design to support child development in parent-young child activities, and design to support collaborative analysis activities of large scale data. Her work has been published in journals, e.g., Information & Organization, Online Information Review, Behaviour & Information Technology, and Journal of Community Informatics. She also presents her work in conferences such as CSCW, CSCL, PDC, ICALT, and iConference. Her most recent work on crowdsourcing study was presented at Collective Intelligence 2012.

Received: November 28, 2011; Accepted: February 02, 2012.

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

004

COMPUTER Science and Information Systems : the International journal / Editor-in-Chief Mirjana Ivanović. – Vol. 9, No 2 (2012) - . – Novi Sad (Trg D. Obradovića 3): ComSIS Consortium, 2012 - (Belgrade : Sigra star). –30 cm

Polugodišnje. – Tekst na engleskom jeziku

ISSN 1820-0214 = Computer Science and Information Systems
COBISS.SR-ID 112261644

Cover design: V. Štavljanin
Printed by: Sigra star, Belgrade

ComSIS Computer Science and Information Systems

Volume 9, Number 2, June 2012

Contents

Regular Papers

Special Section: Emerging Trends in Technology Enhanced Learning

ISSN: 1820-0214

Computer Science and Information Systems

Vol 9, No 2, June 2012