Com
SIS

**Computer Science
and Information Systems**

Volume 18, Number 4, September 2021

Contents

Papers

Special Section on Pattern Recognition, Optimization, Neural Computing and Applications in Smart City

ISSN: 1820-0214 (Print) 2406-1018 (Online)

Computer Science and Information Systems

Vol 18, No 4, September 2021

Com
SIS

Computer Science and Information Systems

Published by ComSIS Consortium

Volume 18, Number 4
September 2021

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
Technical Faculty “Mihajlo Pupin”, Zrenjanin, Serbia

University of Niš:
Faculty of Electronic Engineering, Niš, 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 Obradovića 4, 21000 Novi Sad, Serbia
Phone: +381 21 458 888; **Fax:** +381 21 6350 458
www.comsis.org; Email: comsis@uns.ac.rs

Volume 18, Number 4, 2021
Novi Sad

Computer Science and Information Systems

ISSN: 1820-0214 (Print) 2406-1018 (Online)

The ComSIS journal is sponsored by:

Ministry of Education, Science and Technological Development of the 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. 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 2020 two-year impact factor 1.167,
- 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 20 (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 18, Number 4, September 2021

CONTENTS

Papers

Special Section on Pattern Recognition, Optimization, Neural Computing and Applications in Smart City

Computer Science and Information Systems 18(4): i–ii https://doi.org/10.2298/CSIS210400iI

Editorial

Mirjana Ivanović[1], Miloš Radovanović[1], and Vladimir Kurbalija[1]

University of Novi Sad, Faculty of Sciences
Novi Sad, Serbia
{mira,radacha,kurba}@dmi.uns.ac.rs

Before we introduce this fourth issue of Volume 18 of Computer Science and Information Systems, we are very glad to announce the impact factors of our journal, updated for 2020: the two-year IF rose to 1.167, and the five-year IF to 0.974. This is the first time since our journal's inception that we reached an impact factor higher than 1. This achievement certainly cannot be attributed only to the efforts of the editorial team, but all our creative authors whose high-quality articles, in emergent topics in ICT, attracted the citations needed to increase impact. Also, let us not forget the reviewers who recognized the potential of the articles and in many cases helped improve them with their diligent reviews.

This issue contains 10 regular articles and 6 articles in the "Special Section on Pattern Recognition, Optimization, Neural Computing and Applications in Smart City" guest-edited by Mu-Yen Chen, Jose de Jesus Rubio, and Arun Kumar Sangaiah.

The first regular article, "Buffer-Based Rate Adaptation Scheme for HTTP Video Streaming with Consistent Quality" by Jiwoo Park et al. presents a playback buffer model for rate adaptation and proposes a new buffer-based rate adaptation scheme for HTTP-based adaptive streaming (HAS) of video. Experimental results show that the proposed scheme achieves higher video quality than conventional algorithms and can cope with various environments without tuning of configuration parameters.

The second article, "Deep Semi-supervised Learning with Weight Map for Review Helpfulness Prediction" authored by Hua Yin et al. proposes an end-to-end deep semi-supervised learning model with weight map, which makes full use of the unlabeled reviews in the task of review helpfulness prediction. Training is divided into three stages: obtaining the base classifier, iteratively applying weight map strategy on large unlabeled reviews to obtain pseudo-labeled reviews, training on the combined reviews to obtain the re-trained classifier.

"Cooperation and Sharing of Caught Prey in Competitive Continuous Coevolution Using the Predator-Prey Domain" by Krisztián Varga and Attila Kiss presents a simulation of the predator-prey domain (with carnivores, herbivores, and plants) and continuous (not generation based) neuro-evolution to create a complex environment where two forms of competition arise: between predator and prey, but also between individuals of the same species. The simulation sheds light on questions about the importance of cooperation and sharing in such complex competitive environments.

In the article entitled "Using Honeynet Data and a Time Series to Predict the Number of Cyber Attacks," Matej Zuzčák and Petr Bujok present multiple methods for using real-world time-series data to predict cyber-attacks on home computers, mobile devices, and servers over secure shell (SSH). It focuses on the overall prediction of attacks on the honeynet and the prediction of attacks from specific geographical regions using multiple approaches like ARIMA, SARIMA, GARCH, and Bootstrapping.

Masoud Reyhani Hamedani and Sang-Wook Kim, in "SimAndro-Plus: On Computing Similarity of Android Applications," propose SimAndro-Plus as an improved variant of the SimAndro state-of-the-art method for computing the similarity of Android applications with regards to their functionality. The proposed method introduces two improvements: (1) it exploits two beneficial features to similarity computation that are disregarded by SimAndro, and (2) to compute the similarity of an app-pair based on strings and package name features, SimAndro-Plus considers not only the terms co-appearing in both apps, but also terms appearing in one app while missing from the other.

"Analysis of Entrepreneur Mental Model and Construction of its Portrait," authored by Yongzhong Zhang et al. first summarizes three key factors that affect entrepreneurial mental models: prior knowledge, personality characteristics and opportunity perception. Then, methods for the construction of entrepreneur mental portraits are introduced, which include a cluster analysis method and a fuzzy comprehensive evaluation method, providing a meaningful reference for promoting innovation and entrepreneurship education and training.

Jianjun Li and Jia Liao in their article "Research on Influencing Factors of the Development of Cultural and Creative Industries Based on Grey Factor Analysis" study the influencing factors of cultural and creative industries (CCIs) using Grey Factor Analysis and 30 different indexes to empirically analyze the correlation between the influencing factors and the added value of CCIs in the city of Shanghai, highlighting the importance of technology research and development, policy and government financial support, human resources, social culture, cultural consumption environment, cultural industry basis and development status.

The article "Conversational Agent for Supporting Learners on a MOOC on Programming with Java," by Cristina Catalán Aguirre et al. addresses an important problem in massive open online courses (MOOCs), the lack of personalized support from teachers, by evaluating JavaPAL, a voice-based conversational agent offered on edX for supporting learners on programming with Java. Agent usability, learners' performance and interviews with users are evaluated and used to determine the helpfulness of JavaPAL.

"Assessing Learning Styles Through Eye Tracking for E-Learning Applications" by Nahumi Nugrahaningsih et al. investigates the possibility to distinguish between visual and verbal learning styles from gaze data. In an experiment involving first year students of an engineering faculty, content regarding the basics of programming was presented in both text and graphic form, and participants' gaze data was recorded by means of an eye tracker. Results show a significant relation between gaze data and visual/verbal learning styles for an information arrangement where the same concept is presented in both graphical and text formats.

The final regular article entitled "Compensation of Degradation, Security, and Capacity of LSB Substitution Methods by a New Proposed Hybrid n-LSB Approach," by Kemal Tütüncü and Özcan Çataltaş, proposes a new hybrid n-LSB (Least Significant Bit) eight substitution-based image steganography method in the spatial plane. The previously proposed n-LSB substitution method by the same authors is combined with the Rivest-Shamir-Adleman (RSA), RC5, and Data Encryption Standard (DES) encryption algorithms to improve the security of the steganography as judged by multiple standard criteria.

Computer Science and Information Systems 18(4): iii–iv https://doi.org/10.2298/CSIS210400iiiC

Guest Editorial – Pattern Recognition, Optimization, Neural Computing and Applications in Smart City

Mu-Yen Chen[1], Jose de Jesus Rubio[2], and Arun Kumar Sangaiah[3]

[1] National Cheng Kung University, Taiwan
[2] Instituto Politécnico Nacional, Mexico
[3] Vellore Institute of Technology, India

Machine Learning was coined in 1980's. It comes under the category of Artificial Intelligence. Without being explicitly programmed by human or assistance, Machine Learning gives the opportunity to the computer to learn automatically. The primary aim is to allow the computer learn automatically without the human intervention. But it has the limitation of handling only smaller dimensional data with lesser amount of inputs and parameters. Due to this drawback, Deep Learning was introduced. Deep Learning on the other hand is the enhancement of Machine Learning which can handle any number of high dimensional data as well as greater number of inputs and outputs. Due to this advancement it can handle complex model in easier manner. Since Deep Learning uses multiple layers to extract high level features from input, it can work with various disciplines such as Biomedical, Computer Vision, Handwriting Recognition etc.

The idea of smart city requires connecting every related matter with the Internet tightly; from public facilities to municipal management systems, it has to integrate information technology with the Internet of Things (IoT) to enhance the quality of life and the resource management of the city. Through collecting various types of data by IoT, utilizing cloud spaces or other types of storage equipment to share data, and conducting big data research to analyze relevant issues could support the municipal decision-making. To improve resource efficiency, all of the devices in a smart city system, including transportation, medical care, electricity, disaster prevention etc, could conduct big data and artificial intelligence analyses to understand the usage of user traffic, logistics, and resource. With the high-performance technology in smart city systems, such as cloud computing, fog computing, and high-consumption sensors, to handle massive data for satisfying the demands from the public.

The first article entitled" The Dynamic Two-echelon MSW Disposal System Study under Uncertainty in Smart City", authors construct a grey fuzzy multi-objective two-echelon MSW (municipal solid waste) allocation model. According to the result, the MSW is prior to be allocated to RDF (Refuse Derived Fuel) plant and incineration plant.

The second article entitled "The Application of E-commerce Recommendation System in Smart Cities based on Big Data and Cloud Computing", authors present one comprehensive evaluation system based on improved collaborative filtering recommendation algorithm under the Hadoop cloud computing platform. The experiential results showed the proposed model has more efficient than the traditional single machine environment.

The third article entitled "Optimization of Intelligent Heating Ventilation Air Conditioning System in Urban Building based on BIM and Artificial Intelligence Technology", this research proposes the energy consumption of building HVAC (heating ventilation air conditioning) system by combining back propagation neural network (BPNN) and Adaboost algorithm. The experimental results illustrate the proposed hybrid Adaboost-BP algorithm can be useful to predict the energy consumption of the air conditioning system.

The fourth article entitled" Face Recognition Based on Full Convolutional Neural Network Based on Transfer Learning Model", authors develop an adaptive scale feature extraction method based on convolutional neural network (CNN). This research also adopts the transfer learning approach to construct the sketch face recognition model by using the training sample. The proposed model can reach about 97.4% and the accuracy rate has been outperformed than the traditional sketch face recognition algorithm.

The fifth article entitled "Background Modeling from Video Sequences via Online Motion-Aware RPCA", authors propose a novel online motion-aware RPCA (robust principal component analysis) algorithm, named OM-RPCAT, which adopt truncated nuclear norm regularization as an approximation method for low rank constraint. In this research, the dataset is used scene background initialization (SBI) and the experimental results illustrate the proposed algorithm has the better performance than traditional online RPCA algorithms.

Finally, in the last article entitled “A Novel Network Aligner for the Analysis of Multiple Protein-protein Interaction Networks”, authors present the Accurate Combined Clustering Multiple Network Alignment (ACCMNA) method. It is a novel and accurate multiple network alignment algorithm. After several performance evaluations, the results illustrate the proposed ACCMNA algorithm outperforms better than traditional PPINs (protein-protein interaction networks) of various sizes within an acceptable running time.

Acknowledgments. The guest editors are thankful to our reviewers for their effort in reviewing the manuscripts. We also thank the Editor-in-Chief, Professor Mirjana Ivanovic, for her supportive guidance during the entire process.

Computer Science and Information Systems 18(4):1139–1157 https://doi.org/10.2298/CSIS200820021P

Buffer-Based Rate Adaptation Scheme for HTTP Video Streaming with Consistent Quality

Jiwoo Park, Minsu Kim, and Kwangsue Chung

Department of Electronics and Communications Engineering,
Kwangwoon University, Seoul, South Korea
{jwpark, mskim}@cclab.kw.ac.kr, kchung@kw.ac.kr

Abstract. Recently, HyperText Transfer Protocol (HTTP) based adaptive streaming (HAS) has been proposed as a solution for efficient use of network resources. HAS performs rate adaptation that adjusts the video quality according to the network conditions. The conventional approaches for rate adaptation involve accurately estimating the available bandwidth or exploiting the playback buffer in HAS clients rather than estimating the network bandwidth. In this paper, we present a playback buffer model for rate adaptation and propose a new buffer-based rate adaptation scheme. First, we model the playback buffer as a queueing system that stores video segments. The proposed scheme selects the next video bitrate that minimizes the difference between the current buffer occupancy and the expected value from the playback buffer model. The evaluation results indicated that the proposed scheme achieves higher video quality than conventional algorithms and can cope with various environments without the tuning of the configuration parameters.

Keywords: adaptive algorithms, queueing analysis, streaming media, transport protocols.

1. Introduction

Global Internet video traffic has been growing rapidly with the emergence of popular video streaming services such as YouTube, Netflix, and Amazon Prime. According to Cisco's Visual Networking Index, worldwide video traffic accounted for 75% of total Internet traffic in 2017 and is expected to reach 82% by 2022 [1]. To handle the increasing video traffic, many video service providers adopt adaptive bitrate streaming technology to provide the best possible streaming experience for users. Recently, HyperText Transfer Protocol (HTTP) based adaptive streaming (HAS) technology has attracted attention owing to the simplicity of its implementation and deployment [2]. In contrast to the existing real-time transport protocol (RTP) based streaming technology, which transmits video packets through User Datagram Protocol (UDP), HAS streams video over HTTP/Transmission Control Protocol (TCP), which is a traditional protocol stack used to deliver web messages. Video streaming technologies such as Microsoft’s Smooth Streaming, Apple’s HTTP Live Streaming, and Adobe’s HTTP Dynamic Streaming rely on HTTP-based adaptive bitrate streaming [3-6]. In the HAS system, the video content is encoded at various bitrates, and the encoded video content is divided into small video segments of a certain length and stored in the HTTP web server [7]. The HAS client sends an HTTP GET message to download the video segments. The transmitted video segment is stored in the playback buffer of the client, and when enough video segments are stored, the decoder consumes the first video segment and displays it on the screen.

Research on HAS is being actively conducted to improve the service quality and user experience. A general research topic is a methodology for improving the performance of the rate adaptation algorithm implemented in the HAS client and applying it to various environments [8]. In the conventional video streaming service, quality degradation is caused by the interruption of playback or the distortion of the image in a situation where the network bandwidth is insufficient. Because HAS dynamically adjusts the video bitrate, unnecessary changes in video quality can make users feel uncomfortable.

Recent studies have shown that requesting a segment in an ON-OFF pattern to maintain the buffer occupancy causes the HAS client to incorrectly measure the available bandwidth and repeat unnecessary quality changes in a multi-client environment [9]. To solve this problem, techniques for bandwidth measurement and playback buffer-based adaptation methods have been studied [10-12]. The existing approach is expected to improve the performance of HAS by accurately measuring the available bandwidth or setting a threshold for the buffer occupancy. However, most of the conventional approaches have been designed by targeting to a specific scenario. This leads to require the setting of configuration parameters such as weights and thresholds, degrading adaptability to the various scenarios. The conventional approaches are hard to achieve consistent quality for the media-consumption environments that the network bandwidth, videos watching by users, and number of users are changing over time.

In this paper, we propose a buffer-based rate adaptation scheme to achieve consistent quality for HAS. The main contributions of the proposed scheme are as follows.

- We analyze the relationship among the video bitrate, network bandwidth, and playback buffer occupancy of HAS.
- We present a playback buffer model for rate adaptation by considering the analyzed results for the relationship among the affecting factors to the performance of HAS.
- We then propose a novel rate adaptation scheme that controls the video bitrate by using the current buffer occupancy and the average buffer occupancy predicted by the playback buffer model.
- To compare the performance of the proposed scheme with the conventional approaches, we perform simulations in various network environments by using the ns-3 network simulator.

The reminder of the paper is organized as follows. Section 2 reviews related work on HTTP-based adaptive streaming. Section 3 presents the playback buffer model for HAS. Section 4 describes the proposed rate adaptation scheme and its buffer-based adaptation algorithm. Section 5 presents the results of the proposed scheme, and Section 6 concludes the paper.

2. Background and Related Work

In this section, we describe the basic operation of the HAS system and analyze the behavior of the HAS client. We also classify rate adaptation schemes according to adaptation factors such as the bandwidth, buffer occupancy, and video bitrate.

2.1. HAS System

As the demand for video streaming over the Internet increases, various technologies and standards have been proposed and developed. Recently, HTTP-based adaptive streaming has attracted attention owing to its efficient use of limited network resources and fast start-up time. Conventional streaming technology frequently uses the RTP over UDP, which does not perform error recovery for fast media delivery. By using HTTP, network address translation and firewall problems of existing streaming protocols can be easily solved. The HAS system also has the advantage of a low implementation cost because it can use the existing HTTP web servers and cache servers that are already installed globally.

In the HAS system, the HTTP web server stores video contents encoded with different resolutions, frame rates, and bitrates depending on the quality level. Each video is divided into segments of short length. The HAS client requests consecutive video segments while performing rate adaptation to adapt the video bitrates to the changing network environment. In general, rate adaptation algorithms use segment throughput to estimate the available bandwidth.

2.2. Behavior of HAS Client

At the beginning of the streaming, the HAS client quickly fills the playback buffer by continuously requesting video segments in the buffering state to prevent playback stalling. When the playback buffer is full, it periodically requests video segments in the steady state. Fig. 1 shows that if the video bitrate is lower than the network bandwidth, the HAS client has an ON-OFF pattern in the steady state. Owing to this ON-OFF pattern, the available bandwidth may be inaccurately measured in an environment where multiple HAS clients compete.

There are two typical problems caused by HAS clients having an ON-OFF pattern. The first problem is that the HAS client underestimates the available bandwidth because the TCP connection is idle during the OFF period [13]. When a TCP sender does not send or receive data for more than one retransmission timeout, the TCP congestion window is reduced to the initial value, and the TCP connection restarts slow-start after an idle period [14]. Unnecessary slow-start reduces the TCP throughput. For example, whenever an HAS client restarts slow-start while competing with greedy TCP flows, the throughput of the HAS client gradually decreases, and the client is unable to obtain the fair share of bandwidth.

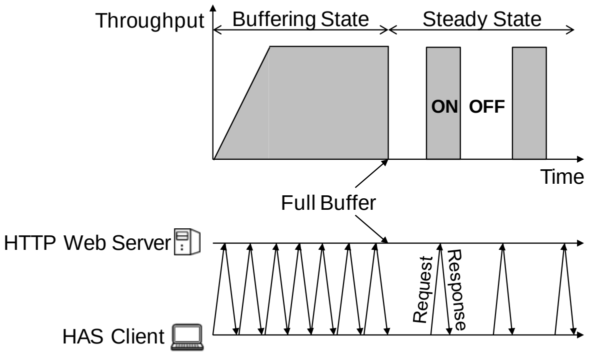


Fig. 1. Request pattern of the HAS client

Another problem is that HAS clients overestimate the available bandwidth when there are multiple HAS clients in the same network and they operate in an ON-OFF pattern [10]. Fig. 2 shows that the download duration varies depending on the overlap of ON periods when two HAS clients request video segments of the same size. Because the available bandwidth is estimated according to past segment throughputs, HAS clients may overestimate the bandwidth if the ON period is not overlapped. For example, if multiple HAS clients overestimate the available bandwidth and unnecessarily improve the video quality simultaneously, the network bandwidth becomes insufficient and network congestion occurs, resulting in poor video quality. In summary, the ON-OFF pattern is known to be a typical factor that degrades the quality of HAS services.

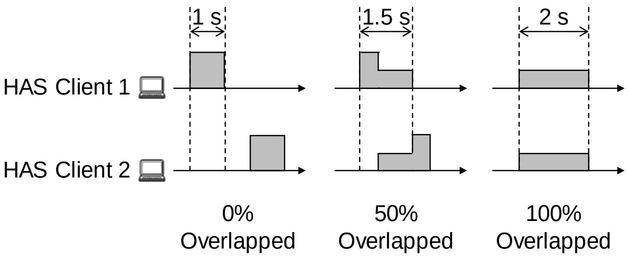


Fig. 2. Download duration of two competing HAS clients in the same network

2.3. Rate Adaptation Schemes

Although HAS is a relatively new application, its popularity has resulted in considerable research. In particular, the rate adaptation scheme is an interesting research topic because it automatically adjusts the video quality to provide video to users at the maximum possible quality. We begin by reviewing the rate adaptation scheme and then describe the key shortcomings of state-of-art solutions.

The most basic method of rate adaptation is to select the highest quality while ensuring a video bitrate lower than the available bandwidth. In general, the rate adaptation scheme is implemented in the client to reduce the load on the server.

- **Estimating**: Estimate the available network bandwidth by measuring the per-segment throughput from the previous segment request.
- **Smoothing**: Remove noise from estimates by applying an exponentially weighted moving average filter or a harmonic mean filter.
- **Quantizing**: Select the video quality using the smoothed version of the estimated bandwidth.
- **Scheduling**: Determine the next request time according to the playback buffer occupancy.

We can classify existing rate adaptation schemes into two main categories: bandwidth-based and buffer-based. Bandwidth-based rate adaptation controls the video bitrate according to the estimated available bandwidth. Early rate adaptation schemes adopted by commercial video providers belong to this category. Because the video quality mainly depends on the accuracy of the bandwidth estimation, measurement techniques considering the network type and traffic characteristics have been proposed. The dash.js video player provided by the DASH Industry Forum estimates the future throughput by using the average throughput of the last three segments to mitigate fluctuations in the bandwidth measurement [15]. PANDA predicts the available bandwidth in a manner similar to TCP congestion control and prevents the problem of ON-OFF patterns when multiple clients share bottleneck links [12]. PANDA updates the segment throughput in an additive increase and multiplicative decrease (AIMD) manner. Under complex network conditions, it is still challenging to accurately predict the network bandwidth.

Buffer-based rate adaptation selects the video bitrate according to the occupancy of the playback buffer implemented in HAS clients. In the buffer-based approach, the video quality is proportional to the buffer occupancy. A few studies have addressed the buffer-based approach to model the playback buffer [16-20]. BBA performs rate adaptation using a function that linearly maps the current buffer occupancy to the video bitrate [21]. It also divides the playback buffer into three sections, and its performance is determined by the length of each section. In [22], the authors modeled the playback buffer as an M/M/1 queue to characterize buffer starvations.

Because rate adaptation schemes are designed to improve the performance of HAS in a specific scenario, they make direct or indirect assumptions regarding the target environment. Most of them also require the setting of configuration parameters, such as weights and thresholds. These are often set arbitrarily through experiments. While fixed parameters may be adequate in certain scenarios, they cannot achieve consistent quality for all scenarios. Therefore, we must identify the factors affecting the video quality and consider these factors for rate adaptation. Clearly, the bandwidth and buffer are the main factors in rate adaptation. However, the bandwidth and buffer are treated separately, and the relationship between them is not well-considered in conventional approaches. In this paper, we present a playback buffer model for HAS clients and analyze the relationship among the bandwidth, buffer occupancy, and video bitrate using queueing theory.

3. Playback Buffer Model for HAS

In this section, we formalize the playback buffer for HAS clients. Before presenting the playback buffer model, we first define the symbols and terms used in the paper, as shown in Table 1.

Table 1. Notation used in this paper

Notation	Definition
r_n	Video bitrate of the n^{th} segment
R_m	Video bitrate of the m^{th} quality level
x	Segment throughput
τ	Segment duration
b	Buffer occupancy
$b_{\max}$	Buffer capacity
k	Number of segments in the buffer
λ	Segment arrival rate
μ	Segment service rate
ρ	Traffic intensity of the buffer
c	Coefficient of variation
N	Average number of segments in the buffer
K	Maximum number of segments in the buffer
W	Average waiting time of buffer

In this paper, the buffer occupancy of HAS clients is expressed in units of time. As shown in Fig. 3, the buffer occupancy is reduced by the time the video is played, and when the segment download is completed, the segment duration is added to the buffer occupancy. We can model the playback buffer as a queue that stores and processes video segments, as shown in Fig. 4.

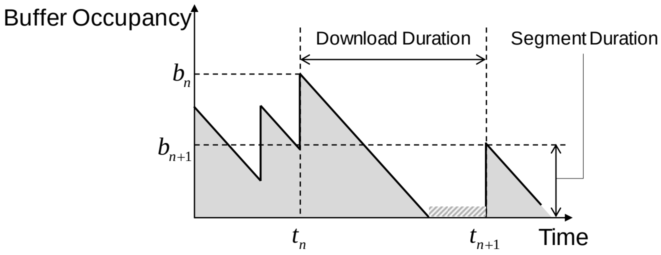


Fig. 3. Buffer occupancy of the HAS client

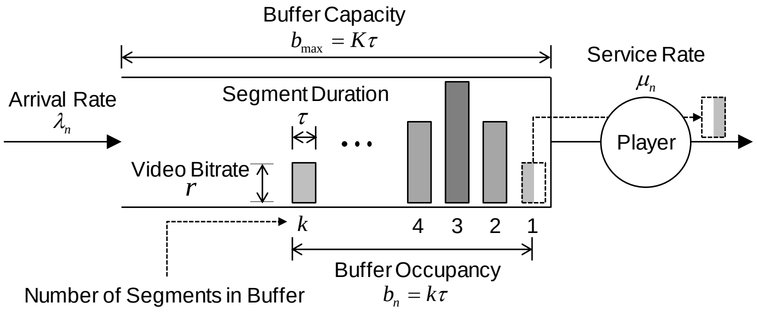


Fig. 4. Queuing model of the playback buffer

In the playback buffer model, the arrival rate is the number of video segments arriving per unit time. The nth arrival rate can be calculated using the estimated network throughput and the size of the nth video segment, as follows.

$$\lambda_n = \frac{x_n}{r_n \cdot \tau} \tag{1}$$

Unless playback is paused, the HAS client consumes one video segment during each segment duration in the steady state. In this case, the service time is equal to the segment duration, and the service rate can be expressed as follows.

$$\mu_n = \frac{1}{\tau} \tag{2}$$

The buffer occupancy is updated according to the following equation.

$$b_n = \max\left(0, b_{n-1} - \left(t_n - t_{n-1}\right)\right) + \tau \tag{3}$$

Here, t_n is the time at which the nth segment download is completed. Assuming that we have a continuous analogue of b_n, the following relationship is satisfied.

$$\frac{db(t)}{dt} = \lambda(t) - \mu(t) \tag{4}$$

Equation (4) shows that the buffer occupancy of HAS clients can be mathematically modeled as a non-linear differential equation.

Because video segments are transmitted over the network and consumed at a constant rate, we suppose that the arrival rate follows a certain probability distribution and that the service rate is fixed. Thus, we model the playback buffer as a G/D/1/K queue, where G represents interarrival times, which have a general distribution; D represents service times, which are deterministic; and K represents the queue size. Because the analytic solution of the G/D/1/K queuing model is unknown and is very difficult to obtain, we use an approximation to predict the average buffer occupancy. In the playback buffer model, the buffer occupancy is equal to the time to wait in the playback buffer until the most recently received segment is decoded.

Kingman’s formula is the most widely used approximation for the mean waiting time in a G/G/1 queue [23].

$$E\left[W_{GG1}\right] \approx \left(\frac{c_a^2 + c_s^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\frac{1}{\mu} \tag{5}$$

Here, $\rho = \lambda / \mu$ is the traffic intensity, which represents how busy a queueing system is. Because the second term in (5) represents the average number of elements in an infinite queue, it must be modified for a finite queue. If the interarrival time and the service time follow the exponential distribution, their coefficients of variation (CVs) are equal to 1, and (5) becomes an equation for M/M/1 queues. An M/M/1/K queue is the finite version of the M/M/1 queue, and its mean waiting time is calculated via summation instead of an infinite series. The mean waiting time for an M/M/1 queue is twice that for an M/D/1 queue. The mean waiting time of M/M/1, M/M/1/K, and M/D/1 queues can be expressed as following equation when $0 < \rho < 1$.

$$\begin{aligned}\lim_{K\to\infty} W_{MM1K} &= \lim_{K\to\infty}\left(\frac{\rho}{1-\rho}\right)\left(\frac{1-(K+1)\rho^K + K\rho^{K+1}}{1-\rho^{K+1}}\right) \\ &= \frac{\rho}{1-\rho} = W_{MM1} = 2\cdot W_{MD1}\end{aligned} \tag{6}$$

According to the relationship among M/M/1, M/M/1/K, and M/D/1 queues, we can predict the mean waiting time of a G/D/1/K queue as follows.

$$E\left[W_{GD1K}\right] \approx \frac{c_a^2}{2}\left(\frac{\rho}{1-\rho}\right)\left(\frac{1-(K+1)\rho^K + K\rho^{K+1}}{1-\rho^{K+1}}\right)\frac{1}{\mu} \tag{7}$$

The CV of the service time c_s is removed because the standard deviation of the service time is equal to 0 in our model. When the playback buffer is in the steady state, where ρ converges to 1, (7) converges to the following equation.

$$\lim_{\rho\to1} E\left[W_{GD1K}\right] = \frac{c_a^2}{2}\cdot\frac{K}{2}\cdot\frac{1}{\mu} = \frac{c_a^2 \tau K}{4} \tag{8}$$

4. Proposed Buffer-Based Rate Adaptation

This section introduces the proposed rate adaptation scheme, which measures the buffer information rather than the network bandwidth for rate adaptation. Fig. 5 shows a block diagram of the proposed scheme in the HAS system.

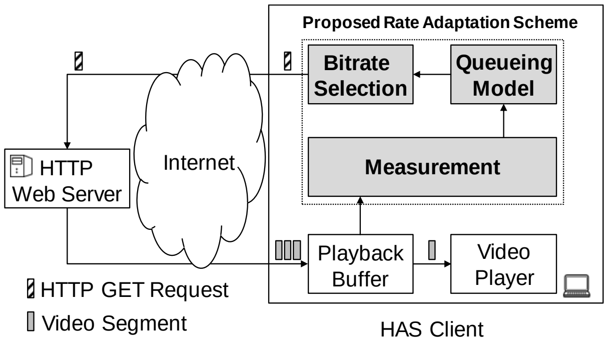


Fig. 5. Block diagram of the proposed scheme

4.1. Measurement

In this step, the proposed scheme measures the available bandwidth and the interarrival time of segments for the playback buffer model. Measurement of the available bandwidth in the HAS client is not accurate, because it is performed in the application layer and involves measurement error. The approximate available bandwidth is predicted via the smoothing of bandwidth samples.

There are many ways to take an average, such as the arithmetic mean, harmonic mean, and moving average. The proposed scheme uses all samples and updates the previous average using the current sample, as follows.

$$A_n = \frac{1}{n}\sum_{i=1}^{n} x_i = A_{n-1} + \frac{1}{n}\left(x_n - A_{n-1}\right) \tag{9}$$

The arithmetic mean of the samples is calculated using (9). The network bandwidth is expressed in terms of the bitrate, i.e. the number of bits transferred per unit of time. When calculating the average of rates, such as speed, bitrate, and bandwidth, the harmonic mean is a more appropriate method than the arithmetic mean. The proposed scheme estimates the network bandwidth using the following equation.

$$H_n = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1} = \left(\frac{1}{H_n} + \frac{1}{n}\left(\frac{1}{x_n} - \frac{1}{H_{n-1}}\right)\right)^{-1} \tag{10}$$

We also calculate the CV of the interarrival time, which is a variable in the playback buffer model. The proposed scheme records the arrival time of each segment and calculates the interarrival time. The average of the interarrival times is calculated using (9). Because only a squared CV is needed, we calculate the variance of the interarrival time, as follows.

$$\sigma_n^2 = \left(1 - \frac{1}{n}\right)\left(\sigma_{n-1}^2 + \frac{1}{n}\left(x_n - A_{n-1}\right)^2\right) \tag{11}$$

In the proposed scheme, the CV of the interarrival time represents the network variability, which indicates how often the network changes. However, because there are insufficient data at the beginning of the streaming, the CV does not contain meaningful information. It may be a smaller than predicted using the proposed playback buffer model. To prevent this, we calculate the CV using (9) and (11) and set the minimum value as follows.

$$c_n^2 = \max\left(1, \left(\frac{\sigma}{A_n}\right)^2\right) \tag{12}$$

4.2. Updating Queuing Model

The proposed scheme estimates the average buffer occupancy through the queuing model to perform buffer-based rate adaptation. We define the expected average buffer occupancy *B* as a function of the estimated bandwidth and the CV of the interarrival time.

$$B_n = \frac{c_n^2 \tau}{2}\left(\frac{H_n}{r - H_n}\right)\left(\frac{r^{K+1} - (K+1) r H_n^K + K H_n^{K+1}}{r^{K+1} - H_n^{K+1}}\right) \tag{13}$$

Equation (13) gives the relationship among the available bandwidth, buffer occupancy, and video bitrate. We observe that the buffer capacity, segment duration, and CV of the interarrival time affect the buffer occupancy and represent the variability of the device, video, and network, respectively. In this model, the buffer capacity and segment duration are fixed before performing rate adaptation. The proposed scheme updates the queuing model using the data measured in the previous step and thus is able to take into account the variability of the surrounding environment.

4.3. Bitrate Selection

When choosing the video bitrate according to the buffer occupancy, it is necessary to prevent buffer underflow and overflow, which adversely affect the performance, and simultaneously improve the average video quality. Buffer underflow and overflow can be resolved by keeping the buffer occupancy constant. By selecting the video bitrate in proportion to the buffer occupancy, the video quality can be improved as the playback buffer is filled.

If we derive a function such as $f(H_n, B_n) = r$ from (13), we can easily select the appropriate video bitrate. However, it is impossible to obtain the inverse of a multivariate nonlinear function in an analytical manner. Therefore, the proposed scheme follows a heuristic method to determine the video bitrate according to the buffer occupancy. If the video bitrate can be selected to set the buffer occupancy to the target occupancy, the playback buffer can remain stable. The proposed scheme selects the next video bitrate that minimizes the difference between the buffer occupancy and the expected value from (13), as follows.

$$r_{n+1} = \arg\min_R \left|(B_{\max} - B_n) - b_n\right| \tag{14}$$

Here, $B_{\max}$ is the maximum predictable value of the average buffer occupancy and satisfies the following equation.

$$B_{\max} = \lim_{x\to\infty} B_n = 2 \cdot \lim_{x\to r} B_n \tag{15}$$

Fig. 6 illustrates the proposed bitrate selection in a two-dimensional space. Suppose that a video is encoded at five bitrates $\{R_1, R_2, R_3, R_4, R_5\}$. Then, we can draw five points on the $B_{\max} - B$ curve. A larger index of R represents a higher video bitrate. The position of each point is determined by the ratio of the bandwidth to the encoded bitrate. For example, as the bandwidth increases, the points move to the right along the curve. In accordance with (13), the buffer capacity determines the shape of the curve, and the slope increases with the capacity. The CV of the interarrival time and the segment duration scale the curve vertically. The proposed scheme finds the closest point to the buffer occupancy line. Thus, video streaming starts with the lowest bitrate, but a higher bitrate is selected as the buffer occupancy increases from 0. If the selected bitrate satisfies $0 < x/r < 1$, the buffer occupancy decreases because the video bitrate exceeds the network throughput. Conversely, the buffer occupancy increases when the network speed is higher than the selected bitrate. If the rate adaptation is performed for a sufficient time in the proposed method, the buffer occupancy converges to $B_{\max}/2$ and remains stable unless the network bandwidth changes significantly.

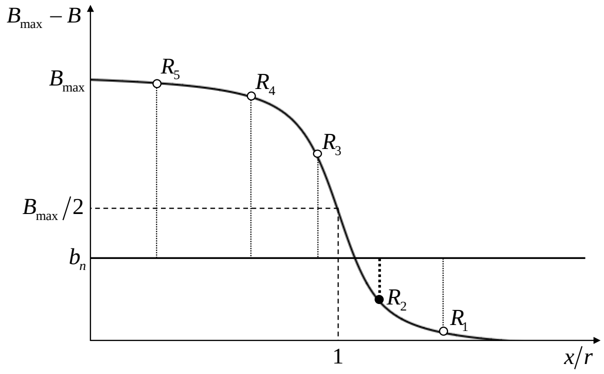


Fig. 6. Proposed buffer-based bitrate selection

5. Simulation Results

We performed a set of simulations to evaluate the proposed scheme in comparison with other conventional algorithms using the ns-3 network simulator [24]. To objectively evaluate the performance of the rate adaptation scheme, we implemented the HAS system in addition to the ABR, PANDA, and BBA algorithms. ABR is a basic rate adaptation scheme, and PANDA and BBA are the most representative algorithms for rate adaptation. A brief description of each algorithm is presented as follows.

- **ABR** estimates the available bandwidth through an arithmetic mean of the three most recent segment throughputs and selects the highest video bitrate that is lower than the measured available bandwidth [15].
- **PANDA** performs AIMD-like bandwidth estimation with an additive increment w and a multiplicative factor κ [12]. We used 0.3 and 0.28 as the defaults for w and κ, respectively.
- **BBA** uses a lower threshold of 90 s and an upper threshold of 24 s, for a buffer capacity of 240 s [21]. We set the thresholds at ratios of 3/8 and 9/10 for the variable buffer capacity.

5.1. Experimental Setup

As shown in Fig. 7, in all the simulation, a simple dumbbell network topology including TCP and UDP applications was used for generating competing traffic according to network profiles.

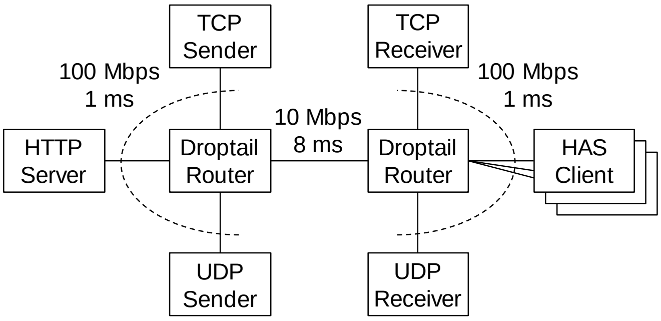


Fig. 7. Network topology used in the simulations

Table 2. Network profiles used in the simulations

Network Profile	Period (s)	Min (kbps)	Max (kbps)	Pattern
1	30	1500	5000	High-low-high
2	30	1500	5000	Low-high-low
3	-	2529	4110	FTP, Exponential ON-OFF, Pareto ON-OFF

We used two basic patterns for the network profile, i.e. high–low–high and low–high–low, according to the guideline of the DASH Industry Forum [25]. To simulate a general network environment, we constructed a network profile that generated highly variable traffic by combining three patterns of traffic models: FTP, Exponential ON-OFF, and Pareto ON-OFF. FTP is a file transfer protocol that sends packets using multiple TCP connections and thus transmits data at the maximum possible speed.

Exponential ON-OFF is a traditional traffic model of circuit-switched networks, whereas Pareto ON-OFF traffic represents a bursty characteristic of packet-switched networks. Detailed information regarding the network profiles is presented in Tables 2 and 3. Fig. 8 shows the bitrate changes of each profile in the simulation.

Table 3. Detailed settings of network profile 3

Pattern	Characteristic	Configuration
FTP	Greedy	TCP NewReno Always ON

Exponential ON-OFF	Poisson/Memoryless	BurstTime = 0.8 s
		IdleTime = 0.2 s
		Rate = 3 Mbps

Pareto ON-OFF	Long-tail/Bursty	BurstTime = 0.5 s
		IdleTime = 0.5 s
		Rate = 3 Mbps

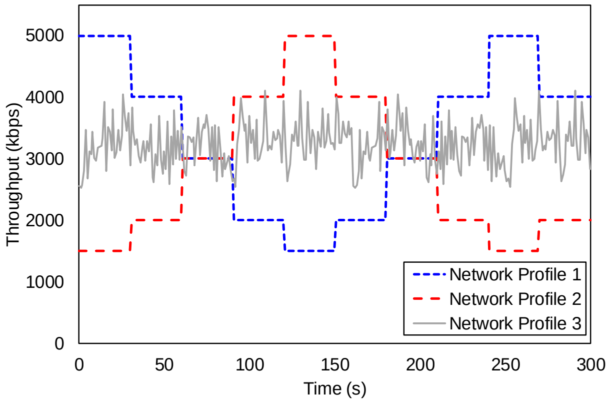


Fig. 8. Bitrate change in the network profiles

The video sample used in the experiment was encoded with six levels of quality, according to YouTube's recommended encoding settings [26]. As the quality level increased, the video bitrate increased exponentially. Table 4 presents the resolution and bitrate for each quality level. For further experiments, we divided the video sample into video segments 2–10 s in length.

Table 4. Configuration of the video bitrates

Quality level	Resolution	Bitrate (kbps)
1	160p	221
2	240p	614
3	360p	1384
4	480p	2462
5	720p	5535
6	1080p	12453

5.2. Performance Metric

We evaluated the rate adaptation scheme with regard to efficiency and stability. In the HAS system, efficiency corresponds to the overall video quality, and stability corresponds to the lack of changes in video quality. To measure the overall video quality, we calculated the average video bitrate for all segments using (9). To evaluate the change in the video quality and the magnitude of the change simultaneously, we defined a metric for the relative difference in the video bitrate, as follows.

$$\frac{1}{N-1}\sum_{n=1}^{N-1}\frac{\left|r_{n+1}-r_n\right|}{\min\left(r_{n+1},r_n\right)} \tag{16}$$

To take into account the variability of the device, video, and network, we performed 15 simulations for each rate adaptation scheme, while the changing buffer capacity and segment duration. For all the network profiles, the buffer capacity was changed to 30, 60, and 120 s, and the segment duration was changed to 2, 4, 6, 8, and 10 s. The simulation results were averaged, the standard deviations were calculated.

5.3. Performance Evaluation

Before comparing the proposed rate adaptation scheme with ABR, PANDA, and BBA, we describe its behavior. The proposed scheme selects the next bitrate according to the buffer occupancy and maintains the buffer occupancy through the playback buffer model. Fig. 9 shows that the proposed scheme maintained a stable buffer occupancy even in a highly variable environment. The proposed scheme tended to select lower bitrates to fill the playback buffer quickly when the buffer capacity was large. Because network profile 3 had increased network variability, the playback buffer model computed a higher value for the average buffer occupancy, owing to the increased CV. Therefore, the proposed scheme behaves conservatively when the network is unstable.

To compare the performance of the rate adaptation scheme, we calculated the average video bitrate and the relative difference in the video bitrate from all the simulation results, as shown in Fig. 10. Bandwidth-based rate adaptation schemes exhibit lower video bitrates and fewer bitrate changes than buffer-based schemes. PANDA exhibited worse performance than ABR when the network bandwidth was insufficient at the beginning of the streaming. This inefficiency indicates that the bandwidth estimation must be swift to catch network changes. Tuning the configuration parameters may solve this problem but should be done on a per-network basis. BBA exhibited a higher average video bitrate than ABR and PANDA, but there were unnecessary bitrate oscillations. Because BBA set thresholds of the buffer occupancy that divided the playback buffer into several areas, it changed the video quality too frequently when the buffer capacity was small. The proposed scheme exhibited a high average video bitrate, similar to BBA, but reduced the number of changes in the video bitrate by using the difference in the buffer occupancy.

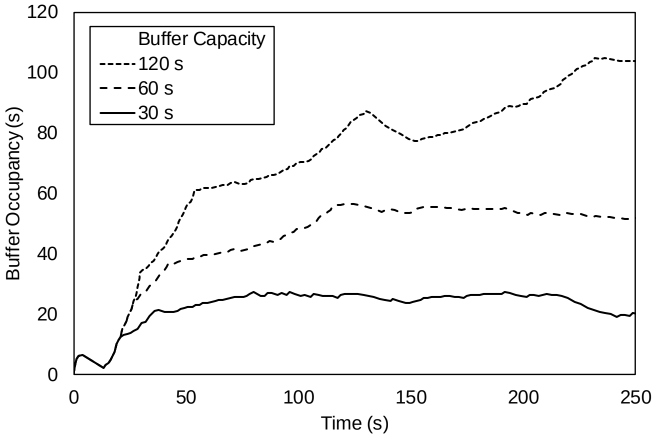
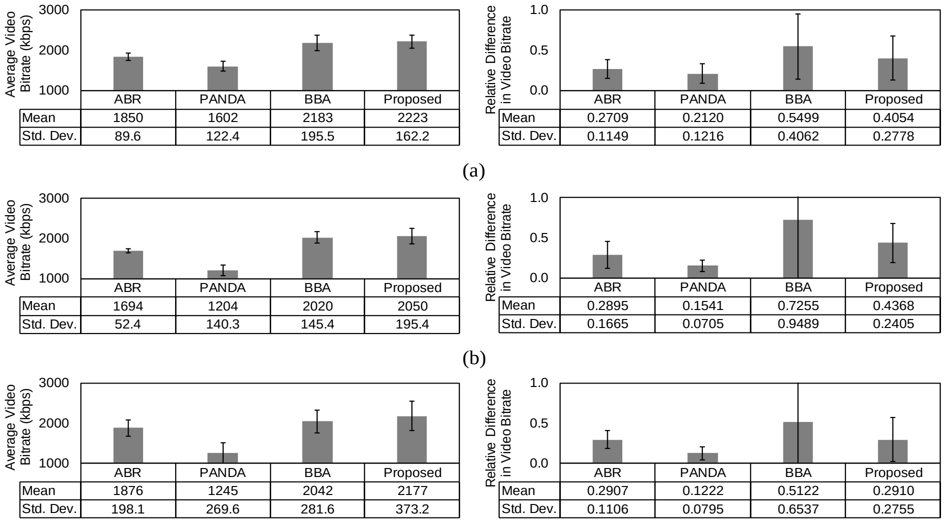


Fig. 9. Buffer occupancy of the proposed scheme in network profile 3



	ABR	PANDA	BBA	Proposed
Mean	1850	1602	2183	2223
Std. Dev.	89.6	122.4	195.5	162.2

	ABR	PANDA	BBA	Proposed
Mean	0.2709	0.2120	0.5499	0.4054
Std. Dev.	0.1149	0.1216	0.4062	0.2778

(a)

	ABR	PANDA	BBA	Proposed
Mean	1694	1204	2020	2050
Std. Dev.	52.4	140.3	145.4	195.4

	ABR	PANDA	BBA	Proposed
Mean	0.2895	0.1541	0.7255	0.4368
Std. Dev.	0.1665	0.0705	0.9489	0.2405

(b)

	ABR	PANDA	BBA	Proposed
Mean	1876	1245	2042	2177
Std. Dev.	198.1	269.6	281.6	373.2

	ABR	PANDA	BBA	Proposed
Mean	0.2907	0.1222	0.5122	0.2910
Std. Dev.	0.1106	0.0795	0.6537	0.2755

(c)

Fig. 10. Average video bitrate and relative difference in the video bitrate for all schemes in (a) network profile 1, (b) network profile 2, (c) network profile 3

Fig. 11 shows the results for all the rate adaptation schemes in network profile 1, where network bandwidth decreased and then increased. All the schemes performed rate adaptation at approximately 100 s owing to the reduction in the available bandwidth or buffer occupancy. ABR and PANDA select the next video bitrate according to the estimated available bandwidth; thus, they responded sensitively to network changes. PANDA performed rate adaptation more conservatively than ABR because of its AIMD-like bandwidth estimation. PANDA exhibited a slow quality improvement at the beginning of the streaming. BBA selects the video bitrate according to the current buffer occupancy and changes video quality when the buffer occupancy exceeds the threshold. BBA frequently changed video quality when the buffer occupancy remained near the threshold. The proposed scheme also selected the video bitrate based on the buffer occupancy but did not change the video quality directly. The playback buffer model computed an expectation of the average buffer occupancy based on indirect information regarding the network, device, and video. Because the proposed bitrate selection method employed this value for rate adaptation, the proposed scheme was able to achieve consistent quality despite the variability.

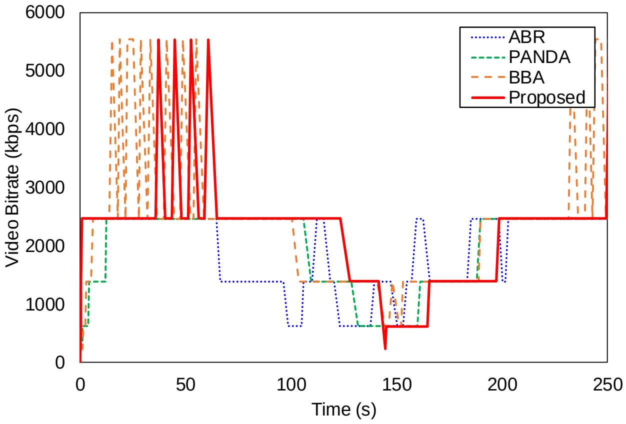


Fig. 11. Rate adaptation in network profile 1

By performing many experiments with various configurations, we observed that there was a tradeoff between the efficiency and stability in HAS. We plotted 12 points representing each rate adaptation scheme in each network profile, as shown in Fig. 12. More points close to the top left of the graph are interpreted as a better rate adaptation scheme. In the experiment, the performance of ABR and PANDA was determined by the accuracy of the bandwidth estimation. BBA could achieve better video quality regardless of the network conditions but made unnecessary quality changes that adversely influenced the user experience. The proposed scheme struck a balance between efficiency and stability and achieved better performance in some cases. The proposed scheme also employed a buffer-based bitrate selection algorithm, but its rate adaptation was performed using the playback buffer model comprising bandwidth, buffer, and video segment information. Therefore, the proposed scheme can provide consistent quality for HAS despite the variability of the network, device, and video.

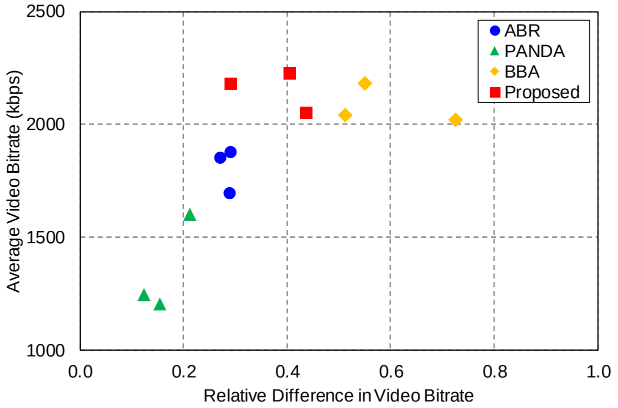


Fig. 12. Comparison of performance in network profiles 1, 2, and 3

6. Conclusion

To investigate the relationship between the bandwidth and the buffer in HAS, we developed a playback buffer model. The playback buffer was modeled based on the queueing theory, which is a proper way to analyze waiting entities. We predicted the average buffer occupancy by exploiting the playback buffer model. Because the average buffer occupancy is determined by the available bandwidth, segment duration, and buffer capacity, we propose a novel bitrate selection algorithm based on the playback buffer model. The proposed scheme sets the average buffer occupancy as a target and thus performs buffer-based rate adaptation for achieving consistent quality despite the variability of the network, device, and video. To evaluate the performance of the rate adaptation scheme, we implemented the HAS system in the ns-3 network simulator and conducted simulations with various configurations. We compared the proposed scheme with well-known rate adaptation algorithms with regard to the average video quality and the change in video quality. The simulation results indicated that the proposed scheme achieves very high video quality on average, even under unstable network conditions. Because the proposed scheme updates the expectation of the average buffer occupancy whenever it receives video segments, it responds to network changes without adjusting any parameters. However, the playback buffer modeling based on queuing theory has a limit in the real-world environments where the network bandwidth, videos watching by users, and number of users are changing more severely than simulation environments. To address this issue, we plan to extend the proposed scheme with a more practical buffer model for the real-world commercial HAS clients as a future work.

Acknowledgements. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. NRF-2020R1F1A1048627). It has also been conducted by the excellent researcher support project of Kwangwoon University in 2021.

References

1. Cisco Public.: Cisco Visual Networking Index: Forecast and Trends, 2017-2022, (2019). [Online]. Available: https://davidellis.ca/wp-content/uploads/2019/05/cisco-vni-feb2019.pdf, last accessed date 2019/5/24
2. T. Stockhammer.: Dynamic Adaptive Streaming over HTTP - Standards and Design Principles. In Proceedings of the Second Annual ACM Conference on Multimedia Systems (MMSys*)*. San Jose CA, USA, 133–144. (2011)
3. Microsoft Azure.: Playback with Azure Media Player, (2019). [Online]. Available: http://www.iis.net/downloads/microsoft/smooth-streaming, last accessed date 2019/7/17
4. Apple Developer.: HTTP Live Streaming (HLS), (2019). [Online]. Available: https://developer.apple.com/streaming, last accessed date 2019/8/12
5. Adobe Live Video Streaming Online.: What is HTTP Dynamic Streaming? (2019). [Online]. Available: http://www.adobe.com/products/hds-dynamic-streaming.html, last accessed date 2019/6/29
6. Multimedia Communication.: HTTP Streaming of MPEG Media, (2012). [Online]. Available: https://multimediacommunication.blogspot.com/2010/05/http-streaming-of-mpeg-media.html, last accessed date 2012/4/26
7. O. Oyman, S. Singh.: Quality of Experience for HTTP Adaptive Streaming Services. IEEE Communications Magazine, Vol. 50, No. 4, 20–27. (2012)
8. J. Kua, G. Armitage, P. Branch.: A Survey of Rate Adaptation Techniques for Dynamic Adaptive Streaming over HTTP. IEEE Communications Surveys & Tutorials, Vol. 19, No. 3, 1842–1866. (2017)
9. S. Akhshabi, L. Anantakrishnan, A. C. Begen, C. Dovrolis.: What Happens When HTTP Adaptive Streaming Players Compete for Bandwidth?. In Proceedings of the 22nd International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV). New York, USA, 9-14. (2012)
10. X. Zhu, Z. Li, R. Pan, J. Gahm, H. Hu.: Fixing Multi-client Oscillations in HTTP-based Adaptive Streaming: A Control Theoretic Approach. In Proceedings of IEEE 15th International Workshop on Multimedia Signal Processing (MMSP). Santa Margherita di Pula, Sardinia, Italy, 230–235. (2013)
11. L. D. Cicco, V. Caldaralo, V. Palmisano, S. Mascolo.: ELASTIC: A Client-side Controller for Dynamic Adaptive Streaming over HTTP (DASH). In Proceedings of the 20th International Packet Video Workshop. San Jose, CA, USA, 1–8. (2013)
12. Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Began, D. Oran.: Probe and Adapt: Adaptation for HTTP Video Streaming at Scale. IEEE Journal on Selected Areas in Communications, Vol. 32, No. 4, 719–733. (2014)
13. T. Huang, N. Handigol, B. Heller, N. McKeown, R. Johari.: Confused, Timid, and Unstable: Picking a Video Streaming Rate is Hard. In Proceedings of the 2012 Internet Measurement Conference (IMC). Boston Massachusetts, USA, 225–238. (2012)
14. M. Allman, V. Paxson, E. Blanton.: TCP Congestion Control, (2009). [Online]. Available: https://tools.ietf.org/html/rfc5681, last accessed date 2020/1/21
15. DASH Industry Forum.: dash.js, (2019). [Online]. Available: https://github.com/Dash-Industry-Forum/dash.js, last accessed date 2020/7/26
16. C. Mueller, S. Lederer, R. Grandl, C. Timmerer.: Oscillation Compensating Dynamic Adaptive Streaming over HTTP. In Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME). Turin, Italy, 1–6. (2015)
17. P. Juluri, V. Tamarapalli, D. Medhi.: SARA: Segment Aware Rate Adaptation Algorithm for Dynamic Adaptive Streaming over HTTP. In Proceedings of the 2015 IEEE International Conference on Communication Workshop (ICCW). London, UK, 1765–1770. (2015)

18. K. Spiteri, R. Urgaonkar, R. K. Sitaraman.: BOLA: Near-optimal Bitrate Adaptation for Online Videos. In Proceedings of the 35th IEEE International Conference on Computer Communications (INFOCOM). San Francisco, CA, USA, 1–9. (2016)
19. R. Huysegems, B. D. Vleeschauwer, T. Wu, W. V. Leekwijck.: SVC-based HTTP Adaptive Streaming. Bell Labs Technical Journal, Vol. 16, No. 4, 25–41. (2012)
20. C. Sieber, T. Hoßfeld, T. Zinner, P. Tran-Gia, C. Timmerer.: Implementation and User-centric Comparison of a Novel Adaptation Logic for DASH with SVC. In Proceedings of the 2013 IFIP/IEEE International Symposium on Integrated Network Management (IM). Ghent, Belgium, 1318–1323. (2013)
21. T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, M. Watson.: A Buffer-based Approach to Rate Adaptation: Evidence from a Large Video Streaming Service. In Proceedings of the 2014 ACM Conference on SIGCOMM (SIGCOMM). Chicago Illinois, USA, 187–198. (2014)
22. Y. Xu, E. Altman, R. El-Azouzi, M. Haddad, S. Elayoubi, T. Jimenez.: Analysis of Buffer Starvation with Application to Objective QoE Optimization of Streaming Services. IEEE Transactions on Multimedia, Vol. 16, No. 3, 813–827. (2014)
23. J. F. C. Kingman.: The Single Server Queue in Heavy Traffic. Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 57, No. 4, 902–904. (1961)
24. NSNAM.: ns-3 Network Simulator, (2019). [Online]. Available: https://www.nsnam.org, last accessed date 2020/10/7
25. DASH Industry Forum.: Guidelines for Implementation: DASH-AVC/264 Test Cases and Vectors, (2014). [Online]. Available: https://dashif.org/docs/DASH-AVC-264-Test-Vectors-v1.0.pdf, last accessed date 2014/3/24
26. Brandee.: Recommended Upload Encoding Settings, (2019). [Online]. Available: https://brandee.edu.vn/glossary/1722171-youtube-en/, last accessed date 2019/11/30

Jiwoo Park received his B.S. and Ph.D. degree from the Electronics and Communications Engineering Department, Kwangwoon University, Seoul, South Korea, in 2009 and 2019, respectively. His research interests include network protocols, multimedia systems, and video communications—in particular, QoS support in adaptive bitrate streaming.

Minsu Kim received his B.S. degree from the Electronics and Communications Engineering Department, Kwangwoon University, Seoul, South Korea, in 2017, where he is currently working toward a Ph.D. degree. His research interests include QoS/QoE support, multimedia systems, and streaming protocols.

Kwangsue Chung received his B.S. degree from Hanyang University, Seoul, South Korea, his M.S. degree from KAIST (Korea Advanced Institute of Science and Technology), Seoul, South Korea, Ph.D. degree from University of Florida, Gainesville, Florida, USA, all from the Electrical Engineering Department. Before joining the Kwangwoon University in 1993, he spent 10 years with the Electronics and Telecommunications Research Institute (ETRI) as a member of the research staff. He was also an adjunct professor at KAIST from 1991 to 1992 and a visiting scholar at the University of California, Irvine from 2003 to 2004. His research interests include communication protocols and networks, QoS mechanisms, and video streaming.

Received: August 20, 2020; Accepted: February 27, 2021.

Computer Science and Information Systems 18(4):1159–1174 https://doi.org/10.2298/CSIS201228044Y

Deep Semi-supervised Learning with Weight Map for Review Helpfulness Prediction

Hua Yin[1,⋆], Zhensheng Hu[1], Yahui Peng[2], Zhijian Wang[1], Guanglong Xu[3], and Yanfang Xu[4]

[1] Information School, Guangdong University of Finance & Economics,
Guangzhou, Guangdong, 510320, China
yinhua@whu.edu.cn
huzhsh6@mail2.sysu.edu.cn
1632646684@qq.com

[2] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou,
Guangdong, 510275, China
13924006758@139.com

[3] School of Statistics and Mathematics, Guangdong University of Finance & Economics,
Guangzhou, Guangdong, 510320, China
guanglongxu2018@gmail.com

[4] School of Art and Design, Guangdong University of Finance & Economics, Guangzhou,
Guangdong, 510320, China
kate20@student.gdufe.edu.cn

Abstract. Helpful online product reviews, which include massive information, have large impacts on customers' purchasing decisions. In most of e-commerce platforms, the helpfulness of reviews are decided by the votes from other customers. Making full use of these reviews with votes has enormous commercial value, especially in product recommendation. It drives researchers to study the technologies about how to evaluate the review helpfulness automatically. Although Deep Neural Network(DNN), learning from the historical reviews and labels, computed by the votes, has demonstrated effective results, it still has suffered insufficient labeled reviews problem. When the helpfulness of a large number of reviews is unknown for lack of votes, or some useful latest reviews with less votes are submerged by the past reviews, the accuracy of current DNN model decreases quickly. Therefore, we propose an end-to-end deep semi-supervised learning model with weight map, which makes full use of the unlabeled reviews. The training process in this model is divided into three stages:obtaining base classifier by less labeled reviews, iteratively applying weight map strategy on large unlabeled reviews to obtain pseudo-labeled reviews, training on above combined reviews to obtain the re-training classifier. Based on this novel model, we develop an algorithm and conduct a series of experiments, on Amazon Review Dataset, from the aspects of the baseline neural network selection and the strategies comparisons, including two labeling and three weighting strategies. The experimental results demonstrate the effectiveness of our method on utilizing the unlabeled data. And our findings show that the model adopted batch labeling strategy and non-linear weight mapping method has achieved the best performance.

Keywords: Semi-supervised learning; Review helpfulness; Pseudo label; Weight map; Labeling strategy.

[⋆] Corresponding author

1. Introduction

Online reviews of products, which are very important to costumers, can provide reference for their purchase decisions [1] [2]. But massive customer reviews on e-commerce websites, including plenty of descriptive, emotional texts with diversified expressions, may lead to information overloading [3]. To highlight the useful reviews, some platforms allow users to vote on their helpfulness. But these votes are imbalanced on some new or unpopular products. The fewer votes on latest products lead to bias errors and are not credible. Therefore, it is necessary to automatically evaluate the review helpfulness[4].

Early studies mainly used hand-crafted features to try to solve the problem of review helpfulness prediction, such as Geneva affect label coder (GALC)[5], linguistic inquiry and word count (LIWC), general inquirer (INQ)[6]. However, due to manual feature engineering and data annotation, using manual features is laborious and expensive. Recently, models built using the convolutional neural networks (CNN)[7] have been applied to review usefulness predictions and showed performance increasing on review helpfulness prediction.

However, most of the current models have poor generalization ability when there is insufficient data. For products with few reviews, it is difficult to obtain enough label data to train effective models using supervised learning methods. In order to alleviate the issues of insufficient labeled data and make full use of adequate unlabeled data, we studies semi-supervised learning for review helpfulness prediction. This paper proposes a deep semi-supervised learning method with weight map for review helpfulness prediction without any hand-craft features and prior knowledge.

The remainder of the paper is organized as follows. Section 2 analysed the related work, Section 3 formally defines the problem and presents our method step by step. Section 4 illustrates the experiment settings including dataset, evaluation standards, experiment design and the experiment results analysis. Finally, Section 5 discusses the conclusions and the future work.

2. Related work

Some pioneering works hypothesize that helpfulness is an internal property of review texts, and try to find new hand-crafted features to study it. Martin used GALC [5] to extract the emotion features from the review texts to build the emotion-based review helpfulness prediction model [8]. Yang leveraged existing linguistic and psychological dictionaries to represent reviews in semantic dimensions [6]. Liu used some argument-based features as the indicators of helpful reviews [9]. However, the performances of these methods depend largely on the hand-crafted features and a mass of manual annotated samples, which are time-consuming and labor-intensive.

Hence, some neural network-based methods have been proposed to solve this problem. Lee and Choeh used multi-layer perceptron neural networks with hand-crafted features [10], similar with the work of Malik and Hussain who used deep neural network with emotion features [11]. Chen and Yang designed a convolutional neural network(CNN) on the raw-text reviews without any hand-crafted features [12]. Saumya used a two-layered convolutional neural network model to predict the best helpful online product review [13]. The models they built showed performance increasing on review helpfulness prediction.

Most of the existing works focus on popular product categories with massive reviews. However, in the case of insufficient data, the model generalization ability is poor. For example, the '*Electronics*' category of the Amazon Review Dataset [14] has more than 1.68 million reviews, while the '*Musical Instruments*' category only has 10k reviews, and most of them have a few votes. For products with a few reviews, it is difficult to obtain enough labeled data to train an effective model with supervised learning method.

Semi-supervised learning [15,16] was prompted to alleviate the issues of insufficient labeled data and make full use of adequate unlabeled data. The most classic and simplest form of semi-supervised learning is self-training[17,18,19,20]. It is an iterative process, which firstly trains a supervised classifier on the labeled data, and utilizes this classifier to label the unlabeled data, then enlarges the training set with the most confident predictions (also named pseudo labels[21]). This method can improve classifier's performance, especially when the labeled training data is obviously scarce[22]. However, classification errors might be accumulated along the process when the pseudo labels are not predicted correctly. The essential problem of self-training is how to make the baseline classifier more precisely and decrease the impact of false predicted reviews in the training process.

This paper proposes a deep semi-supervised learning method with weight map to predict review helpfulness automatically.The contributions of this paper are as follows:

1) The method is a new end-to-end self-training model for review helpfulness predictions, and the performance is considerable well in insufficient labeled reviews situation.
2) It proposes a novel deep semi-supervised learning framework with different labeling and weight mapping strategies,which guides the model to choose more reliable pseudo labels.
3) It develops an algorithm and conducts a series of experiments from the aspects of baseline and strategies comparisons on Amazon Review Dataset. The experimental results demonstrate the effectiveness of our method on ultilizing the unlabeled data.

3. Methodology

In this paper, A deep semi-supervised learning method is proposed for review helpfulness prediction. The flowchart of the proposed method is shown in Figure 1.The procedure is divided into three phases, including: (1)Training base classifier. (2) Weighting unlabeled reviews. (3) Re-training classifier. First, A base classifier is trained on the small labeled dataset by deep neural network. Second,the large unlabeled dataset is predicted by the base classifier for getting the pseudo labeled dataset. A probability selection is applied on this pseudo labeled dataset to get the selected labeled dataset. To balance the instances, a weighting map is proposed and applied on the selected labeled dataset. Then the weighted labeled dataset is combined with the original small labeled dataset to construct the re-train labeled dataset. Finally, the classifier is built on the expanded dataset. The details of each step are described in the following sections.

3.1. Preliminary

Given a small labeled review dataset $D_l = \{(x_i, y_i)|i = 1, 2, \ldots, n\}$ including n reviews, where x_i represents the ith user review and y_i represents the label of this review. The

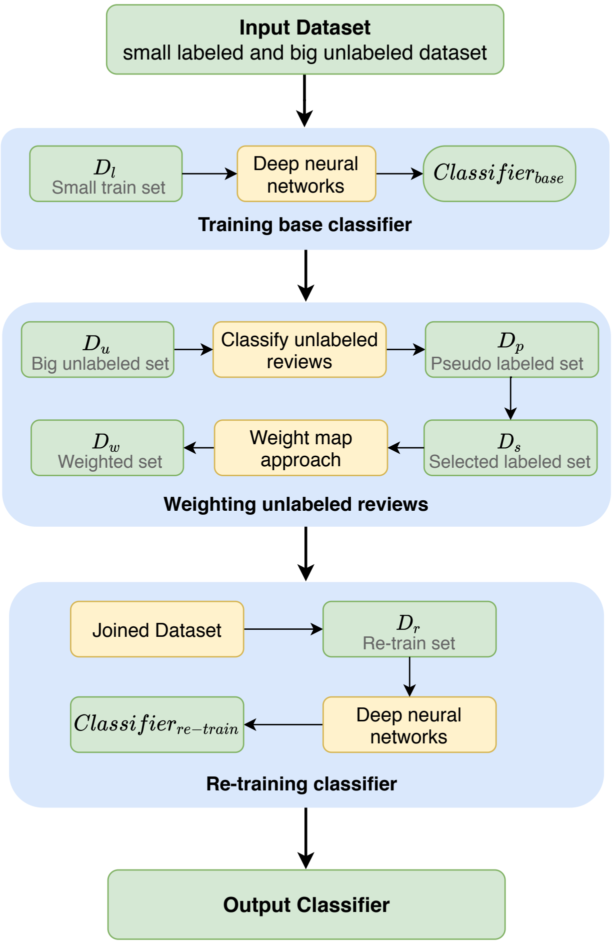


Fig. 1. The flowchart of the method

value of y can only be 0 or 1. When $y_i = 1$, it means that the review is helpful, otherwise it is unhelpful. $D_u = \{(x_j)|j = 1, 2, \ldots, m\}$ is the unlabeled large dataset including m reviews, where $n \ll m$. The review helpfulness prediction problem is defined as a binary classification problem to output a classifier which makes full use of both of the two datasets.

According to the flowchart, the pseudo labeled dataset $D_p = \{(x_j, y'_j, p_j)|j = 1, 2, \ldots, m\}$, where p_j represents the probability of $y'_j(y'_j = 0or1)$, is firstly obtained by the base review classifier built on D_l . Based on p_j, parts of reviews from D_p are selected to get $D_s = \{(x_j, y'_j, p_j)|j = 1, 2, \ldots, m'; m' < m\}$, where $D_s \subset D_p$. The rest reviews of D_p is D'_u. After D_s is processed by weighting map approach, the weight set $D_w = \{(x_j, y'_j, w_j)|j = 1, 2, \ldots, m'; w \in (0, 1]; m' < m\}$ is produced, where w_j is the weight of x_j. Then let $D_r = \{(x_k, y'_k, w_k)|k = 1, 2, \ldots, n + m'; w \in (0, 1]\}$ be the

Re-train set, which combined D_l with D_s. The symbol descriptions are showed in Table 1.

Table 1. Symbol Descriptions

Symbol	Description
D_l	the original small labeled training set with n reviews
D_u	the larger unlabeled training set with m reviews
D_p	pseudo labeled training set on D_u with m reviews
D_s	selected pseudo labeled training set on D_p with m' reviews
D_w	weighted pseudo labeled training set on D_s with m' reviews
D_r	new training set after combined D_l with D_w

3.2. Processing unlabeled reviews

In the training base classifier phrase, the base classifier $Classifier_{base}$ is built by the deep neural networks, not limited to CNN [7], Gated CNN [23,24,25] and RNN [26] to minimize the cross entropy error function on D_l. For the pretraining language models have demonstrated the state-of-the-art performance on a wide range of natural language processing tasks [27], BERT[28] is applied as the word embedding layer. As a key phrase, processing the unlabeled reviews is divided into three steps.

1) Generate the pseudo labeled set

The architecture of the model with deep neural networks are shown in figure 2. Given an unlabeled review x_j in D_u, we predict this review's label y'_j by $classifier_{base}$. As a binary classification problem, the output of the $Classifier_{base}$ is a two-dimensional vector o_{jk}, where $k = 0, 1$,for x_j. After training on D_u, we get an output matrix $Output$. For weighting process, we transfer matrix $Output$ to a probability matrix $Prob_{Output}$ computed by formula (1).

$$Output = \begin{bmatrix} o_{10} & o_{11} \\ o_{j0} & o_{j1} \\ \vdots & \vdots \\ o_{m0} & o_{m1} \end{bmatrix}$$

$$Prob_{Output} = \begin{bmatrix} p_{10} & p_{11} \\ p_{j0} & p_{j1} \\ \vdots & \vdots \\ p_{m0} & p_{m1} \end{bmatrix}$$

$$\begin{aligned} p_{jk} &= p(y'_j = k|x_j) \\ &= \frac{e^{o_{jk}}}{\sum_{i=0}^{i} e^{o_{ji}}} \end{aligned} \tag{1}$$

where $k = 0, 1$ $i = 1$, $\sum p_{jk} = 1$.

Then we obtain the pseudo label y'_j and p_j for each unlabeled review x_j in D_u by formula (2), (3).

$$y'_j = \begin{cases} 1, & if\ p_{jk} > 0.5, k = 1 \\ 0, & if\ p_{jk} \geq 0.5, k = 0 \end{cases} \tag{2}$$

$$p_j = p(p_{jk} | k = y'_j) \tag{3}$$

2) Select the pseudo label set

It is an iteration process to select the reviews from D_p to get D_s. We firstly choose the reviews whose p_j is larger than the mean of the p_j. These reviews construct D_s, and the rest of reviews in D_p construct D'_u,which is used in the next iteration.

$$p_{mean} = \frac{\sum_{j=1}^{m'} p_j}{m'} \tag{4}$$

3) Generate the weighted set

In the following retraining step, we combine the new labeled dataset with the original labeled dataset D_l to produce a new classifier. There are two ways to use the labeled dataset D_s. One is that all the reviews in D_s are treated as the same confidence,and another one is to treat them as different weights. Therefore, we set a weight factor on D_s, and transfer D_s to D_w. w_j is generated by formula (5).

$$w_j = \begin{cases} 1, & \text{No weight} \\ p_j, & \text{Hard weight} \\ f(p_j), & \text{Mapping weight} \end{cases} \tag{5}$$

When w_j is set to 1, D_s is transferred to D_w as following.

$$D_w = \left[\begin{array}{cc|c} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots \\ x_{m'} & y'_{m'} & 1 \end{array} \right]$$

The hard weight utilizes the p_j in D_s. It sets p_j as the weight.

$$D_w = \left[\begin{array}{cc|c} x_1 & y_1 & p_1 \\ x_2 & y_2 & p_2 \\ \vdots & \vdots & \vdots \\ x_{m'} & y'_{m'} & p_{m'} \end{array} \right]$$

We propose a new soft weight mapping approach. It adjusts the weight based on hard weight to a more reasonable way.

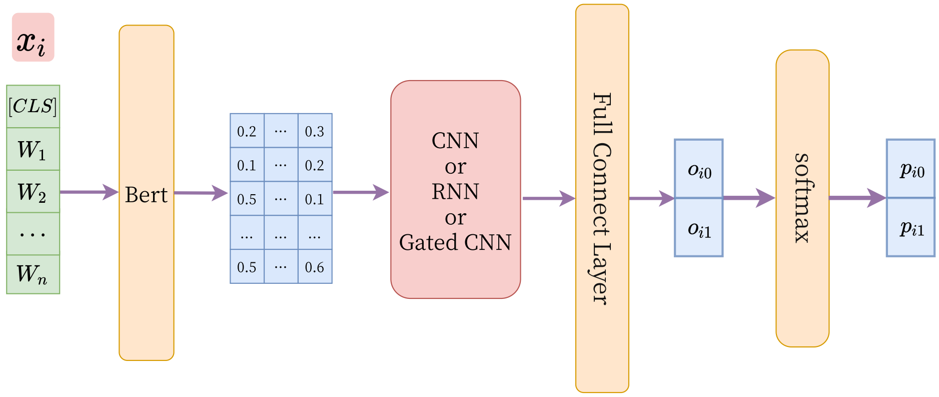


Fig. 2. The architecture of deep neural networks

3.3. Weight mapping approach

For the lower value of p_j, its influence should be lowered to the future classification model. On the other hand, more attention should be paid to the instances which have higher value of p_j. Weight mapping works.

Before being mapped, the range of p_j is $[p_{mean}, p_{max}]$. The confidence of each instance is very close. So the paper tries to map the original p_j to the range of $[0.5, 1]$. Linear mapping and non-linear mapping are two different mapping way to be compared. The linear mapping is done by formula (6). This mapping way doesn't change p_j's distribution density showed as Figure 3. The non-linear mapping is done by formula (7). The aim range is still $[0.5, 1]$. From Figure 3, it is found that non-linear mapping makes the instances diffuse to the side way and changes the original distribution. It makes more instances in the two sides of the new distribution.

$$w_j = \frac{(p_j - p_{mean}) * 0.5}{p_{max} - p_{mean}} + 0.5 \tag{6}$$

$$w_j = \min(\max(p_j(p_j - \frac{p_{max} - p_{mean}}{2} + 1), 0.5), 1) \tag{7}$$

3.4. Re-training classifier

Retraining phrase, shown as figure 4, is a semi-supervised learning process, which iterative utilize the processed unlabeled reviews. It includes three parts:join datasets, retrain classifier and terminate training.

The paper combines D_w with D_l to get a new dataset D_r,which has $n + m'$ reviews. For the reviews in D_l, its weight w_k is set to 1. Then the neural network is trained on the combined dataset D_r and the new classifier $Classifier_{re-train}$ is gotten. L_{ce} is the cross entropy loss function.

$$L(D_r) = L_{ce}(w_k \odot x_k, y_k) \tag{8}$$

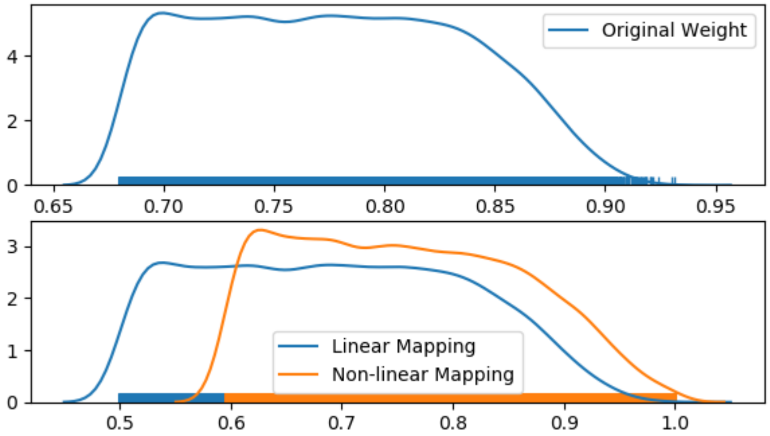


Fig. 3. Linear Mapping and Non-linear Mapping

$$L_{ce}(x_k, y_k) = -[y_k \cdot log(p_k) + (1 - y_k) \cdot log(1 - p_k)] \tag{9}$$

Retraining on the D_r is an iterative process. It will be terminated and output the final classifier when the accuracy of classifier does not increase any more.Or the number of reviews in D'_u is smaller than in D_l, then the training process will be terminated.

4. Experiment

In this part, the detailed experiments,including dataset processing, evaluation metrics, experiment settings and results analysis,are illustrated.

4.1. Dataset

The benchmark dataset is from Amazon Product Review Dataset. It has product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014 [14,29]. The reviews information contains ratings of products, texts of reviews, helpfulness votes and total votes of reviews. The paper chooses Electronics as a representative category to verify the proposed method. This category has the most reviews and is the most frequently used product category in related work.

In order to avoid data bias, the reviews with a total of less than 6 votes are removed, the proportion of helpful votes and unhelpful votes is set to 0.5 [30]. The paper randomly selects some of the helpful reviews to make the training dataset distribution to reach a balanced state, that is, the reviews marked as helpful and unhelpful are both half of the data set. To satisfy with the data setting requirement which the number of D_l is largely smaller than D_u, 1% of the original dataset is selected as the training dataset [20], and 20% of

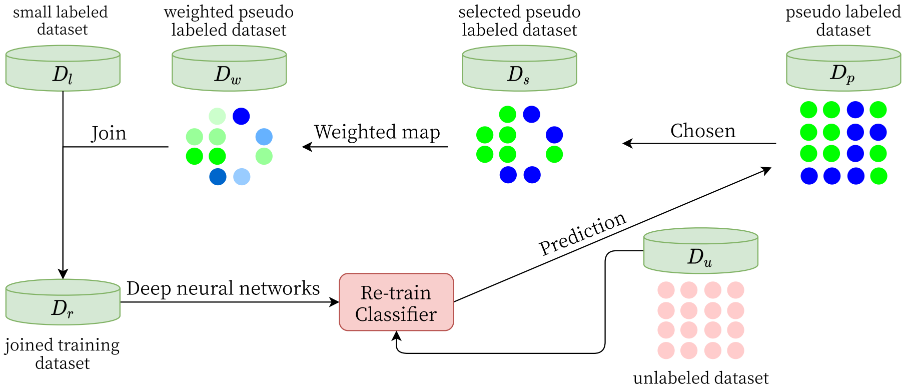


Fig. 4. Process of re-training classifier

the trainning dataset as the development set, and 10 times of the trainning dataset as the test set. The left of the original dataset is the unlabeled dataset. The detailed description is shown in Table 2.

Table 2. Datset Divisions

	Reviews	Tokens(unique)	Tokens
Train Set	500	10670	59378
Dev set	100	3585	11723
Test set	5000	54592	654264
Unlabeled	42934	232729	4789641

4.2. Evaluation metrics

Accuracy, Precision, Recall, F1-Score are chosen as the performance measures for evaluating the classification performance of our approach. The values of these evaluation criteria range from 0 to 1. The larger of these evaluation criteria, the performance of the model better.

Accuracy refers to the proportion of reviews correctly classified by the model. The calculation method is ass:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

F1-score considers both the Precision(P), which refers that the proportion of helpful reviews identifications is actually correct, and the Recall(R), of the test to compute the score, which are defined as:

$$
\begin{aligned}
F_1\text{-}score &= 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \\
Precision &= \frac{TP}{TP + FP} \\
Recall &= \frac{TP}{TP + FN}
\end{aligned}
\tag{11}
$$

Where a true positive (TP) is an outcome where the model correctly predicts the helpful reviews. Similarly, a true negative(TN) is an outcome where the model correctly predicts the unhelpful review. A false positive(FP) is an outcome where the model incorrectly predicts the helpful reviews. And a false negative(FN) is an outcome where the model incorrectly predicts the unhelpful review.

4.3. Experiment setting

To validate the efficiency of the deep semi-supervised learning method, it designs the experiments to answer the three important questions. Has the different deep neural network impacted on the final classifier? Which is the best choice in the three weight mapping approaches? What are the influences of the two labeling strategies?

1) Baseline setting

The paper chooses three typical deep neural network as the baseline network including Convolutional Neural Networks (CNN) [7], Gated CNN [23,24,25] and Recurrent Neural Networks(RNN) [26] , and sets the parameters of these three network to select better-performing network. For the Convolutional Neural Networks, the model consists of one convolutional layer with the 256 channels. The paper adopts multiple sizes of kernels 2, 3, 4, followed by ReLU activation [31]. It sets dropout rate to 0.3 for regularization [32], and concatenates them after every max-pooling layer, then trains the model using AdamW optimizer[33] with 1e-4 learning rate.The setting of model with Gated CNN is same as [23,24], and the model with RNN mainly refers to [26]. The word embedding model used in all experiments in this paper is Bert [28], and the max length of the review text used is 420, covering 95% of the effective review data. All the experiments are conducted on NVIDIA GeForce GTX 2080Ti GPU and implemented using PyTorch.During the training process, due to that the number of training data is very small, the early-stopping strategy [34] is adopted to prevent the model from overfitting.

2) Strategy comparison experiments description

The comparative experiments mainly verify the relevant strategies proposed above. One is a comparison of labeling strategies, and the other is a comparison of different weight mapping strategies.Labeling strategies includes overall labeling and batch labeling strategy. The overall labeling strategy means that the unlabeled set Du is used as a whole for prediction processing, and the remaining ones are filtered out and iteratively predicts the labeling process again. In contrast, the batch labeling strategy is to divide Du into batches in advance into $Du_1, Du_2, \ldots$, and then predict and filter each subset. Experiments will be conducted to analyze the impact of different labeling strategies. The weighting strategies includes three different strategies:no weight, hard weight and mapping weight. The effects of different weight strategies will be compared by experiments.

The table 3 is an explanation of the models and related strategies built by each comparative experiment. The neural network used in all experiments is the one performing well in the above baseline experiment.

Table 3. Comparative experiments description

Experiments	Weight Method	Labeled Method
Exp.1	No weight	Overall labeling
Exp.2	No weight	Batch labeling
Exp.3	Hard weight	Overall labeling
Exp.4	Hard weight	Batch labeling
Exp.5	Liner mapping weight	Overall labeling
Exp.6	Liner mapping weight	Batch labeling
Exp.7	Non-liner mapping weight	Overall labeling
Exp.8	Non-liner mapping weight	Batch labeling

4.4. Results and Analysis

After getting the results of baseline experiments, the better neural network was selected, then we analysed the influences of the two labeling strategies and the final results of the experiments.

1) Baseline experiments results

In order to select a better neural network, Three benchmark experiments have been conducted. The results of these experiments are shown in the following table 4. Based on the results, it can be concluded that the CNN network is generally better than the others from values of F1-score and accuracy.

Table 4. Baseline experiments Results

Model	Precision	Recall	F1-score	Accuracy
CNN	68.12	67.21	**67.66**	**67.33**
Gated CNN	67.51	66.32	66.91	67.10
RNN	63.64	70.02	66.67	65.11

2) Strategies experiments Analysis

When using the overall labeling strategy, in each training loop, the number of pseudo-labeled samples reduce and its F1-score is as follows Figure 5. It can be shown from Figure 5 that when the overall labeling strategy is adopted, the amount of pseudo-labeled samples added is no more than half of the previous loop. During the process of training model, the pseudo labeled samples introduced become less and less. It causes more errors accumulated in the early stage and even leads to the model performance degradation.

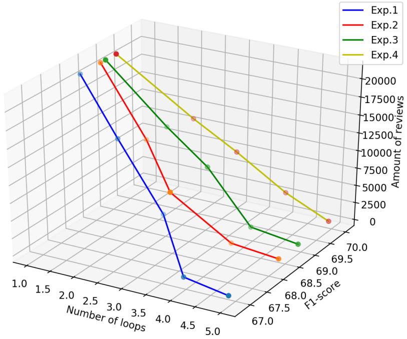


Fig. 5. F1-score and amount of reviews changed by loops on overall labeling strategy

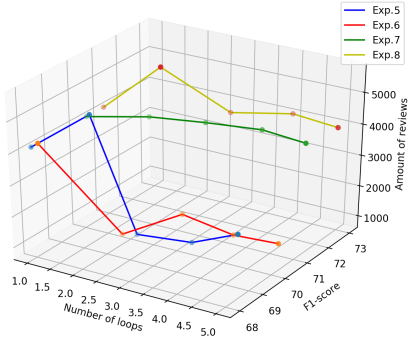


Fig. 6. F1-score and amount of reviews changed by loops on batch labeling strategy

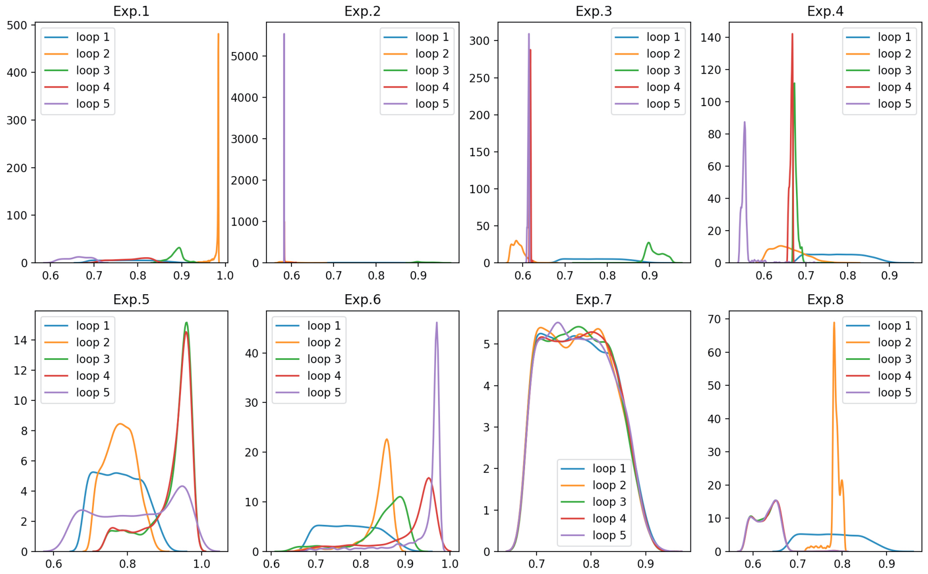


Fig. 7. Distribution density of p_j by loops (Exp.1-8)

When using the batch labeling strategy, the amount of pseudo-labeled reviews in each training loop and its F1-score are shown in the figure 6.

It can be demonstrated from Figure 6 that the experiments that only adopting batch labeling strategy without weight mapping processing will filter out more pseudo-labeled samples in the first loop. In the subsequent loops, the amount of reviews are slightly reduced, but remain stable. At the same time, experiments using batch labeling strategy and weight mapping method can prevent the model from a sharp decrease in the amount of pseudo-labeled training reviews, and make the amount of newly added pseudo-labeled reviews more stable in the entire training process. Thus the model performance can be stably improved.

The distribution and changes of the pseudo-labeled sample probability p_j in each semi-supervised training loop of Exp.1-8 are shown in the figure 7. When overall labeling strategy is used, in the semi-supervised training loop, the amount of pseudo-labeled training set added to the training loop decreases sharply. It results in a very large change in the probability distribution, which is extremely centralized. The stability and generalization of the model are both not enough. When the batch labeling strategy is adopted, the overall performance and improvement of the model are more stable, because the amount of pseudo-labeled samples added is relatively stable, and the model is more robust with a better generalization performance.

3) Analysis of the final results of experiments

The final results of the experiment are shown as table 5. It can be concluded that the model adopted batch labeling strategy and non-linear weight mapping method has the best experimental results. It's F1-score increases by 5.27% and accuracy increases by 4.96%, compared with the baseline model, which demonstrates obvious improvement.

Table 5. Final results of the experiment

Model	Precision	Recall	F1-score	Accuracy
Exp.1	67.82	68.53	68.17	68.06
Exp.2	66.39	71.26	68.73	68.27
Exp.3	67.21	71.01	69.05	67.93
Exp.4	71.25	68.91	70.06	69.42
Exp.5	69.46	67.63	68.53	68.34
Exp.6	69.23	71.80	70.47	70.48
Exp.7	71.12	72.21	71.66	71.21
Exp.8	70.12	76.22	**73.03**	**72.29**

5. Conclusion and future works

This paper studies semi-supervised learning for review helpfulness prediction. It proposes a deep semi-supervised learning method with weight map for review helpfulness prediction without any hand-craft features and prior knowledge. As the experiments demonstrated,the batch labeling strategy can effectively alleviate the problem of the sharp decrease in the pseudo-labeled sample size and make the pseudo-labeled data set flattened in the semi-supervised learning loop, and the weight mapping strategy can effectively improve the model effect, the stability and generalization of the model. In the future work, we will further explore the method of adjusting the pseudo-labeled sample weight in the semi-supervised learning process, and the application of semi-supervised learning in text classification.

Acknowledgments. This work was supported by Science and technology program of Guangzhou (No. 201707010495), the project of philosophy and social science planning of Guangzhou

(No. 2018GZYB77), MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 21YJCZH202) and the major project of basic and applied research in Guangdong universities (No. 2017WZDXM012).

References

1. Shuiqing Yang, Jianrong Yao, Atika Qazi, et al. Does the review deserve more helpfulness when its title resembles the content? locating helpful reviews by text mining. *Information Processing & Management*, 57(2):102179, 2020.
2. Juheng Zhang and Selwyn Piramuthu. Product recommendation with latent review topics. *Information Systems Frontiers*, 20(3):617–625, 2018.
3. Muhammad Shahid Iqbal Malik. Predicting users' review helpfulness: the role of significant review and reviewer characteristics. *Soft Computing*, pages 1–16, 2020.
4. Xiaoru Qu, Zhao Li, Jialin Wang, Zhipeng Zhang, Pengcheng Zou, Junxiao Jiang, Jiaming Huang, Rong Xiao, Ji Zhang, and Jun Gao. Category-aware graph neural networks for improving e-commerce review helpfulness prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2693–2700, 2020.
5. Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.
6. Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. Semantic analysis and helpfulness prediction of text for online product reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 38–44, 2015.
7. Yoon Kim. Convolutional neural networks for sentence classification, 2014.
8. Lionel Martin and Pearl Pu. Prediction of helpful reviews using emotions extraction. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI-14)*, pages 1551–1557, 2014.
9. Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. Using argument-based features to predict and analyse review helpfulness. *arXiv preprint arXiv:1707.07279*, 2017.
10. Sangjae Lee and Joon Yeon Choeh. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, 41(6):3041–3046, 2014.
11. MSI Malik and Ayyaz Hussain. Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, 73:290–302, 2017.
12. Cen Chen, Yinfei Yang, Jun Zhou, and etc. Li. Cross-domain review helpfulness prediction based on convolutional neural networks with auxiliary domain discriminators. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 602–607. Association for Computational Linguistics, June 2018.
13. Sunil Saumya, Jyoti Prakash Singh, and Yogesh K. Dwivedi. Predicting the helpfulness score of online reviews using convolutional neural network. *SOFT COMPUTING*, 24(15, SI):10989–11005, AUG 2020.
14. Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.
15. Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.
16. Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

17. David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.
18. Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42(2):245–284, 2015.
19. Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in neural information processing systems*, pages 3235–3246, 2018.
20. Hwiyeol Jo and Ceyda Cinarel. Delta-training: Simple semi-supervised text classification using pretrained word embeddings. *arXiv preprint arXiv:1901.07651*, 2019.
21. Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pages 1–6, 2013.
22. Heereen Shim, Stijn Luca, Dietwig Lowet, and Bart Vanrumste. Data augmentation and semi-supervised learning for deep neural networks-based text classifier. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, page 1119–1126, New York, NY, USA, 2020. Association for Computing Machinery.
23. Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941, 2017.
24. Cen Chen, Minghui Qiu, Yinfei Yang, Jun Zhou, Jun Huang, Xiaolong Li, and Forrest Bao. Review helpfulness prediction with embedding-gated cnn. *arXiv preprint arXiv:1808.09896*, 2018.
25. Cen Chen, Minghui Qiu, Yinfei Yang, Jun Zhou, Jun Huang, Xiaolong Li, and Forrest Sheng Bao. Multi-domain gated cnn for review helpfulness prediction. In *The World Wide Web Conference*, pages 2630–2636, 2019.
26. Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*, 2016.
27. Zhensheng Hu, Hua Yin, Guanglong Xu, Yi Zhai, Danbei Pan, and Yongkang Liang. An empirical study on joint entities-relations extraction of chinese text based on bert. In *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, pages 473–478, 2020.
28. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
29. Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.
30. Xianshan Qu, Xiaopeng Li, and John R Rose. Review helpfulness assessment based on convolutional neural network. *arXiv preprint arXiv:1808.09016*, 2018.
31. Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, 2000.
32. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
33. Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ICLR*, 2018.
34. Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

Hua Yin born in 1981. Associate Professor and Master supervisor in the Information School,Guangdong University of Finance and Economics. Her main research interests include machine learning, deep learning and natural language processing.

Zhensheng Hu born in 1991. Master student in the Information School, Guangdong University of Finance and Economics. His main research interests include deep learning and natural language processing.

Yahui Peng born in 1980. Ph.D.candidate in the School of Electronics and Information Technology, Sun Yat-sen University. His main research interests include data mining and medical image processing.

Zhijian Wang born in 1970. Professor and Master supervisor in the Information School, Guangdong University of Finance and Economics. His main research interestis business intelligence.

Guanglong Xu born in 1993. Master Student in the School of Statistics and Mathematics, Guangdong University of Finance and Economics. His main research interests include machine learning and natural language processing.

Yanfang Xu born in 1993. Master Student in the School of Art and Design, Guangdong University of Finance and Economics. Her main research interests include art design theory and data analysis and visualization.

Received: December 28, 2020; Accepted: May 05, 2021.

https://doi.org/10.2298/CSIS201122022V

Cooperation and Sharing of Caught Prey in Competitive Continuous Coevolution Using the Predator-Prey Domain

Krisztián Varga[1] and Attila Kiss[2]

[1] Eötvös Loránd University
Budapest, Hungary
scout@inf.elte.hu
[2] Eötvös Loránd University department of Information Systems
Budapest, Hungary
kiss@inf.elte.hu

Abstract. Competition is one of the main driving factors of evolution and can be observed in nature as well as in simulations. Competition can occur between predators and preys causing an arms race, but it can also happen between individuals of the same species. Our simulation uses the predator-prey domain (with carnivores, herbivores and plants,) and continuous (not generation based) neuro-evolution to create a complex environment where both forms of competition arise. The characteristics of the simulation make it hard for a predator to catch prey alone, this creates a dependence on cooperation. The predators can share the caught prey with nearby members, in order to help them work together. We explore how sharing affects the cooperation of the hunters, and compare the effectiveness of one and two predator populations.

Keywords: artificial life, neural networks, evolution, coevolution, cooperation, arms race, competition, predator, prey.

1. Introduction

Evolution is a critical part of nature that makes complex behaviours and higher level lifeforms possible. A lot of reasons can lead to evolution including environmental changes [11], competing species [8], and competition between individuals [3]. A relevant example today for the first one is the ongoing climate change, which has already shown its effects in nature: a 2 degrees Celsius increase in spring temperature over 10 years caused a red squirrel population in the southwest Yukon, Canada, to have an 18 days advancement in the timing of breeding [22]. It has been shown that competitive arms races arise in nature (bats and moths [12], whelk and Mercenaria [4]) as well as in simulated environments [20] [23]. It is important to study these processes to better understand one of the main driving factors of evolution.

We were interested in the performance and behaviour of simple agents that have limited knowledge about their local surroundings in a competitive and changing environment, where cooperation is key for survival. Studies confirm [5] [14] that even microbiological lifeforms usually cooperate by releasing enzymes into the common space, which helps everyone to get more nutritions. But can such low level lifeforms really cooperate? Marshall [17] who studied the “wolf-pack” hunting of myxobacteria disagrees. He argues that they do not necessary cooperate with each other, because more bacteria indeed helps killing prey more effectively, but the released nutritions must be shared. Since such low lifeforms have no concept of teamwork and can only sense other individuals near them, the “cooperating” behaviour might only come from the loose definition of cooperation. The experiment proposed here is similar to the strategy of these microscopic predators, but instead of chemical warfare, they have to catch prey by surrounding it.

There are a number of hypothesises we want to examine. The first one is that arms race can be sustained in our continuous simulation without distinct generations. The second is that two smaller hunter populations will be more effective at catching prey, because they can specialize two different roles to capture prey and there is a possibility of arms race between the two populations. The last hypothesis is that a higher sharing percentage will increase cooperation between hunters.

2. Related Work

2.1. Evolutionary Computation

In computer science evolutionary computation refers to a family of algorithms that are used for solving global optimisation problems. They start with a population of solutions, which are usually randomly generated, and evolve them based on natural selection, crossovers and mutations. Natural selection means survival of the fittest, therefore there is often a fitness function that evaluates every solution and then the best performing ones will be selected for crossing or mutation. Crossing two or more population members means that the result will contain parts of all of the parent members, this helps to faster converge to the optimal solution. Mutation is also an important part of the process, by having a low chance for altering small parts of members we can ensure that we do not get stuck in a local optima. Evolutionary computation have led to interesting discoveries and showed solutions that exceeded the researchers expectations [15].

2.2. Neuro-evolution

Neuro-evolution is a subset of the evolutionary algorithms. It is a form of artificial intelligence, which focuses on training artificial neural networks. The training can involve the weights of the network, the structure of the network or both, in this paper we are using the first one. Artificial neural networks are often trained by a form of stochastic gradient descent algorithm, but one of the greatest advantages of neuro-evolution over them is that it does not need training sets of correct input-output pairs. It is a powerful tool for solving problems, where the system or environment is highly complex, for example controlling autonomous agents [2], rockets [6] or freeway traffic [27], which is exactly what we need.

2.3. Coevolution

Coevolution in evolutionary computation means that multiple agents belonging to two or more populations are evolving while interacting with each other and this is represented in their fitness evaluations. Coevolution can be cooperative, competitive or both at the same time. In cooperative coevolution [16] agents work together to achive a common goal and they share a common fitness evaluation. In competitive coevolution [13] [24] agents are working against each other and one's gain in fitness means loss to it opponent(s). A good example of competitive and cooperative coevolution is Balch's work [1] where he trained robot soccer teams with reinforcement learning. He showed that global reinforcement for the teams results in better performance and more heterogeneous behaviour, because an agent is not punished even if he did not contribute as much to the overall fitness. Teams evolved with local reinforcement show homogeneous behaviour, because they are competitive against their teammates. Another study [21], where behaviour of spotted hyenas was modeled using coevolution, showed that communication and reward sharing increased cooperation in a simple simulated environment. In this paper we wanted to achieve a mix of global and local reinforcement for hunters by introducing sharing of caught prey. This way each agent will have their own fitness value, but they can benefit by working together. Another important aspect of coevolution is that it often leads to arms races between the populations and evolutionary trade-offs can also be discovered [8]. Evolutionary trade-offs are adaptations, where one feature or behaviour of an agent becomes better while other areas experience decreased performance.

2.4. Predator-Prey Domain

Predator and prey systems have been extensively studied by many scientific fields, including biology [25], mathematics [30] and computer science [7] [13]. In predator prey systems there are usually two populations where the predators' goal is to eat preys. Different aspects of this system can be analysed depending on the simulation or the real life ecosystem. One of the most studied concept is coevolution between predators and preys [13], because the model fits perfectly with its two distinct subpopulations, where one's gain is the other's loss.

Our study was inspired by a simulation [23] that was used to create competitive and cooperative behaviours using coevolution. In this study they used 3 predators and 2 preys and they managed to keep up an arms race between cornering and fleeing strategies. For the neuro-evolution a multi agent ESP [29] architecture was used, where the hidden neurons inside a network come from different sub-populations and each network inside one agent contains its own neuron sub-populations. They used multiple networks to keep track of all of the enemies' coordinates and combined them by summing their output neuron values. With this structure they created multilevel cooperation and competition. Cooperation between the hidden neurons in a network, cooperation between the neural networks inside agents, cooperation between agents of the same kind, and competition between hunters and preys.

3. Rules and Mechanisms of the Simulation

We wanted to scale up the number of agents in the simulation and to run it continuously, without distinct generations, but still benefit from evolution. The original architecture was not suited for our needs, because every agent had global knowledge of all of their enemies' positions and by scaling up to hundreds of agents one individual's movement would require hundreds of neural networks to cooperate. The spatial environment was a good starting point, because spatial coevolution proved to be effective [18], [19], [26] in overcoming the main arising problems in coevolution: loss of gradients, (which means that one population is either too weak or too strong for the other for meaningful change to occur), over-specialization, (which means that the evolutionary process gets stuck in a local optimum), and red queen dynamics, (which means the populations continue to change, but these changes do not force more general solutions). To avoid extinction we used populations with fixed sizes. This means that whenever an agent dies, it is immediately replaced by a new one and the same happens to plants. This does not mean that there are infinite resources, because in practice plants have to be found first, which takes time. The new agent will be constructed from multiple currently alive agents. These agents are chosen from the top performing individuals in the population, which means that they are in the top 10% based on their fitness values. (The selection process is discussed in detail in section 3.5.) The 10% threshold was chosen, because after experimenting with different values, we found this to be high enough that not only a few lucky agents will be considered and low enough to ensure that only competitive genes will be passed on. This parameter could have been set to any value, but this is not in the scope of our study. This system combined with local interactions will keep the simulation running forever and ensure that rapid as well as slow evolution will take place. Since we are using agents with no memory and there is no "Hall of Fame" to preserve the best networks of all time, the evolution only reacts to the current environment. The complexity of the environment makes it highly unlikely to reach a state that was visited before, which prevents the simulation to be stuck in a periodic repetition of states and strategies.

To experiment and to draw conclusions a suitable environment is needed. We built the simulation around the predator-prey domain, which has been studied by the scientific community for over decades. Multiple studies confirm, that arms race between predators and preys can be observed in nature and that this can be reproduced in an artificial environment. Each individual in the simulation is controlled by 3 different artificial neural networks. Fitness based selection and random mutations ensure the continuous evolution.

3.1. Playground and Walls

The artificial world consists of a 2 dimensional rectangle, where the size of each dimension can be configured. There are maps that can be loaded into the simulation. A map is a collection of walls, the simplest map consists of 4 walls around the rectangle. A wall has two points, these points' coordinates are defined on the unit square. When a map is loaded all the walls' coordinates will be multiplied by the worlds dimensions, this way maps are not tied to sizes. Walls do not move and do not have thickness. If a creature comes into contact with a wall, then it dies. Walls represent static danger to all agents, it does not change throughout the simulation contrary to dynamic danger for the preys, the moving and evolving carnivores. It adds complexity at the neural networks' level, where walls have to be considered when fleeing as well as catching prey. They can be interpreted as uncrossable environmental features, like deep canyons, fast flowing rivers or in the microscopical sense a patch of toxins. Walls create a structure to the map and separate small ecosystems. They can provide shelter against predators too, if prey can stay in close proximity most hunters will not risk bumping into a wall or scaring the prey into killing itself. This alone can create an arms race, where preys and predators try to get more and more closer to the walls.

3.2. Plants

Plants are represented as disks on the map. If a prey comes into contact with a plant, then the plant will be eaten. There is always a constant amount of plants, because when one is consumed it reappears on a random location. Predators have no effect on plants and their vision is not blocked by them in order to help them find prey more easily.

3.3. Predators and Preys

Predators and preys are very similar. Both of them have a square as shape, and both of them have 5 sensors they can use to navigate. The sensors are rays starting from the creature and going to 5 different directions. Every creature can move in 8 different directions on the plane: N, E, W, S, NE, NW, SE, SW, (N = North, E = East, W = West, S = South). They have an orientation in the direction they are moving and can only turn left, right or keep going straight. The 5 sensor rays follow these directions too, one goes where the creature is facing and 2-2 go left and right relative to that. A sensor tells the creature that in the relevant direction what kind of object can be found and how far it is. For preys the objects can be walls, plants, predators or nothing if there is nothing inside the view range. For predators the objects can be walls, preys, other predators or nothing. There is always a constant amount of both types of agents, because when a creature dies it is replaced with a new one immediately.

3.4. The Brain

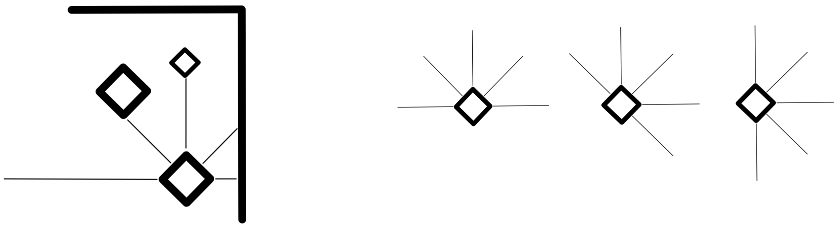
All predators and preys have a brain, which they use to decide which way they should move based on the sensor inputs. The brain is responsible for finding food and escaping static or dynamic obstacles, to accomplish this it is divided into 3 different artificial neural networks. The “food-network” is responsible for locating food, the “wall-network” avoids walls and the “predator-network” knows where predators are. For predators the food means prey, but for the preys the food means plants and another important distinction is that for the prey, predators mean danger, but for predators they are necessary companions to catch prey. The networks have the same architecture, because all of them are used the same way. The input neurons get the data from the sensor rays and the 3 outputs are the directional changes (turn left/right, keep going straight). The architecture consist of 5 input neurons, 7 hidden neurons and 3 output neurons with sigmoid activation function. Only one hidden layer was chosen, because the task is simple and more layers would have added unnecessary complexity. Since the environment needs to be optimized first (choosing default values) to be able to properly evaluate our architecture and there is no training data, we couldn’t use pruning and constructing algorithms [28] [9]. There is no danger of overfitting our network, because of the lack of a training dataset and the environment will change as evolution progresses, but too few neurons could restrict our agents abilities. There is no standard way of determining the network architecture in this case, but there are some rule of thumbs that we can use [10]:

- “The number of hidden neurons should be between the size of the input layer and the size of the output layer.”

- "The number of hidden neurons should be 2/3 the size of the input layer, plus the size of the output layer."
- "The number of hidden neurons should be less than twice the size of the input layer."

These rules work to some extent, but every task is different and the chosen architectures have to be tested. We choose the highest number of neurons where while using 100 plants and 300 agents the simulation still runs smooth on the test machine. Although our decision makes output calculations more resource intensive, this way our agents will not be hindered by too few neurons. After the environment was balanced using this architecture, the configuration performed well during testing and allowed complex behaviours to emerge. Finding the most optimal number of hidden neurons will be a future research.

The different networks inside a single agent only get the relevant inputs for them (for example the wall-network is only concerned about detected walls). The value for a given input neuron is 0 if in the given direction there is no relevant object or $v - d$ where v represents the maximum view range and d represents the distance of the detected object. The 3 outputs represent the 3 decisions the creature can make: turn left, right, or keep going straight. The outputs of the 3 different networks is summed together and then the decision with the maximum summed weight will be selected. This creates an internal cooperation between the networks, because they are all necessary to survive and they share the host's fitness. (The fitness function is discussed in the next section.) The outputs could have been combined other ways, for example with another neural network, but this would add additional complexity to our simple agents and this method already proved to be sufficient [23].



(a) Detection of objects with rays

(b) Behaviour of rays when turning right two times

Fig. 1. Ray mechanics

3.5. Selection and Mutations

Selection occurs within each population. Since the number of agents is constant in the populations, when a creature dies another takes its place. This new agent will have 3 different neural networks and a starting position, all of these features come from other agents. The donor agents come from the top 10 percent of the population based on their fitness value at the time the replacement occurs. The randomly selected individuals from the elite fitness group are not necessary different, for example 2 or 3 networks can be inherited from the same agent. The new position will be the average of the position of an elite agent and a random coordinate, the average will be weighted 2 to 1, this way the new agent will spawn in the near of the selected high performing teammate. This will ensure that reproduction will also happen locally, but with enough variety to not get stuck in one place.

The fitness value is calculated by how long an agent has been alive combined with how many plants or preys they have consumed. An agent starts out with a predefined amount of health. Every timestep, (the smallest time interval in the simulation or in other words one update of the artificial world,) the health declines, and if it reaches zero then the agent dies. For every timestep it survives, it gets 1 fitness point. Hunters and preys can extend their lifespan by consuming food, for preys every plant eaten will give them a predefined amount of fitness and health points, for predators every caught prey gives them a fitness score boost based on the fitness of the prey. If they catch one with a high fitness score, then they will get more fitness and health themselves for defeating a successful opponent. This is called "competitive fitness sharing" [24] and it helps avoiding stagnation.

3.6. Parameters

The simulation can be configured using a configuration file with many different parameters:

- screen_size_x: size of the x axis in pixels
- screen_size_y: size of the y axis in pixels
- updates_per_second: Simulation updates per second
- food_amount: Amount of plants on the field
- herbivore_amount: Amount of preys on the field
- carnivore_amount_1: Amount of hunters in the first group on the field
- carnivore_amount_2: Amount of hunters in the second group on the field
- map: Map file to load
- seed: Random seed, so the simulations can be repeated
- herbivore_speed: Speed of the preys
- carnivore_speed: Speed of the hunters
- initial_herbivore_health: Starting health when spawning new prey
- initial_carnivore_health: Starting health when spawning new hunter
- food_nutrition: Nutrition (health and fitness) the prey gets when it consumes a plant
- herbivore_nutrition: Base nutrition value (health and fitness) the hunter gets when it consumes a prey
- threshold_herbivore_score: The threshold fitness score for caught prey, in case of a successful hunt the nutrition the hunter gets depends on the preys fitness (new fitness = old fitness + prey nutrition * prey fitness / threshold fitness)
- sharing_percentage_1: Amount of food a hunter in the first group is willing to share if other hunters are nearby (in percentage)
- sharing_percentage_2: Amount of food a hunter in the second group is willing to share if other hunters are nearby (in percentage)
- share_range: The sharing range in pixels from a hunter that caught prey, other hunters in this range will get an equal share from the prey (the part that the hunter was willing to share will be divided equally between the others)

- mutation_rate: Probability of mutation
- herbivore_size: Size of the preys in pixels
- carnivore_size: Size of the hunters
- thinking_time: Number updates after the agents make their next decision
- view_range: Length of the sensor rays in pixels
- start_recording: Elapsed time from the start of the simulation to start recording (in minutes)
- recording_duration: Elapsed time from the start of the recording to end recording (in minutes)
- record_all_details: Recording all details of the simulation or just the events and averages (true or false)

The planning for the default simulation values started with the map. We decided that the size should be 800 by 800 pixels, because this fits on most modern monitors regardless of orientation. The map can be seen in *Figure* 2. This configuration was chosen, because there are vast spaces as well as corners where prey can hide. The prey's diameter is half the size of the hunter's, this helps prey to escape through small paths between hunters and makes hunters easier to detect. The hunters benefit from their increased size, because they can cover more area of the map. The diameters were set to be 40 and 80 pixels and the initial plan was to have 100 plants, prey and hunters in the simulation, because a population of 100 individuals is large enough for the algorithm to not get stuck in a local optima. The population of hunters had to be reduced to 90, because they took up too much area of the map and blocked advanced strategies from emerging. The prey are slightly faster than the hunters (0.8 vs 1.2 pixels/update), in order to make them hard to catch for one hunter alone and still make it possible using teamwork. At first giving the agents a lot of initial health seemed to be a good idea to make them last longer and to have enough time to find food. This worked well when there were only prey and plants on the map, but when hunters were introduced their behaviour changed. The prey were hiding from the predators and not catching any plants, making it harder for both populations to evolve. By reducing their health to 1000 and setting the reward for one plant to 200 they tended to seek food even when danger was introduced. On the other hand hunters needed as much initial help as they could get, therefore their health was set to 5000. More health equals more time to find teammates and hunt together, but if we would set it any higher they wouldn't explore the map, because it is more beneficial to them to just wait out their lifetime than to accidentally die early by colliding with a wall. The base reward for catching prey is set to 200, but the final reward depends on the fitness value of the caught agent. An agent is considered average, if it reaches 750 fitness, this means that it survived three quarters of its initial lifespan or that it managed to eat some plants. The exact reward value comes from the formula: agent's_fitness_value / 750 * 200. This ensures that killing a skillful individual is rewarded and that agents who are results of bad mutations or crossovers do not feed the hunters too much. There are two more important parameters left, the view range and the share range. The view range was not limited at first, but after a short time of running the simulation this resulted in reduced movement, by decreasing its value the agent were more keen to explore the map. With 190 agents on a map with 640000 pixels, there is approximately a 60 by 60 pixels square for each one, therefore a view range of 50 pixels is sufficient for them to be aware of their surroundings. The sharing range comes from the view range, an agent should only get a share of the price if it contributed to the hunting process, which means that it was close when the prey died, therefore the sharing range should be less than then view range. When the sharing range was too low (10-30 pixels) the hunters did not have enough room to corner prey and still be close enough to get a share, in the end 40 pixels showed the best results.

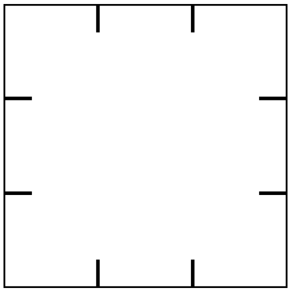


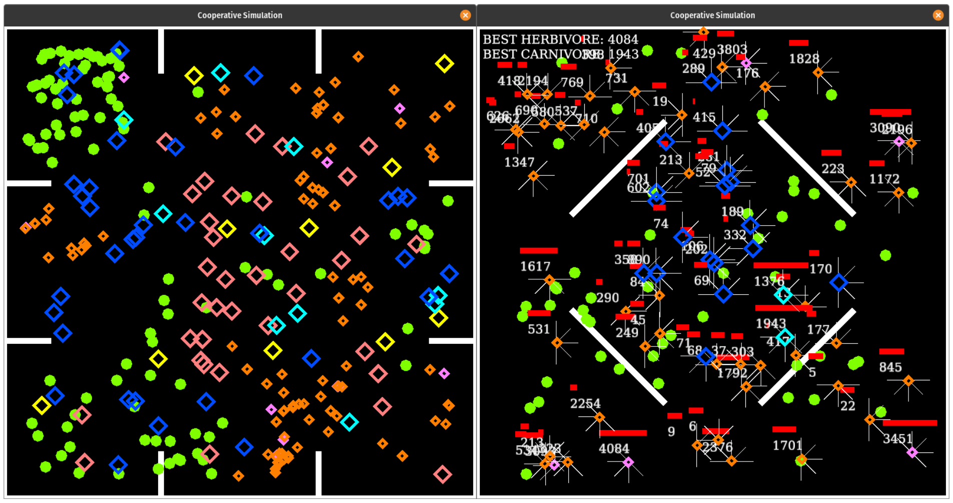
Fig. 2. The map used for the simulations

4. Experiments

The simulation was implemented in Rust, a fast and high level language that has many safety guards for memory management and concurrent programming. The code is open source and it is published on github[3]. The program does not have huge system requirements, but the performance will depend on the number of agents used. The experiments were carried out on a 6 core 3.6GHz processor. If simulation data is being recorded, we advise using an SSD.

There are a lot of parameters in this simulation that can be adjusted, therefore we had to come up with an experimenting plan. After the default values have been chosen for the parameters that we will use for all simulation runs, (more details in the *Parameters* section,) we focused on two main factors: the number of predator populations (1 or 2) and sharing percentages (0, 25, 50, 75 or 100). When there are two predator populations we can also test different sharing percentages for them (0-100, 20-80, 40-60). This gives us $5 * 2 + 3 = 13$ kinds of simulation runs. All simulations are run for 8 hours (or approximately 700000 simulation updates). The results shown here are individual runs, not averaged over multiple simulations from the same kind, because each run is different with huge spikes in the graphs at different places and we do not have the necessary time and resources to run enough simulations to smooth the graphs out.

[3] https://github.com/kvarga-research/Competitive-and-Cooperative-Evolution



(a) Agents inner state is hidden

(b) Agents inner state is visible

Fig. 3. Screenshots of the simulation. The visibility of the agents' inner state and detection rays can be toggled. The smaller squares are preys, the bigger ones are predators. The first hunter group is red and their elite agents are yellow. The second hunter group is blue and their elite agents are cyan. The elite preys are purple.

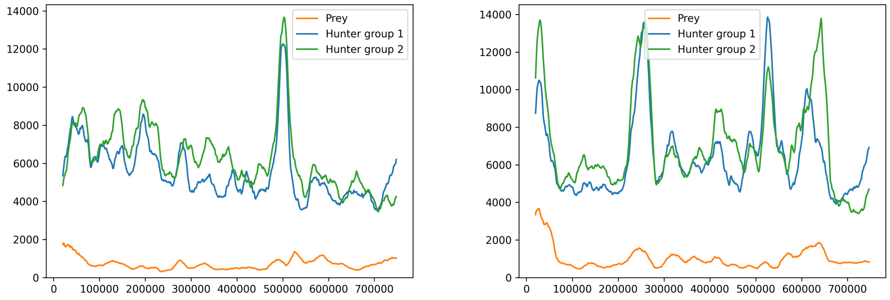
4.1. Results

The 13 runs with different configurations will be discussed in this section. To avoid repetition of words, we introduce a naming system for the different configurations. 'PNSM' means that N population(s) was/were used with $M\%$ sharing. For example P2S50 means two populations with 50% sharing. If there are different sharing percentages, then the additional percentage will be added to the end: P2S20S80. All graphs were smoothed out with rolling average over 20000 timesteps.

The results for P2S25, P2S50 and P2S75 can be seen in *Figure* 5. The graphs on the right side show the cooperation between hunters and the first thing we can notice is that hunting involving only one individual is always the smallest and mostly 2 to 3 sized hunting groups dominate. Its interesting to see that the performance of the two hunter groups are very close throughout the entire run in P2S25 and P2S50, the green and blue lines follow each other closely, this means that they have been working and moving together, reacting the same way to the environmental and behavioural changes. If we take a look at *Figure* 4, we can see that there is an arms race between the two hunter populations. because their average fitness scores are taking turns at having the most prey captured. Looking at the P2S75 run we can see that green group is clearly more successful than the blue one, but the spikes in the graph are at the same places, this most likely means that the two teams are cooperating in a way: they form huge packs (which is beneficial when there is 75% sharing), but the green team is the one actively catching preys meanwhile the blue team might be blocking fleeing paths, or just simply having a parasitic relationship with the greens. The cooperation graph confirms this theory, because the “equal or greater than 5” sized groups are massively dominating. These are most likely much bigger groups than 5, because if they were near 5, then statistically the 4 sized groups’ graph should be closer to it. The plant eating is much lower than the prey catching throughout the simulation, but this does not mean that they are suppressed, this only tells us that eating plants is not worth it that much. Another interesting phenomenon that we observed in most of the simulations is that since prey are running from hunters, there are areas where hunters are dominating, therefore at such places the spawned plants will not be eaten. This leaves much fewer plants for other areas and when hunters move after prey, they free up plant-rich fields, which explains periodic bumps in plant eating and prey capture.

The results for P1S25, P1S50 and P1S75 is shown in *Figure* 6. The main difference compared to the “two population” version is that the overall hunter performance is worse. The performance also declines the more sharing is introduced, this is most apparent with 75% when the average preys captured are a third of the multi population counterpart. Such low number of caught prey is caused by the hunters forming big “wolf packs” with lot of agents on a very small area. These high density groups leave a lot of free space for preys to roam around and make it relatively easy to evade them. With such good conditions for preys we could expect the plant eating to skyrocket, but this is not the case, because as mentioned before, these hunter rich areas are hogging plant resources. Preys might seem poorly performing looking at the average plants eaten, but keep in mind that they have to balance eating and avoiding predators. When predators are truly dominating, the number of caught prey increases meanwhile the plant consumption drops to almost zero, because preys are eaten so fast they have no chance of finding plants. A spectacular example can be seen of this with 50% sharing between the 350000th and 400000th timesteps.

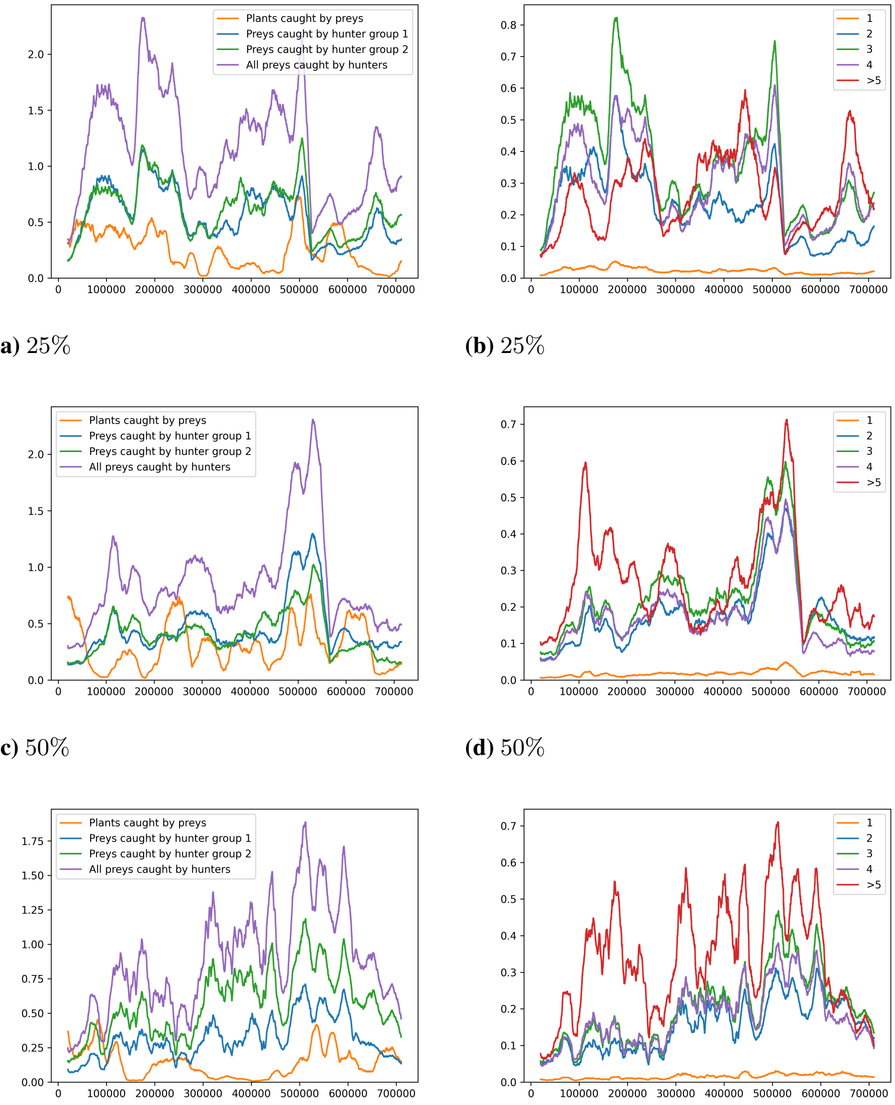
The cooperating diagrams are mostly dominated by hunter groups of size 2 or 3, but size 4 and 5 are not far behind and successful solo hunting remains almost nonexistent. These ratios are not changing much during simulations and they do not seem to be affected by different sharing levels.



(a) 25%

(b) 50%

Fig. 4. Average fitness of the top 10% with 25% and 50% sharing



(a) 25%

(b) 25%

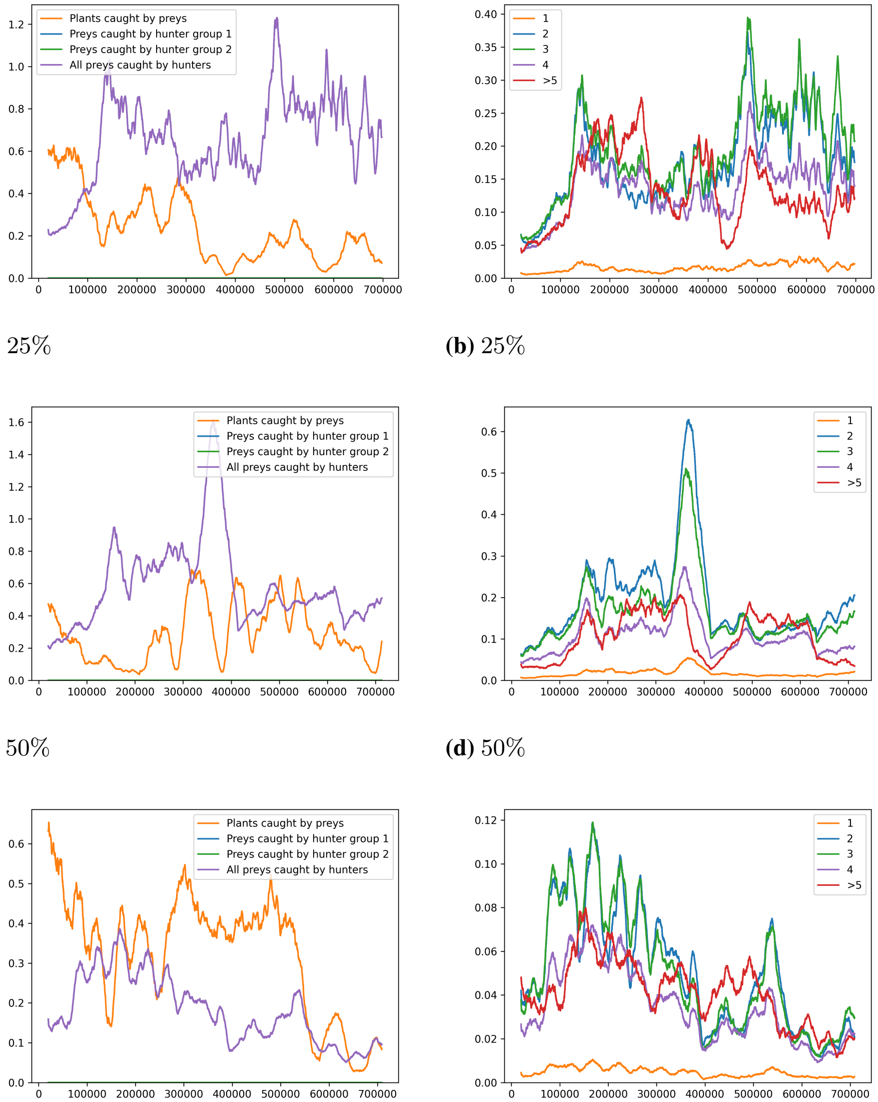
(c) 50%

(d) 50%

(e) 75%

(f) 75%

Fig. 5. Successful hunting and teamwork comparison between 25, 50 and 75 percent sharing with two predator populations



(a) 25%

(b) 25%

(c) 50%

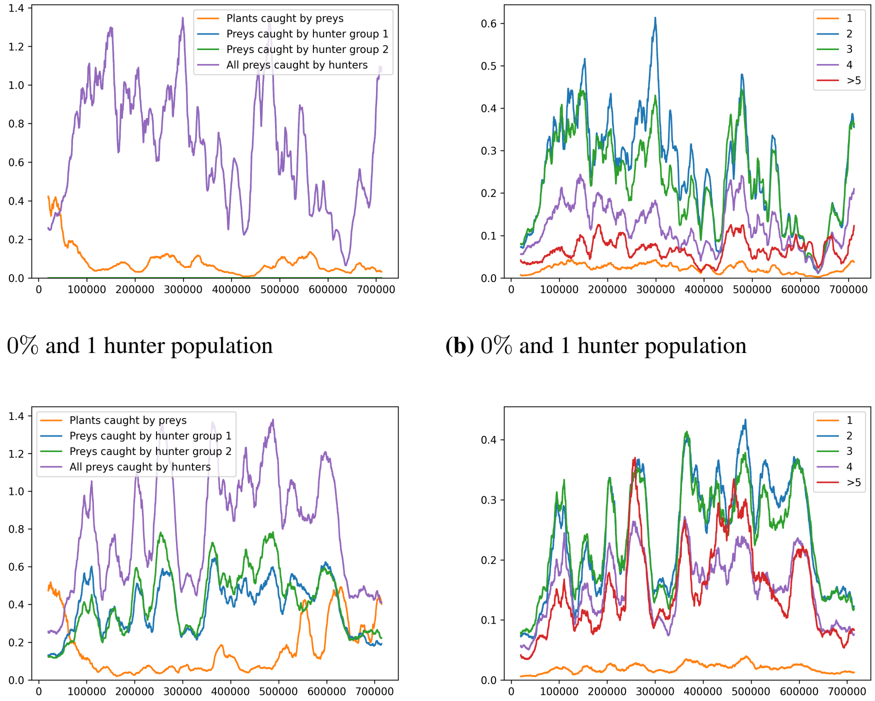
(d) 50%

(e) 75%

(f) 75%

Fig. 6. Successful hunting and teamwork comparison between 25, 50 and 75 percent sharing with one predator population

Sharing can be also set to 0%, which means that hunters will only get rewards if they are the ones catching the prey and there is no compensation for teamwork and blocking escape paths. The corresponding simulation runs are shown in *Figure* 7. This configuration often makes hunters avoid each other as much as possible to maximize their chance at being the one actually eating prey. This results in homogeneous hunter density over the map, which explains the consistently low amount of plant eating, because hunters are everywhere evenly spread out, making avoiding them priority over finding food. This behaviour can be observed best with one population, where all agents use very similar strategies and therefore hunters are distributed uniformly across the map. This affects co-operation by favoring smaller hunter groups, which is clearly visible in *Figure* 7/(b). The only exception is solo hunting which is still the lowest. There are two reasons for this, first of all hunting alone is a difficult task and prey caught this way probably has a low amount of fitness, which means that the hunter will not get enough nutrition out of it to survive. The second reason is the even distribution of hunters that makes it hard to be alone while hunting. With two populations the results are not as organized, but smaller groups are dominating in this case too with a few exceptions time to time. Once again with multiple populations the performance of the two hunter teams are closely following each other and their arms race is observable by examining the health and fitness of their top performing agents in *Figure* 8. The overall hunter performance is similar with both one and two populations and compared to previous configurations it is at the same level as P1S25 and P1S50.



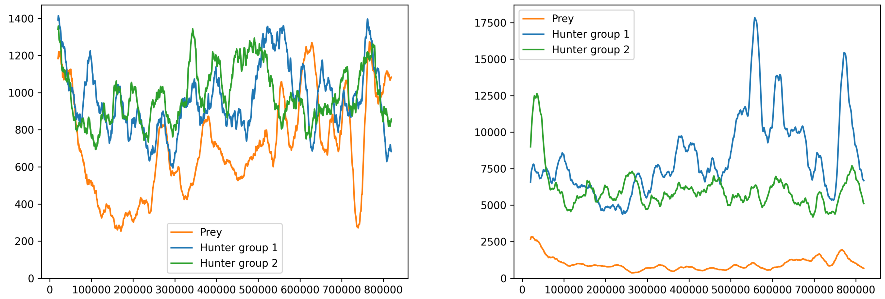
(a) 0% and 1 hunter population

(b) 0% and 1 hunter population

(c) 0% and 2 hunter populations

(d) 0% and 2 hunter populations

Fig. 7. Successful hunting and teamwork comparison with 0% sharing



(a) Health

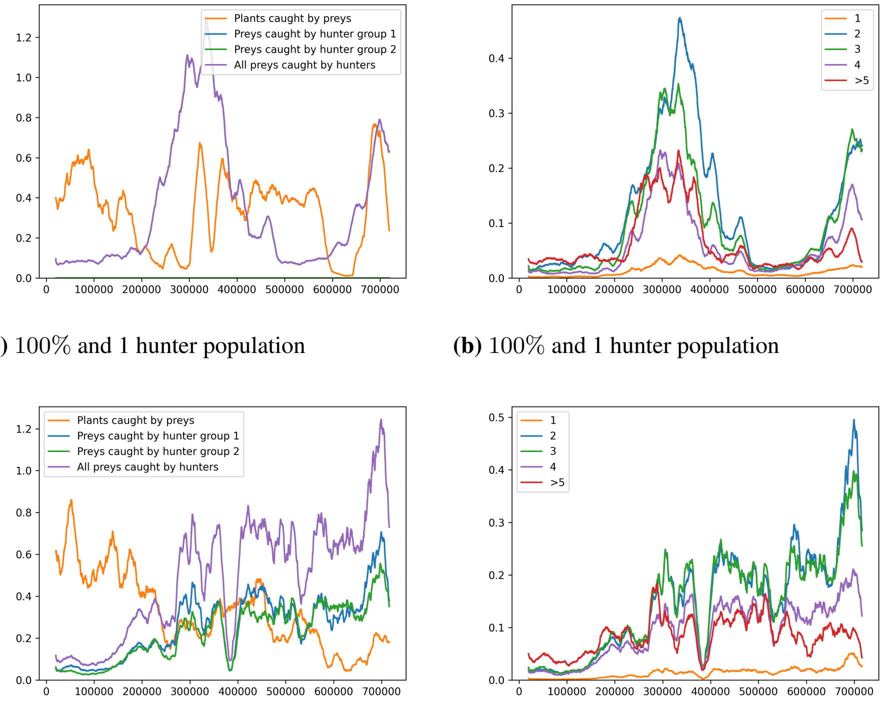
(b) Fitness

Fig. 8. Average health and fitness of the top 10% with 1 population and 0% sharing

Another special case is when there is 100% sharing. This means that whoever catches a prey will not get anything out of it, because it will be divided between other hunters nearby. If there is no one else inside the sharing radius, then the solo hunter gets all the nutritions. Even with this huge benefit for hunting alone, the results in *Figure* 9 show that this is still the least common method of successful hunting. With one population there is one huge spike in hunting from the 200000th time step to the 500000th, then at 600000 another spike starts growing. It looks like the hunters came up with a strong strategy to catch prey, but this is not the case. If we take a look at *Figure* 10 we can see that even though hunting numbers are high, a superior strategy is not reflected in the fitness of the best performing hunters. This can occur when preys are all easily caught and they do not prioritise avoiding enemies, therefore they can only provide small amounts of nutrition. The two small spikes in hunters' fitness correlates to the two spikes in plant eating at 320000 and 380000.

An interesting phenomenon can sometimes occur in these simulations, because of the locality of reproduction predators can force all preys into a corner and surround it, but not eat all of them at once. Given these circumstances plants quickly run out in the area and any attempt at escape is impossible. This produces high amount of capture and weak preys not providing enough energy. The situation can be broken up either by a successful escape attempt or by chance, when the new spawn point is generated behind the enemy lines.

In the case of two populations we can see a steady increase in prey capture throughout the whole simulation. In the first part huge hunter groups are dominating, but later smaller groups with 2 or 3 predators start to rise higher and higher, meanwhile the hunts including large groups are stagnating. At first the hunters were not successful, this led to preys eating a lot of plants, but as the evolution progresses this completely changes to the opposite. It looks like a gradual overtaking from the predators, but their progress was set back multiple times, the most noticable at the 380000th timestep and a smaller one at the 530000th. The recorded data of simulation did not tell us anything of these events, so we had to rely on our own observations. During these periods preys became more and more cautious and avoid predators further, meanwhile predators did not spread out leaving space to do so. Such a huge drop could have caused the extinction of predators, but in our system they recovered and started catching prey again.



(a) 100% and 1 hunter population

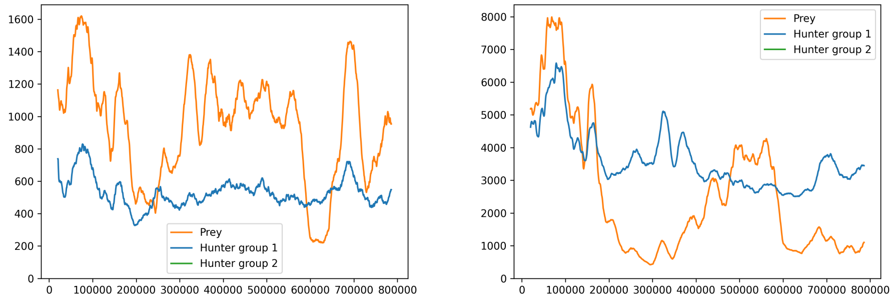
(b) 100% and 1 hunter population

(c) 100% and 2 hunter population

(d) 100% and 2 hunter population

Fig. 9. Successful hunting and teamwork comparison with 100% sharing

When we run simulations with two predator populations their sharing percentages can be set to be different. The results for P2S100S0, P2S80S20 and P2S60S40 is shown in *Figure* 11. P2S100S0 has the most diverse population, it consist of selfish and selfless agents. Surprisingly, the selfish population worked harder and caught more prey, most likely because the reward was high and it did not needed to be shared. Members of the selfless population can only get nutritions if another selfless agent hunts prey near them, this was not enough motivation for hunting. At first bigger groups were more common, but as the simulation progressed groups of size 2 and 3 took over. This change separated behaviours in the two populations. The smaller groups performance aligns almost perfectly with the selfish team's performance, and success of groups with 4 or more hunters aligns with the selfless team.



(a) Health

(b) Fitness

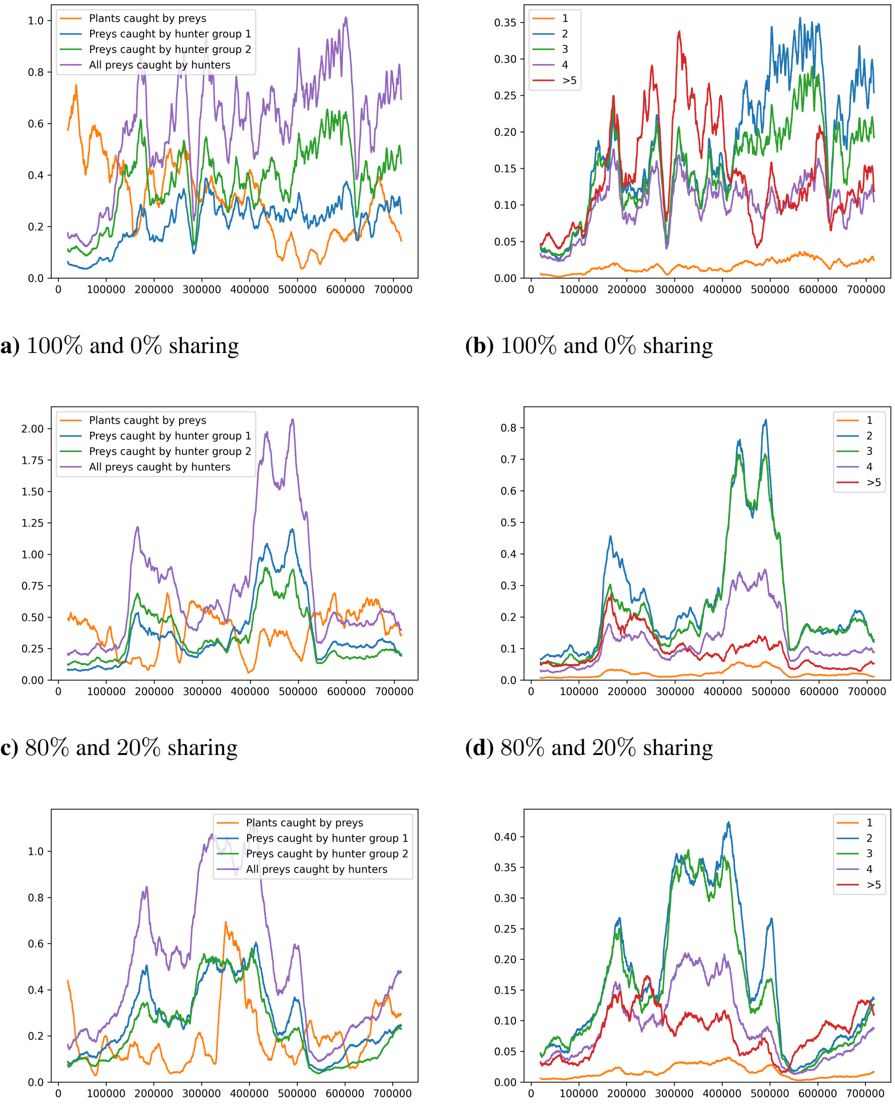
Fig. 10. Average health and fitness of the top 10% with 1 population and 0% sharing

With configuration P2S80S20 the selfless population is still less successful at first, but it closely follows the other teams performance and eventually it takes the lead. Overall the hunters have a higher efficiency than in P2S100S0, similar to P2S25, P2S50 and P2S75. Interestingly, huge groups were not as common as expected with half the population having high amount selflessness, teams containing 2 or 3 predators dominated.

One would expect that configuration P2S60S40 is very similar to P2S50, but the results tell otherwise. The overall performance is similar to P2S100S0 and there is visible arms race between the two hunter teams. A more interesting dynamic can be spotted between the hunters and preys, from 330000th to 410000th timestep the preys valued catching plants more than running from predators. This resulted in a spike in both plant eating and prey catching and then when the priority focused again on survival, both of them dropped significantly. Unlike with P2S50, here large hunter groups were not dominant at first, but this changed when the prey behaviour changed and a new strategy was needed to catch them.

5. Summary

So far we have discussed the individual runs and the hunters ability to catch prey. However, the agents were not trying to maximize the amount of food caught, they were trying to survive including eating and avoiding dangers, and their fitness values represent this. The simulations are diverse regarding the agents' average fitness at any specific timestep, but we needed to come up with a way to represent them easily for easier comparison between them. To measure the performance of the predators, we calculated the overall average of the predators' fitness score averages at every timestep. This was done twice, once with every hunter included, and then with only best performers (top 10%) at every timestep. To visualize the cooperation between hunters, we calculated the average number of hunters that played a role in a successful prey capture. The results for the simulations can be seen in *Figure* 12.



(a) 100% and 0% sharing

(b) 100% and 0% sharing

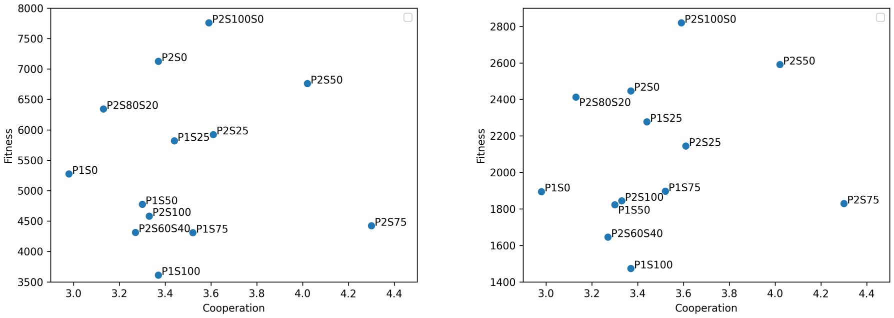
(c) 80% and 20% sharing

(d) 80% and 20% sharing

(e) 60% and 40% sharing

(f) 60% and 40% sharing

Fig. 11. Successful hunting and teamwork comparison between simulations, where populations have different sharing percentages



(a) Best performing hunters (top 10%) only **(b)** Fitness of all hunters

Fig. 12. Comparison of different configurations. The fitness is the average of the average fitness values of the hunters in every timestep. The cooperation is calculated by averaging the number of hunters in every successful hunt during the simulation. The results where the fitness was calculated only using the best hunters in every timestep can be seen in (a), the results for all hunters is shown in (b).

First take a look at the configurations with 1 population. The extreme case P1S0 is an outlier in performance as well as in cooperation. All others have a cooperative score at around 3.4 while P1S0 falls behind significantly and all the others follow the rule: more sharing equals less fitness. Following this rule P1S0 should have had the best fitness performance, but it only came in second. Comparing the 1 and 2 population results we can see that configurations with 2 populations always performed better regarding their fitness, often by a significant margin. Looking at the cooperation values it is the same situation, with only one exception (P1S100 and P2S100, which are both extreme cases). Focusing on only those with 2 populations and homogeneous sharing we can see that more sharing helps cooperation, again with one exception P2S100. Their fitness scores are not so consistent, overall with more sharing the performance decreases, but this is noticeable only with high sharing values (P2S75 and P2S100). The last three configuration with heterogeneous sharing show that the more diverse the population is the better the performance, however they are not consistent compared to their homogeneous counterparts and they have under average cooperation values. They do not fit into the picture painted by all the other configurations, the reason for this could be that more diverse populations have more potential to be volatile and our sample size was too small for them.
Overall our results suggest that a complex environment with a hybrid type of reinforcement (local for the individual agents, but with more ”global” (involving multiple agents in an area) rewards for cooperating) has a significant impact on performance and cooperation. Contrary to Balch’s [1] work, where his study resulted in better performance when global sharing was used, our findings show the opposite, but we have to keep in mind that our approach did not use global reinforcement for the whole population. Although performance decreased, cooperation was more common with reward sharing, similarly to Rajagopalan’s work [21]. We were also able to utilize Rawal’s [23] work and create our own scaled up environment with high number of agents, where we were able to keep up

an arms race in different configurations. These arms races often led to evolutionary trade offs and agents constantly tried to balance out feeding, avoiding danger and cooperation.

6. Conclusion

We successfully built a simulation framework where special aspects of artificial life and evolution can be observed, using cooperating neural networks in simple and memoryless agents. This was ensured by never letting the species go extinct, because instead of distinct generations a continuously changing "elite-fitness" population was used for selection and reproduction. We have shown that arms races arise in this complex continuous simulation, the behaviours of agents is constantly changing and they do not get stuck in a local optimum. We found that predators had better performance with two populations compared to one when the same amount of sharing percentage was used. We also showed that simulations with two populations resulted in more cooperation, but more sharing also decreased the fitness score. Overall the simulations never stagnated and were always changing thanks to the localized interaction and reproduction and fixed population sizes. There is also room for improvement, for example finding the optimal neural architecture or quantifying the diversity of the agents in different runs. We hope this framework will provide ground for more research in the future and helps understand rapid evolution and cooperation in competitive environments better.

Acknowledgments. The project has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002).

References

1. Balch, T.: Learning roles: Behavioral diversity in robot teams. Multiagent Learning: Papers from the AAAI Workshop (1997)
2. Beer, R.D., Gallagher, J.C.: Evolving dynamical neural networks for adaptive behavior. Adaptive Behavior 1(1), 91–122 (1992)
3. Dawkins, R., Krebs, J.R.: Arms races between and within species. Proceedings of the Royal Society of London. Series B, Biological Sciences 205(1161), 489–511 (1979)
4. Dietl, G.P.: Coevolution of a marine gastropod predator and its dangerous bivalve prey. Biological Journal of the Linnean Society 80(3), 409–436 (2003)
5. Folse, H., Allison, S.: Cooperation, competition, and coalitions in enzyme-producing microbes: Social evolution and nutrient depolymerization rates. Frontiers in microbiology 3, 338 (2012)
6. Gomez, F.J., Miikkulainen, R.: Active guidance for a finless rocket using neuroevolution. In: Genetic and Evolutionary Computation Conference. pp. 2084–2095. Springer (2003)
7. Gras, R., Devaurs, D., Wozniak, A., Aspinall, A.: An individual-based evolving predator-prey ecosystem simulation using a fuzzy cognitive map as the behavior model. Artificial Life 15, 423–463 (2009)
8. Hague, M.T., Toledo, G., Geffeney, S.L., Hanifin, C.T., Brodie Jr, E.D., Brodie III, E.D.: Large-effect mutations generate trade-off between predatory and locomotor ability during arms race coevolution with deadly prey. Evolution letters 2(4), 406–416 (2018)
9. Hassibi, B., Stork, D.G., Wolff, G.J.: Optimal brain surgeon and general network pruning. In: IEEE International Conference on Neural Networks. vol. 1, pp. 293–299 (1993)
10. Heaton, J.: Introduction to neural networks with Java. Heaton Research, Inc. (2008)

11. Hoffmann, A.A., Parsons, P.A. (eds.): Extreme environmental change and evolution. Extreme environmental change and evolution, Cambridge University Press (1997)
12. Hofstede, H., Ratcliffe, J.: Evolutionary escalation: The bat-moth arms race. The Journal of Experimental Biology 219, 1589–1602 (2016)
13. Ito, T., Pilat, M.L., Suzuki, R., Arita, T.: Population and evolutionary dynamics based on predator–prey relationships in a 3d physical simulation. Artificial Life 22(2), 226–240 (2016)
14. Kovács, Á.: Impact of spatial distribution on the development of mutualism in microbes. Frontiers in Microbiology 5, 649 (2014)
15. Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P.J., Bernard, S., Beslon, G., Bryson, D.M., et al.: The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. Artificial life 26(2), 274–306 (2020)
16. Lesser, V.: Cooperative multiagent systems: A personal view of the state of the art. Knowledge and Data Engineering, IEEE Transactions on 11, 133 – 142 (1999)
17. Marshall, R., Whitworth, D.: Is “wolf-pack” predation by antimicrobial bacteria cooperative? cell behaviour and predatory mechanisms indicate profound selfishness, even when working alongside kin. BioEssays 41 (2019)
18. Mitchell, M.: Coevolutionary learning with spatially distributed populations. Computational intelligence: principles and practice 400 (2006)
19. Mitchell, M., Thomure, M.D., Williams, N.L.: The role of space in the success of coevolutionary learning. In: Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems. pp. 118–124. MIT Press (2006)
20. Nolfi, S., Floreano, D.: Coevolving predator and prey robots: Do arms races arise in artificial evolution? Artificial Life 4, 311–335 (1998)
21. Rajagopalan, P.: The evolution of coordinated cooperative behaviors. Ph.D. thesis, Department of Computer Science, University of Texas at Austin (2016)
22. Réale, D., Mcadam, A., Boutin, S., Berteaux, D.: Genetic and plastic responses of a northern mammal to climate change. Proceedings. Biological sciences / The Royal Society 270, 591–596 (2003)
23. Rawal, A., Rajagopalan, P., Miikkulainen, R.: Constructing competitive and cooperative agent behavior using coevolution. In: Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games. pp. 107–114 (2010)
24. Rosin, C.D., Belew, R.K.: New methods for competitive coevolution. Evolutionary Computation 5(1), 1–29 (1997)
25. Savino, J.F., Stein, R.A.: Predator-prey interaction between largemouth bass and bluegills as influenced by simulated, submersed vegetation. Transactions of the American Fisheries Society 111(3), 255–266 (1982)
26. Williams, N., Mitchell, M.: Investigating the success of spatial coevolution. In: Proceedings of the 7th annual conference on Genetic and evolutionary computation. pp. 523–530 (2005)
27. Wu, Y., Tan, H., Jiang, Z., Ran, B.: Es-ctc: A deep neuroevolution model for cooperative intelligent freeway traffic control. arXiv preprint arXiv:1905.04083 (2019)
28. Yann L. Chun, John S. Denker, S.A.S.: Optimal brain damage. Advances in Neural Information Processing Systems pp. 598–605 (1990)
29. Yong, C.H., Miikkulainen, R.: Coevolution of role-based cooperation in multiagent systems. IEEE Transactions on Autonomous Mental Development 1, 170–186 (2009)
30. Zhang, S., Meng, X., Feng, T., Zhang, T.: Dynamics analysis and numerical simulations of a stochastic non-autonomous predator–prey system with impulsive effects. Nonlinear Analysis: Hybrid Systems 26, 19–37 (2017)

Krisztián Varga received his BSc degree in computer science at the Faculty of Informatics, Eövös Loránd University in Hungary and currently doing his master’s degree with the specialization of information systems. During his studies he attended several projects related to machine learning. His research interest is focusing on algorithms, programming, artificial intelligence. https://www.researchgate.net/profile/Krisztian-Varga-2

Attila Kiss defended his PhD in the field of database theory in 1991. His research is focusing on information systems, data mining, artificial intelligence. He has more than 165 scientific publications. Seven of his supervised students got their PhD degree. Since 2010 he has been the head of Department of Information Systems at Eötvös Loránd University, Hungary. He is also teaching at J. Selye University, Slovakia.
https://www.researchgate.net/profile/Attila-Kiss-3

Received: November 22, 2020; Accepted: April 04, 2021.

Computer Science and Information Systems 18(4):1197–1217 https://doi.org/10.2298/CSIS200715040Z

Using Honeynet Data and a Time Series to Predict the Number of Cyber Attacks

Matej Zuzčák and Petr Bujok

Department of Informatics and Computers, Faculty of Science, University of Ostrava
30. dubna 22, 701 03 Ostrava, Czech Republic
{matej.zuzcak,petr.bujok}@osu.cz

Abstract. A large number of cyber attacks are commonly conducted against home computers, mobile devices, as well as servers providing various services. One such prominently attacked service, or a protocol in this case, is the Secure Shell (SSH) used to gain remote access to manage systems. Besides human attackers, botnets are a major source of attacks on SSH servers. Tools such as honeypots allow an effective means of recording and analysing such attacks.However, is it also possible to use them to effectively predict these attacks? The prediction of SSH attacks, specifically the prediction of activity on certain subjects, such as autonomous systems, will be beneficial to system administrators, internet service providers, and CSIRT teams. This article presents multiple methods for using a time series, based on real-world data,to predict these attacks. It focuses on the overall prediction of attacks on the honeynet and the prediction of attacks from specific geographical regions. Multiple approaches are used, such as ARIMA, SARIMA, GARCH, and Bootstrapping. The article presents the viability, precision and usefulness of the individual approaches for various areas of IT security.

Keywords: cyber attacks, honeynet, honeypot, SSH, time series, prediction.

1. Introduction

Besides common users, servers providing various services are the target of virtually unceasing cyber attacks. These servers are most commonly managed using the SSH protocol. SSH provides the administrator with a remote access console offering the same functionality as if they were at the server site. It is one of the most commonly attacked protocols, both by human attackers and by automated bots that are a part of extensive botnets. The SSH protocol was selected as it is among the most frequently attacked protocols, according to the following reports: F-Secure Attack landscape H2 2018[1], Akamai - The State of the Internet Q4 2014[2]. Botnets most commonly use the computers of unaware users, connected to the internet via various technologies and internet service providers across the world.

Server administrators must inevitably protect their systems from a variety of attacks. To do so effectively, they must know and analyse the threats and use that knowledge

[1] F-Secure Attack landscape H2 2018 – `https://blog.f-secure.com/attack-landscape-h2-2018/`

[2] Akamai - The State of the Internet Q4 2014 – `https://www.akamai.com/us/en/multimedia/documents/state-of-the-internet/akamai-state-of-the-internet-report-q4-2014.pdf`

to, for example, expand the databases of IDS [3]/IPS [4] systems. Honeypots, which can be likened to lures or traps, are an ideal tool for this task. Besides being able to analyse historical attacks, a certain foresight of what to expect is also useful to administrators. This will allow them to prepare appropriate protective measures in advance and estimate what types of attacks are going to be prevalent.

In addition to server administrators, ISPs should also be aware of botnets and any infected computers on their networks. These companies would also benefit from an effective estimation of attack rates on their networks. It would allow them to more effectively deploy countermeasures to such attacks, such as dynamic IP address management, since the reputation of these addresses could suffer damage if they were assigned to an infected device conducting malicious activities. Therefore an overview of situational development enables ISPs to make appropriate decisions.

Other groups that can benefit from such foresight are the Computer Security Incident Response Teams (CSIRTs) and researchers to whom the ability to predict potential attacks is imperative. For instance, CSIRTs can effectively predict from which autonomous systems or IP address ranges intensive attacks can be expected, and how attacks with certain identifying features will progress. This would allow CSIRTs to prepare countermeasures and contact the operators of the affected autonomous systems ahead of time. Predicting specific details of how a threat spreads through the world, such as which RIR or country it will likely spread from, allows researchers to deploy monitoring tools appropriately to gain as much data as possible.

Various methods can be used to make predictions, with a time series being one of the most commonly used. There are multiple approaches to setting the necessary parameters. This paper analyses and compares the approaches with the goal of identifying the most effective one in predicting the attacks on a system over time. Predictions by every approach were made over the same period of time and compared with real-world data collected over the same period, a period of approximately one year. The real-world data was collected by the author's honeynet.

2. Honeypot and Honeynet

A honeypot [1] is a system for analysing activity taking place within itself. The activity is commonly malicious, with the goal of using the infected system to spread itself or other threats and conducting other malicious activity such as DDoS attacks or sending spam. A honeypot can consist of software, hardware, or an entire network [2]. Such a system is usually made intentionally vulnerable, and it provides no real-world services. It is usually operated with the intention of analysing and assessing the activity taking place within it. Such a system has to be closed insofar as no activity taking place within it could possibly negatively influence other systems or spread via LAN, WAN, or the internet in general. At the same time, the system must be sophisticated enough to allow the minimal possible contact between the attacker and the outside. The goal here is to give the attacker the impression of a real-life system it can conduct its activity within, without realising it is actually restricted. A compromise between the security and realism of the system has to be achieved, depending on what specific threats the honeypot is focused on.

[3] IDS – instruction detection system

[4] IPS – instruction prevention system

The term ”honeynet” [6] tends to be context-dependent. It is commonly used in reference to a honeypot with a high level of interaction. In this case, it means a specific type of network that, besides the honeypot, may contain other components, such as a special firewall called a honeywall, an IDS/IPS system, and various database systems for data collection, etc.

An additional meaning for a honeynet, is a system of honeypots forming a logical but non-physical system. This meaning is commonly applied to collections of honeypots with a low to medium level of interaction. The use of special tools such as firewalls is not necessary in this case. Data from all the honeypots in a honeynet are commonly collected into a single database. A honeynet can provide a large amount of threat data for analysis.

3. Related Work

The prediction of the development of attacks using time sets, applying various algorithms and methodologies is the focus of several papers. The paper [8] directly deals with predicting attacks detected by honeypots. It uses data from the CZ NIC honeynet that is composed of Kippo honeypots running on port 22. The paper proposes a model that predicted attacks on an emulated SSH protocol, providing the attacker with the ability to log in to the shell and execute some commands. Overall, 179 540 records from the period between 2.11.2014 and 8.5.2016 were analysed. Data from 75 weeks was used to train the model, and data from 5 weeks was used to compare the prediction of that period with the real data from it. An AR(1) - AR model of the 1st order time series with bootstrap point prediction was used. The paper concludes by stating the model is viable for predicting future attacks based on the demonstration.

In the paper [9], a large series with a large amount of data from security incidents is used. It compares the possible ways to predict attacks using a model based on a time series and using the Non-Homogeneous Poisson Process (NHPP) software reliability growth model.

The paper [10] proposes a prediction of the intensity of attacks based on known data regarding the number of attacks per day using the ARIMA model. Four types of attacks are identified: Denial of Service (DoS), malicious emails, malicious URLs, and attacks on the Internet facing service (AOIFS).

In the paper [11], an IDS system for wireless networks for process automation (WIA-PA) is proposed. It is based on recorded network traffic, processed using a model based on the ARMA time series.

In the paper [12], a framework for the prediction of vulnerabilities based on a statistical analysis using a time series between January 1999 and January 2016 is introduced. The ARCH, GARCH, and SARIMA models were used. The data was taken from the National Vulnerability Database (NVD) [5] in its 2016 state. The results of the predictions were mainly useful for the risk management of vulnerabilities.

In the paper [13], predictions based on two approaches, the Extreme Value Theory (EVT) and Time Series Theory (TST) are presented. The TST used the FARIMA + GARCH model. It concludes that EVT is more effective for predicting a longer time period, 24 hours and more, whereas the TST is better for immediate threats, such as within

[5] National Vulnerability Database – `https://nvd.nist.gov`

the hour. It uses data from a honeypot recording the network activity during five time periods in 2010 and 2011. The honeypot emulated several services using the following solutions: Amun [6], Dionaea, Mwcollector [7]a Nepenthes [8]. Data was extracted from the PCAP files generated by capturing network traffic, where every TCP connection, including an unsuccessful TCP handshake, was considered an attack.

In the paper [14], the prediction of attacks based on past event logs is studied. Various methods are applied and evaluated, such as the historical communication between the attacker and the victim, models for neighbour searching, techniques searching for global patterns using Singular Value Decomposition (SVD), and a time series using the Exponential Weighted Moving Average (EWMA) model. Logs from the Dshield project[9] over a period of one month, formed the data set used. The solution was proposed as a framework for a Blacklisting Recommendation system (BRS) as a linear combination of three approaches, namely a time series and two approaches from a neighbouring model area - K-Nearest Neighbor (KNN) algorithm and a co-clustering algorithm. Using SVD showed no significant improvement in the predictions, and it was therefore not included in the proposed solution. However, the solution could be useful for improving the generation of lists of the addresses of attackers.

The content of the paper [15] is not directly concerned with predicting attacks, but a time series is used to represent captured attacks and to demonstrate analytical outputs. Specifically, it proposes a framework for clustering captured attacks on honeypots to as many similar clusters as possible. Symbolic aggregate approximation (SAX) is the technique used for clustering, providing the ability to reduce the dimensionality of the data, therefore ignoring insignificant details. As a result, a cluster may contain attacks against different ports but represents the same network worm that spreads by using multiple ports.

In the paper [16], various aspects of prediction methods used in cyber security are analysed. The methods are divided into three categories: data mining, dynamic network entity reputation, and the use of time series'. The time series methods used were: ARIMA models, exponential smoothing models, "naive approach", and the average of the ARIMA and exponential smoothing models. The paper also looks at and evaluates the accuracy of the categories from the point of view of blacklisting. The data used in the paper was acquired from the SABU platform, which gathers intrusion detection alerts. The data covers a period of seven days. The authors conclude that attack prediction is an approach useful for estimating the number of attacks in the near future and can be used by the given system's operator to optimise its countermeasures. We consider the time period of seven days to be too short.

In the paper [17], a deep, state-of-the-art overview of the current approaches, taxonomy, and methods used for cyber security attack prediction are presented.

The content of the paper [18] is focused on "data-driven incident prediction" methods, and the shift from reactive to proactive approaches of protection.

The authors of the paper [19] propose an IACF framework focused on alert aggregation and correlation, and attack prediction and detection.

[6] Amun honeypot – https://github.com/zeroq/amun

[7] Mwcollector part of – https://sourceforge.net/projects/honeybow/files/honeybow/0.1.0/

[8] Nepenthes honeypot – Deprecated honeypot solution. It is no longer developed nor supported.

[9] Dshield project – https://www.dshield.org/

Only the papers [8], [13], and [15] contain data gathered by honeypots.

The aim of paper [8] is conceptually the closest to this one, due to the chosen approach, but it uses very few methods of prediction, only AR(1) and Bootstrapping. It also only predicts the number of attacks on honeypots, and does not deal with predicting the behaviour of individual attackers over time, and the relationship to geographical location and autonomous systems.

The authors of the paper [13] analysed a time series and an EVT approach. They only used a single time series method, and by considering every TCP connection to be an attack, it is arguably too broad a definition of an attack.

In the paper [15], honeypot gathered data is used, but the attacks are not directly predicted, but rather clustered using a time series.

In papers [9], [12] attacks are not predicted, but vulnerabilities and security incidents are predicted using the ARIMA, SARIMA, ARCH, and GARCH approaches.

In papers [10], [11], [16], [18], and [19], the intensity of cyber attacks in a wider context is predicted, for example, DoS attacks or malicious emails. It uses the ARIMA and ARMA approaches.

The analysis in the paper [14] is specific, as it analyses event logs using the SVD and EWMA approaches.

None of the available related research uses an approach utilising a range of time series based prediction methods and also do not focus on predicting attacks based on the geographical location or other clustering variables of the attackers, such as address ranges. Due to this fact, this paper focuses on these specific aspects.

4. The Honeynet Used and Delineation of the Relevant Time Period

Individual honeypots are based on various types of networks, with the captured connections, and the potential attacks, being sent to a central server where they are saved in a central MySQL[10] database for further analysis. The honeypots are presented in table 1. Each node, or sensor, is running an instance of the Cowrie honeypot emulating an SSH server.

Table 1. Honeypots composing the honeynet.

Sensor ID	Network type	Port
HP1	CESNET - Czech academic network	22
HP2	Czech VPS hosting - grey zone – grey zone	22
HP3	Regular Czech VPS hosting	22
HP4	Czech ISP	22
HP5	Slovak ISP - dynamic IP	2222
HP5-B	Slovak ISP - dynamic IP	22
HP6	VPS hosting - India	22

[10] MySQL– https://www.mysql.com/

4.1. Analysed Data

The honeynet captured all connections heading mostly to port 22, an SSH shell emulation. In one case port 2222 (Tab. 1) was used. Every connection established between a honeypot and a potential attacker is called a session. If during a session the potential attacker logs into the shell and conducts additional activity by inputting commands, such as downloading files and executing them, or uploading files from the emulated system, such a session is considered to be an attack in the context of this paper.

The article focuses on two main areas. The first is predicting the overall number of attacks against the honeynet in the given time period, described in Chapter 6.1. The second is predicting attacks based on their source, or from the point of view of their source, and is subdivided into three areas: Regional Internet Registry (RIR) in Chapter 6.2, country of origin (Chapter 6.3), and the activity of the autonomous systems (AS), specifically, where the attack originated from (Chapter 6.4).

A detailed analysis of the sources of attack from a geographical and analytical point of view is considered important. This is due to the needs of AS administrators often only concentrating on gathering and estimating the development of attacks in the area relevant to them. Therefore assessing the effectiveness of predicting attacks from the point of view of the source area is one of the main goals of this paper.

Before any prediction took place, the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test [24] was applied to the data to verify its stationarity. The null hypothesis of the test is as follows: the data is stationary around a deterministic trend, as opposed to the alternative where they are not stationary. Table 2 presents the calculated p-values, in the case of p being < 0.05 it means a rejection of the null hypothesis. For every evaluated aspect, such as the overall number of sessions or the source of attacks, a search for the best available prediction method was conducted, based on the evaluation of the Mean absolute percentage error (MAPE) in Chapter 5. Individual aspects were modelled by using a variety of established approaches: the Holt-Winters algorithm, ARIMA, SARIMA, GARCH models (for some situations where the data allowed it), and Bootstrapping models. With Bootstrapping, the predictions were always made using three approaches: stationary, fixed, and one based on a model. The one based on a model was not always viable. The tables in the following chapters only present the Bootstrapping model, which achieved the lowest MAPE error.

4.2. Time Period Used for the Prediction

Real data captured by the author's honeynet in the time period between 30.7.2017 and 7.11.2018 was used for predicting and for training the methods. Considering the time period is 16 months, an accumulation of the daily data had to be considered, mainly for reducing the zero value observation for some days. (i.e. there are 466 daily-measured values). Cumulating it into weeks (i.e. seven daily-measured values into one) seems appropriate, as months or quarters of the year would result in too few data points for prediction, therefore, dramatically decreasing the accuracy.

As a result of this, a time period of seven days was chosen to accumulate the measured values to a weekly aggregate. This resulted in 66 weekly data points, out of which 58 weeks were used to train the models, and the last eight weeks were used to test the accuracy of the predictions.

5. Methods Used for Prediction

There is an entire gamut of methods for predicting future observations using a time series. Traditional approaches are mainly based on the decomposition of values in a time series, or by using the Box-Jenkins methodology. Besides the more traditional approaches, other, less conventional ones are available, for example, those based on Bootstrapping. This paper applies several approaches to obtain as accurate a prediction of future observations for a time series as reasonably possible, while also demonstrating the robustness of the methods used. To predict the future values of a time series, $Y_{t+\tau},\ \tau = 1, 2, \ldots$, with sufficient accuracy, a standard deviation of prediction from the real value, an error, has to be introduced:

$$e_t = Y_t - Y_t^{'}(t-1)\ , \tag{1}$$

where Y_t is the value of the time series in the time t, and $Y_t^{'}(t-1)$ is the prediction of that value from the value of the time series in the time $t-1$. Using the error, we can evaluate the quality of the predictive model based on the values of the time series using the *Mean absolute percentage error* (MAPE):

$$\mathit{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\frac{|e_t|}{Y_t} \tag{2}$$

5.1. Holt-Winters

In 1957 Holt introduced a general algorithm of exponential smoothing [25], which was subsequently expanded by Holt and Winters three years later [26]. The Holt-Winters algorithm is based on three components of a time series: level, trend, and a seasonal component. Based on the application of the components, there are two variants of the algorithm, *additive* and *multiplicative*. In the additive model, the components add up, with each being measured in the same units as the time series itself. In the multiplicative model, only the level is in the time series' units, the trend and the seasonal component are factors within the interval $(0,\ 1)$. Even though the Holt-Winters algorithm is rather simple, the results show it achieves very accurate predictions in many different contexts and areas.

5.2. ARIMA

The stationary mixed model of Box-Jenkins methodology ARIMA(p,d,q) was introduced in 1970 [23] and it can be symbolically expressed using the following equation:

$$\phi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t \tag{3}$$

where ϕ is the autoregressive (AR) process, θ is the process of the moving average (MA), B is the lag operator, d is the differentiation operator, and ε_t is white noise.

Besides the autoregressive process AR(p) and the moving average process MA(q), the model also contains the differentiation operator I, which is used to stationarise the non-stationary time series.

The ARIMA model can be calibrated by adjusting the values of the parameters p, d, and q. Setting a parameter value to zero leaves the parameter out, so for example, if $d = 0$ the model is ARMA(p,q) and so on.

5.3. SARIMA

SARIMA is a variant of the ARIMA model expanded to include the seasonal part, allowing it to model a time series influenced by a seasonal component. The model is inscribed as SARIMA(p,d,q)(P,D,Q)$_{Sz}$, where the symbols in the first pair of brackets represent the parameters of the standard ARIMA, while those in the second pair represent the seasonal variants. The Sz parameter is the number of seasons per year. The SARIMA model can also be inscribed using a lag operator:

$$\phi(B)\Phi(B^{12})\Delta^d\Delta_{12}^D Y_t = \theta(B)\Theta(B^{12})\varepsilon_t \tag{4}$$

As with ARIMA, the model can be calibrated by adjusting the values of the parameters (p, d, q, P, D, Q, Sz), with ε_t again being white noise. Setting a parameter to zero omits it.

5.4. GARCH

The GARCH (Generalized Autoregressive Conditionally Heteroscedastic) model was introduced in 1982 [27] and is a generalisation of the ARCH model. GARCH assumes variable volatility, the heteroscedasticity, of a time series. The value of the series in time t can be inscribed using the GARCH(m, s) model as:

$$Y_t = \mu_t + \varepsilon_t\sqrt{\sigma_t} \tag{5}$$

with

$$\sigma_t^2 = \alpha_0 + \alpha(B)Y_t^2 + \beta(B)\sigma_t^2 \tag{6}$$

where $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ are the parameters of the model.

5.5. Bootstrapping

The Bootstrapping technique was introduced in detail in paper [28] by Bradley Efron. Bootstrapping is a very straightforward technique. In order to calculate the confidence interval *CI* for a statistic $T = t(X)$ with a set of n elements $X = x_1, x_2, ..., x_n$, it just repeats the following scheme R times:

- For the i-th iteration: sample n elements from the available sample, while allowing for the repeated choice of the same elements.
- Based on the sample created in the previous step X_i, calculate the new statistics $T_i = t(X_i)$.

There are several modifications of the Bootstrapping method, one is known as block Bootstrapping. Here, the data vector $(X_1, ..., X_n)$ is divided into blocks of length l.

$$Y_1 = (X_1, \ldots, X_l), Y2 = (X_{l+1}, \ldots, X_{2l}), \ldots Y_k = (X_{(k-1)l+1}, \ldots, X_n) \tag{7}$$

This is followed by an independent random sampling from the population of vectors $Y_1, \ldots, Y_N$, providing the sampled vectors $Y_1^*, Y_2^*, \ldots, Y_N^*$. A vector of random variables $(X_1^*, \ldots, X_n^*) = (Y_1^*, \ldots, Y_k^*)$ is considered a Bootstrap selection.

For fixed block sampling (BS_fixed), the date of the beginning of the block is generated first, followed by the date of the point chosen, this allows the time series to have the same length as the block.

For stationary block sampling (BS_stationary), the date of the beginning of the block is generated first using geometric distribution. The new block of data is then drawn onto a new time point and added to the series. This process is also repeated until the new series has the same length as the original one.

In this paper, Bootstrapping methods from the R[11] [30] language and $tsboot$[12] function containing several methods for the resampling of a time series were used. The function $auto.arima$[13] was used, with the parameters: $max.p = 25$, $max.d = 0$, $max.q = 0$, $max.P = 0$, $max.Q = 0$, $max.D = 0$, $ic = `aic'$, $max.order = 25$, $seasonal =$ *TRUE*. Resampling was set to the value of 100. More detail about the residual bootstrap method is found in [29] and a description of auto.arima in R is found in [31].

6. Prediction of Attacks on the Honeynet

This paper applies various models of time series to predict either the overall number of attacks, or the behaviour from individual sources of attacks. The stationarity hypothesis of the data was tested using the standard Kwiatkowski–Phillips–Schmidt–Shin test (KPSS) described in Chapter 4.1 for every specific time series. The most accurate predictions of the given aspect evaluated by the lowest Mean absolute percentage error (MAPE) are plotted as a graph, effectively demonstrating the predictions through real time. In all figures except Fig. 5, the real time series is presented by the green plot, with the orange plot representing the teaching run, and the yellow plot illustrating the predicted values of the time series provided by the most accurate method (the model with the lowest MAPE error).

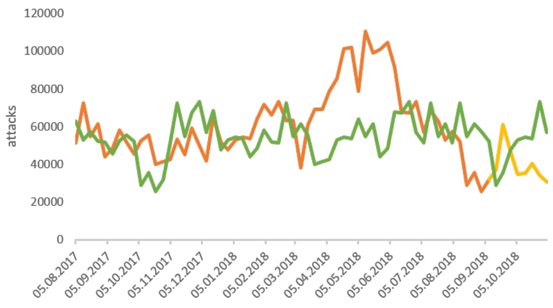


Fig. 1. Overall number of attacks in the given time period

[11] R language – https://www.r-project.org/

[12] Tsboot function – https://www.rdocumentation.org/packages/boot/versions/1.3-23/topics/tsboot

[13] Auto.arima function – https://www.rdocumentation.org/packages/forecast/versions/8.9/topics/auto.arima

6.1. Prediction of the Total Number of Attacks

The first aspect to be analysed was the overall activity of the attacks and sessions directed at the honeynet. This data shows the activity to be quite variable and unstable, as the sources are often home computers being used as part of a botnet. A more detailed analysis is available in [22]. The KPSS test applied to the measured data achieves a significance level of 0.02, see Table 2. Therefore the null hypothesis for the stationarity of the data is rejected. The results obtained from applying each individual method are presented in Table 3. The ARIMA(1,1,0) method achieves the lowest error and thus also the lowest deviation from the real time data of the predicted time period. It is represented in Fig. 1. The results of the predictions can be assessed, given the dynamic behaviour of the attackers, as satisfactory. Analysts can use this to predict trends in the future activity of attacks on the honeynet. However, for a more detailed analysis, the predictions must be made in shorter time frames, as highlighted in the following chapters.

Table 2. KPSS for all time series data.

data	**All_attacks**	**AFRINIC**	**APNIC**	**ARIN**	**LACNIC**
KPSS	0.02	0.05	0.07	>0.10	<0.01
RIPENCC	**AS4134**	**AS4837**	**AS16276**	**AS14061**	**AS45899**
>0.10	>0.10	>0.10	0.04	>0.10	>0.10
China	**Russia**	**Netherlands**	**USA**	**France**	
>0.10	<0.01	>0.10	<0.01	<0.01	

Table 3. An overview of the MAPE values when predicting the overall number of attacks on the honeynet.

Attacks on honeynet	**MAPE (%)**
Holt-Winters$_A$	50.3
SARIMA(1,1,0)(2,0,0)	43.5
SARIMA(0,1,0)(2,0,0)	42.1
SARIMA(1,1,0)(0,1,0)	22.4
ARIMA(1,0,0)	36.2
ARIMA(1,1,0)	22.0
GARCH(2,2)	26.9
BS_fixed	50.6

The values of significance for the KPSS test relating to the individual time series modelled, are presented in Table 2. If the value of significance is less than 0.05, the assumption of the stationarity of the time series is rejected. Based on the KPSS test, the following time series are stationary: AFRINIC, APNIC, ARIN, RIPENCC, AS4134, AS4837, AS14061, AS45899, China, and the Netherlands.

6.2. Activity of Attackers from the Point of View of RIR

When considering RIR, the predictions were very close to the geographical distribution of the continents, allowing the prediction of attacking trends from certain regions. As shown in Table 4 and Fig. 2, which present the best models, attacks from each RIR were best predicted by a different approach.

Table 4. Overview of MAPE values when predicting attacks on a honeypot from individual RIRs.

AFRINIC	MAPE	APNIC	MAPE	ARIN	MAPE
Holt-Winters$_A$	109.2	Holt-Winters$_M$	31.2	Holt-Winters$_A$	82.4
SARIMA(1,0,0)(0,1,0)	87.3	SARIMA(1,1,0)(1,0,0)	41.4	SARIMA(2,1,0)(0,1,0)	104.8
SARIMA(1,1,0)(2,0,0)	39.7	SARIMA(1,1,0)(2,0,0)	30.3	SARIMA(2,1,0)(1,0,0)	81.9
ARIMA(1,0,0)	32.3	ARIMA(2,0,0)	110.8	ARIMA(1,1,0)	34.7
GARCH(1,1)	35.3	GARCH(1,1)	130.1	GARCH(1,1)	55.7
BS_stationary	64.9	BS_stationary	73.1	BS_stationary	33.6
LACNIC	**MAPE**	**RIPENCC**	**MAPE**		
Holt-Winters$_M$	57.3	Holt-Winters$_A$	41.7		
SARIMA(1,0,0)(0,1,0)	37.7	SARIMA(0,1,0)(1,0,0)	28.0		
SARIMA(1,0,0)(2,0,0)	31.9	SARIMA(1,0,0)(1,0,0)	26.7		
ARIMA(1,0,0)	33.6	ARIMA(1,0,0)	27.0		
BS_fixed	35.6	BS_stationary	30.0		

From the graphs representing the individual aspects of attacks, it is apparent these are not easily predictable variables, as their expected value and variance change over time.

The best models were able to predict the development over time with a MAPE error of roughly 30%. The most accurate prediction with the lowest error of 26.7% was achieved by RIPENCC. Again, this suggests the prediction of incoming attack trends from individual RIRs can be rather easily predicted. The accuracy of individual RIR predictions does have a 30% error, although the results are still accurate enough to be useful to researchers for analysis.

According to KPSS, the AFRINIC time series is stationary, resulting in very good results for the ARMA(1,0) model with only a 32% error of prediction. With the APNIC time series, the rather simple Holt-Winters approach was able to achieve a very good level of prediction with an error of only 31%. This success can also be attributed to the series being stationary according to KPSS. Even though the ARIN series is stationary, the ARIMA(1,1,0) stationary model, has the best results here.

The model of non-stationary series LACNIC - SARIMA(1,0,0)(2,0,0) with an error of 30%, being non-stationary, is also surprising.

The last stationary region is RIPENCC, with very similar results achieved by the two models of the Box-Jenkins methodology. However, the seasonal SARIMA(1,0,0)(1,0,0) achieves an even smaller error of 26.7%.

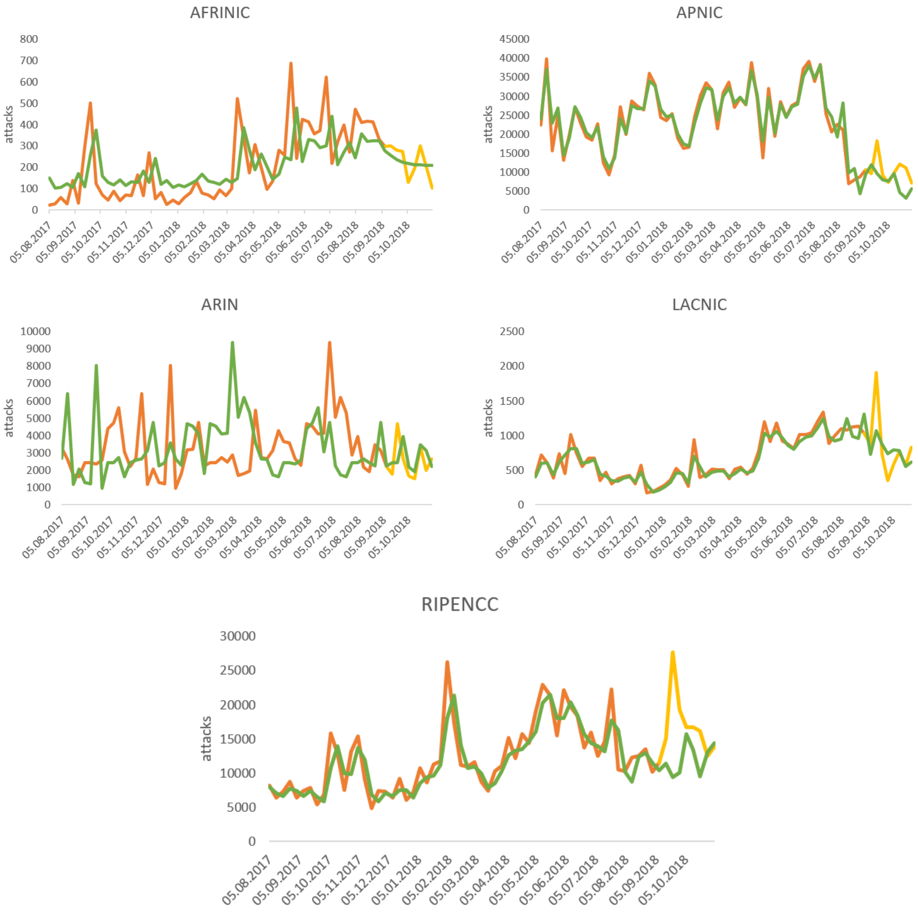


Fig. 2. Prediction of attacks from individual RIRs. The model with the lowest MAPE error according to Table 4 is always presented. The green plot represents the real progression of attacks over time, with the orange representing the teaching run and the yellow in the foreground representing the model.

6.3. Activity of Attackers from the Point of View of Individual Countries

The prediction of attacks based on the country of origin proved to be the most accurate in this research. Given the limited extent of this text, the five most active countries are presented in Table 5 and Figure 3, with the graphs of the best models. The MAPE error of the best models is between 20% (China) and 54% (France), and compared with the errors for RIR, they more accurately predict attacks. The errors for variants of the Bootstrapping model are in the 5th and the 10th line of Table 5.

Analysing the models of prediction and their errors in detail, it shows that three out of the five states obtained their lowest error by using ARIMA. In the case of attacks from China, the ARIMA(1,1,0) model with differentiation has the best result, which is surprising since this time series had weak stationarity according to KPSS. For comparison, ARIMA(1,0,0) a non-stationary model has a 4% higher error. The same model, ARIMA(1,1,0), was also the best for predicting attacks from Russia, which is more fitting since the series is not stationary according to KPSS. The error of prediction for the most successful model for the USA, ARIMA(2,1,0), is 21.4%. It is a very accurate model meant for non-stationary series, which the US one is, according to KPSS. The least accurate of the five countries are the models predicting attacks from France, with the most accurate one being the seasonal SARIMA(0,1,0)(1,0,0) model with differentiation, with an error of 54%. The only model based on Bootstrapping achieving the highest attack prediction accuracy, was for the Netherlands, with an error of 36%.

Apart from the pure research aspect, this information can be very useful to national CSIRT teams, allowing them to prepare appropriate countermeasures in their country well in advance. From a global point of view, predicting attacks based on their source, especially a country based prediction, seems to be the most accurate. This is probably influenced by each country having its own specific predictable variables, like the number of connected computers for common users, the number of servers, habits of the users, and security standards. The predictions were most successful for China, the USA, and Russia. The reason is probably because the USA and China have a proportionally large number of devices connected to the internet. The USA belongs to the ARIN RIR, and China to the APNIC RIR. As shown here, both of these regions obtain a rather good prediction level with an error of roughly 30%. These two countries are also major parts of their regions, allowing for the successful prediction for their RIRs as a whole.

The predictions for the cases of European countries, France and the Netherlands, is less accurate. The activity is more dynamic, and in the case of the Netherlands, it is also influenced by a disproportionately large number of data centres being located there, yet managed from other countries. A detailed analysis of this issue with the Netherlands is found in the paper [22]. France, the Netherlands and Russia belong to the RIPENCC RIR. As mentioned above, the prediction for this region was the most accurate of all the RIRs. In the case of RIPENCC compared to ARIN and APNIC, the main reason for the high accuracy is probably because it contains a large number of small countries and the impact of these is not as large individually as China or the USA in their respective RIRs.

Table 5. An overview of MAPE results for the prediction of attacks from the five most active countries.

China	**MAPE**	**Russia**	**MAPE**	**USA**	**MAPE**
Holt-Winters$_M$	32.6	Holt-Winters$_M$	34.0	Holt-Winters$_M$	59.7
SARIMA(0,0,0)(1,1,0)	29.6	SARIMA(1,1,0)(1,0,0)	47.8	SARIMA(0,0,0)_(1,1,0)	52.1
ARIMA(1,1,0)	20.1	ARIMA(1,1,0)	27.2	ARIMA(2,1,0)	21.4
BS_autoARIMA(1,0,0)	61.5	BS_fixed	33.4	BS_stationary	40.6
France	**MAPE**	**Netherlands**	**MAPE**		
Holt-Winters$_M$	64.4	Holt-Winters$_A$	61.8		
SARIMA(0,1,0)(1,1,0)	54.0	SARIMA(1,0,0)(0,1,0)	52.8		
ARIMA(2,1,0)	87.9	ARIMA(1,1,1)	44.0		
BS_fixed	66.5	BS_fixed	36.2		

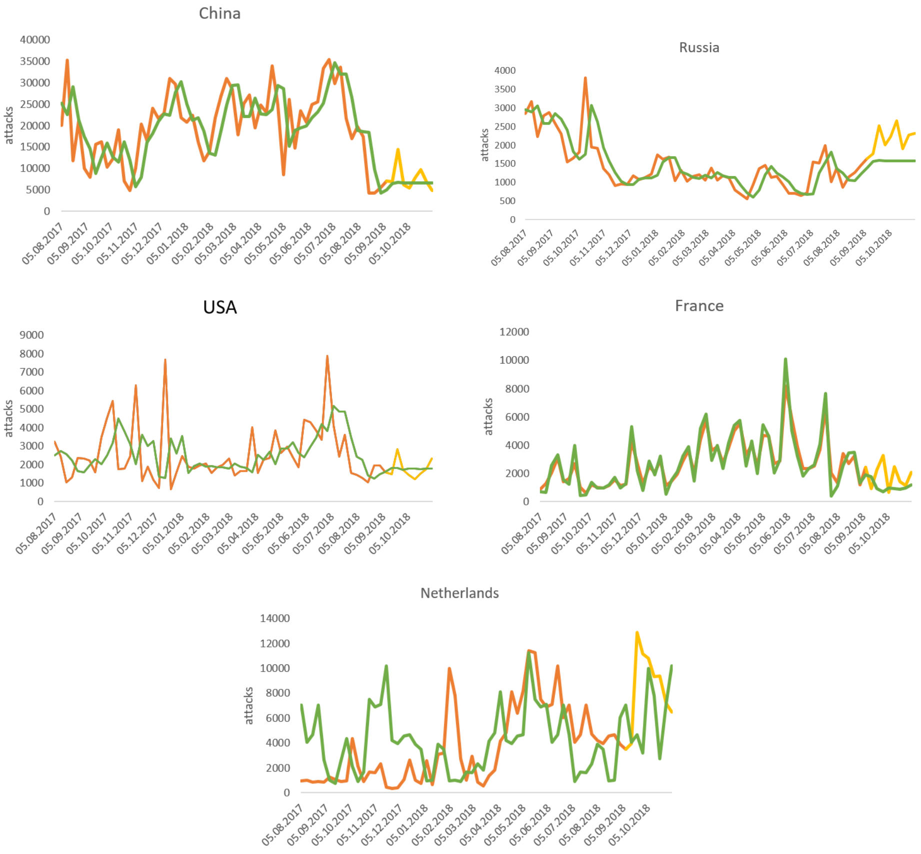


Fig. 3. An overview of MAPE results for predicting the number of attacks from the five most active countries. The green plot represents the real progression of attacks over time, with the orange representing the teaching run and the yellow in the foreground representing the model.

6.4. Activity of Attackers from the Point of View of Autonomous Systems

The prediction of attacks ascertained by the autonomous systems[14] was shown to be the least useful. Table 6 and Figure 4 show that the MAPE error for the five most active autonomous networks varies significantly.

The best error values were between 49.7% and 58%. These are not very accurate estimates, considering most of the series for autonomous systems are stationary. Considering the worst MAPE errors, values higher than 1000% were obtained. With three out of the five autonomous system models, the best predictions were achieved using ARIMA. With the stationary series AS4134 the model ARIMA(1,0,0), achieved the most accurate prediction, with the seasonal SARIMA model achieving more than four times the standard error of about 220%. Large differences in the accuracy of the models are also shown in the case of AS4837, for which ARIMA(1,1,0) was the best, achieving 58% accuracy, even though it is a stationary series, while Bootstrapping predicted values with an error of over 1060%. It should be added that with all the methods of prediction, the best ones were chosen based on the analysis of the settings of the model. System AS16276 achieved a reasonable error level ranging from 51.6% with SARIMA, to 69.2% with the multiplicative Holt-Winters algorithm. The graph for this model shows a very accurate approximation of attacks for the teaching part of the model, with a noticeable reduction in accuracy for the verification part. The next chapter presents the difference between the numerical and factual accuracy of a prediction. Rather balanced errors (compared to errors of other autonomous systems) were also achieved in the case of AS14061, from 49.7% for Bootstrapping, to 76.1% for SARIMA. The stationary series AS45899 was predicted the most accurately by an ARIMA model using a differencing step, with an error of 57.8%, and it was predicted the least accurately by the Bootstrapping model, with an error of 116.4%.

The instability of systems connected within specific autonomous systems is high, whether they are home computers, workstations, or IoT devices. Users turn them on and off at various times, for variedly long periods, with the ISP often mitigating DDoS and spam activity. The common dynamic of assigning addressees should be considered as well. Based on the predictions obtained, it can be concluded that predicting attacks based on autonomous systems is not very effective, with the ISP and AS providers being better off choosing a different approach to predict attacks on their infrastructure.

6.5. Representation of Accuracy and Applicability of the Predictions

The previous chapter presents models for the prediction of attacks using the most accurate methods, specifically, the methods that achieved the lowest MAPE error. When analysing the difference between the predicted and real values of attacks, it appears that even when the model has the lowest error, the prediction is often not very reliable when compared to real values. This is because it predicts the values by a linear or an exponential curve. Therefore, graphs displaying the prediction using the various methods for the chosen models were created, as presented in Figure 5.

[14] Autonomous system (AS) – is a collection of connected Internet Protocol (IP) routing prefixes under the control of one or more network operators on behalf of a single administrative entity or domain that presents a common, clearly defined routing policy to the internet. RFC 1930 - `https://tools.ietf.org/html/rfc1930`

Table 6. An overview of MAPE results for the prediction of attacks from autonomous systems.

AS4134	**MAPE**	**AS4837**	**MAPE**	**AS16276**	**MAPE**
Holt-Winters$_M$	159.4	Holt-Winters$_A$	146.7	Holt-Winters$_M$	69.2
		GARCH(1,1)	696.2		
SARIMA(0,0,0)(1,1,0)	219.5	SARIMA(0,0,0)(1,1,0)	216.6	SARIMA(1,1,0)(1,1,0)	51.6
ARIMA(1,0,0)	52.6	ARIMA(1,1,0)	58.0	ARIMA(1,0,0)	60.8
BS_autoArima(1,0,0)	90.4	BS_fixed	1061.8	BS_fixed	57.6
AS14061	**MAPE**	**AS45899**	**MAPE**		
Holt-Winters$_A$	74.1	Holt-Winters$_M$	64.5		
SARIMA(1,1,0)(0,1,0)	76.1	SARIMA(0,0,0)(1,1,0)	76.5		
ARIMA(1,1,0)	56.2	ARIMA(1,1,0)	57.8		
BS_fixed	49.7	BS_stationary	116.4		

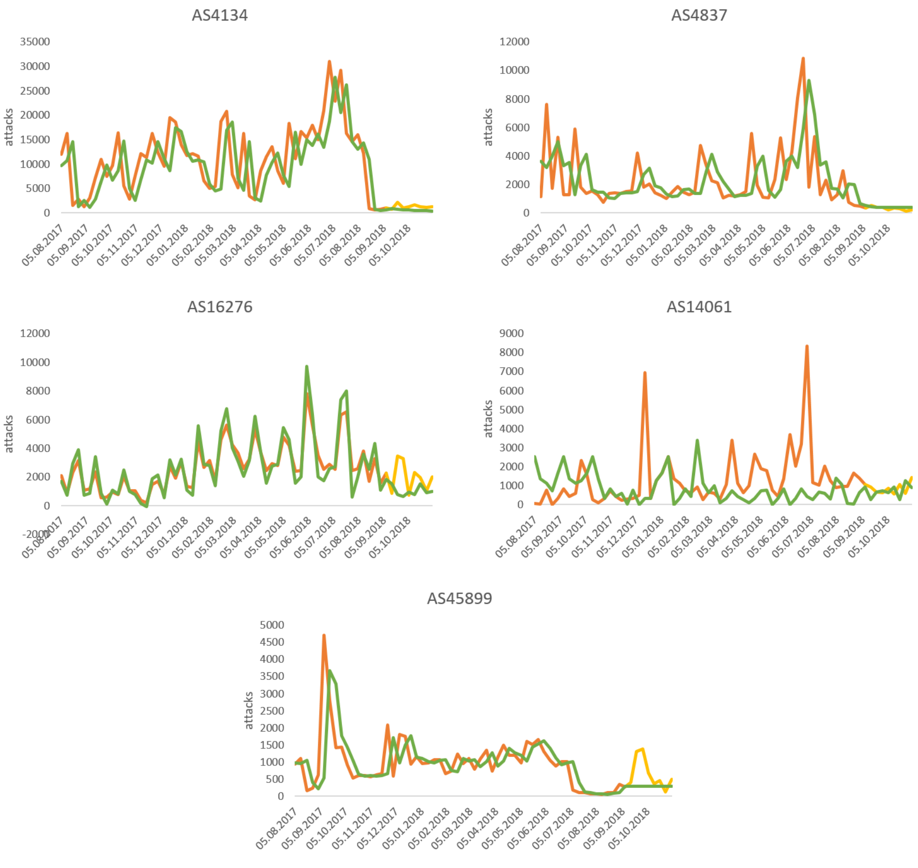


Fig. 4. An overview of MAPE results for predicting the number of attacks from individual autonomous systems. The green plot represents the real progression of attacks over time, with the orange representing the teaching run and the yellow in the foreground representing the model.

The lowest error of prediction, 27% for Russia, was achieved with the ARIMA model, and the curve is nearly constant. Conversely, the second most accurate, the Bootstrapping model, with 33% of error, or the Holt-Winters multiplicative model, with 34% of error, have variable curves over time. It is evident with certain predicted values that these two models are further away from the real values than ARIMA, however, in some other examples, they very accurately predict the values of the real series.

A similar situation occurred with the number of attacks from China, where the most accurate model is ARIMA, with 20% of error, predicting almost constant values. In contrast, SARIMA, with 30% of error, and the Holt-Winters multiplicative model, with 33% of error, are often far closer to the values of the real series.

The prediction of the France time series is the most accurate with the SARIMA model, with 54% of error. Even though the predicted values are not in a linear nor an exponential curve, it is evident it mostly covers the bottom peaks of the real values. The model with the second lowest error of 64%, the Holt-Winters multiplicative algorithm, predicts very similarly. However, while the Bootstrapping model achieves a large percentage error of 66%, it is evident that besides the first value, the prediction is rather close to the real values of the series.

In the case of the APNIC time series, the ARIMA and GARCH models achieve the worst predictions with the largest errors, 110% and 130% respectively, and their curves show no relation to the real values. Alternatively, both the lowest errors, 30% and 31%, and the closest curves were achieved by SARIMA and the Holt-Winters multiplicative model.

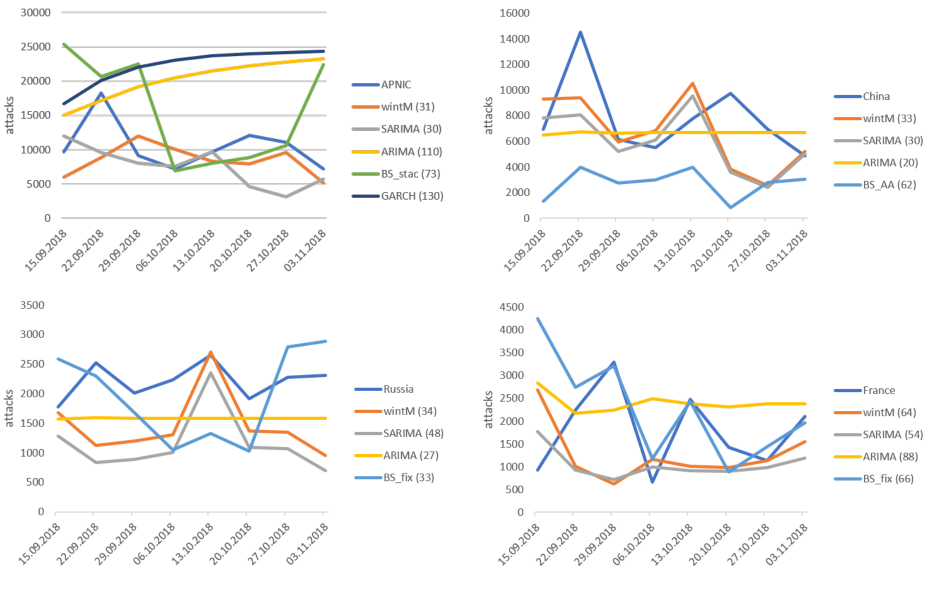


Fig. 5. Representation of accuracy and applicability of prediction.

6.6. Overall Evaluation

Overall, it can be concluded that the most effective predictions of attacks on the honeynet were achieved with a time series predicting the number of attacks from individual countries (i.e. the lower error values were achieved mostly for the data from individual coutries). Such predictions are relatively accurate and can be useful to national CSIRT teams as well as researchers. However, the least effective prediction was achieved with a time series predicting attacks from autonomous systems. With the RIR time series, the predictions can be accurate using an appropriate method for the given RIR. The overall prediction of attacks on a honeynet, regardless of the source, provides a reasonably accurate prediction and potentially useful prediction of attacking trends.

The influence of user and provider behaviour on systems located in a specific country is conclusively strong. Aspects such as user behaviour, the number of provided services such as VPS servers, and security measures are very specific to individual countries, so that when they are grouped, as in a RIR for example, the similarities are not sufficient to increase prediction accuracy. For example, RIPENCC consists of countries from both Western and Eastern Europe, with rather large differences in the behaviour of users and IT services. Even though the predictions from some individual countries such as Russia or France are not very accurate, their influence on the entire model is not sufficient to counter the better predictions of attacks from RIPENCC as a whole. In the case of the USA and ARIN, the prediction is good in both cases, since the USA is a major part of ARIN, reflecting and influencing the predictions of the entire RIR.

While predicting based on countries can be useful to researchers and CSIRT teams, it has its limits. It is important to realise it can only accurately predict for a relatively stable time period, a period during which no new rapidly spreading threat emerges, for example due to a newly found vulnerability. In such a case, it would cause a rapid drop in accuracy. With this in mind, it signifies the necessity to use rather short time periods for prediction to both maximise prediction accuracy and minimise the impact potentially new, rapidly spreading malware will have. Fortunately, the emergence of such new malware is not a very common situation.

From a statistical point of view, it is valid to say that in the case of some of the time series, the assumption that a non-stationary series is best predicted by methods using a differencing step, and vice versa was proven incorrect. This is probably caused by the unpredictable changes in the number of attacks, even though the prediction of the constant expected value, the variability of the series in time, and its weak stationarity, was often confirmed.

It was also established that even though some methods of prediction have a lower error value, such a prediction is less useful than another model with a higher error that more closely matches the curve of the real values over time. The latter models may be more successful with further application, as future attacks will likely not be constant either. The error of prediction in this experiment is highly dependent on the particular series, ranging from 20% to more than 1000%.

When it is considered that nearly all known approaches to time series prediction were applied, with many different settings, it is safe to conclude that the number of attacks is a rather hard series to predict, especially with an economic time series, for example. Another reason for the low achievement in the accuracy of the predictions, is due to the short length of the time series, not allowing for very accurate estimates of the parameters for the used models. Namely those using the seasonal character of the data. Despite this, the series' APNIC, LACNIC, RIPENCC, France or AS16276 achieved the most accurate predictions using the seasonal model SARIMA. What was also surprising was the very good results obtained by the simple Holt-Winters algorithm, achieving the lowest error with the APNIC and China series.

Despite the large error values of some models, the results of the analysis of this experiment can be considered successful, as they helped to reveal other potential areas that should be researched further.

7. Conclusion and Further Research

The paper shows the possibilities and reliability of predicting attacks on a honeynet based on real-world data. The prediction was analysed as the overall attacks, and based on the source of the attacks from specific geographic locations. From a usability point of view, it could provide an analyst with useful predictions and information. It can also provide valuable, directly applicable information to CSIRT teams, mainly at a national level. In most cases, it will provide at least a useful short term prediction of the trends of attacks, often providing accurate predictions. The most accurate predictions were achieved with individual countries used as the source of attacks. The predictions with RIRs as sources and for the overall number of attacks on the honeynet were also acceptably accurate. The predictions with autonomous systems as the source were the least accurate.

The results of the analysis show that even despite using multiple methods and calibrating them, it is impossible to reach acceptable accuracy for all observed aspects. In most cases, the prediction accuracy is acceptable, given the length of the time series used. The methods of prediction using a seasonal component of the time series increase their efficiency with the growing number of seasons they have at their disposal. In the end, it is safe to conclude that using a time series to predict future attacks on a honeynet has proven to be beneficial and, in some cases, effective.

Further research in this area will be focused on the application of soft-computing methods for the prediction of a time series in the area of cyber-security, such as with neural nets.

References

1. Spotzner, L., Honeypots: Tracking Hackers, Addison Wesley Longman Publishing Co., Inc., USA (2002)
2. Joshi, C. R. and Sardana, A., Honeypots A New Paradigm to Information Security, Science Publishers, USA (2011)
3. Provos N., Holz T., Virtual Honeypots: From Botnet Tracking to Intrusion Detection, Addison Wesley Professional, USA (2007).
4. Ligh Hale M., Adair S., Hartstein B., Matthew R. Malware Analyst's Cookbook and DVD - Tools and Techniques for Fighting Malicious Code, Wiley Publishing, Inc, USA (2011)
5. Grudziecki T., Jacewicz P., Juszczyk Ł., Kijewski P., Pawliński P. and ENISA editors, Proactive Detection of Security Incidents Honeypots, ENISA publication, Greece (2012)
6. Abbasi, F.H., Harris, R.J. Experiences with a generation III virtual honeynet, Australasian Telecommunication Networks and Applications Conference, ATNAC 2009 - Proceedings, art. no. 5464785 (2009)

7. Balas, E., Viecco, C., Towards a third generation data capture architecture for honeynets, Proceedings from the 6th Annual IEEE System, Man and Cybernetics Information Assurance Workshop, SMC 2005, art. no. 1495929, pp. 21-28 (2005)
8. Sokol P., Gajdoš A., Prediction of Attacks Against Honeynet Based on Time Series Modeling, Silhavy R., Silhavy P., Prokopova Z. (eds) Applied Computational Intelligence and Mathematical Methods. CoMeSySo 2017, Advances in Intelligent Systems and Computing, vol 662, Springer (2017)
9. Condon, E., He, A., Cukier, M., Analysis of computer security incident data using time series models, 19th International Symposium on Software Reliability Engineering, ISSRE 2008, pp. 77–86, IEEE (2008)
10. Werner, G., Yang, S., McConky, K., Time series forecasting of cyber attack intensity, Proceedings of the 12th Annual Conference on Cyber and Information Security Research, p. 18. ACM (2017)
11. Wei, M., Kim, K., Intrusion detection scheme using traffic prediction for wireless industrial networks. J. Commun. Netw. 14(3), 310–318 (2012)
12. Tang, M., Alazab, M., Luo, Y., Exploiting vulnerability disclosures: statistical framework and case study, Cybersecurity and Cyberforensics Conference (CCC), pp. 117–122. IEEE (2016)
13. Zhan, Z., Xu, M., Xu, S., Predicting cyber attack rates with extreme values, IEEE Trans. Inf. Forens. Secur. 10(8), 1666–1677 (2015)
14. Soldo, F., Le, A., Markopoulou, A., Blacklisting recommendation system: using spatio-temporal patterns to predict future attacks, IEEE J. Sel. Areas Commun, 29(7), 1423–1437 (2011)
15. Thonnard O. snf Marc D., A framework for attack patterns' discovery in honeynet data, Digital Investigation, Volume 5, Supplement, S128-S139, ISSN 1742-2876 (2008)
16. Husák M., Bartoš V., Sokol P., Gajdoš A., Predictive methods in cyber defense: Current experience and research challenges, Future Generation Computer Systems, Volume 115, 517-530 (2021)
17. Husák M., Komárková J., Bou-Harb E., Čeleda P., Survey of attack projection, prediction, and forecasting in cyber security, IEEE Commun. Surv. Tutor. 21 (1) 640–660 (2019)
18. Sun N., Zhang J., Rimba P., Gao S., Zhang L. Y., Xiang Y., Data-driven cybersecurity incident prediction: A survey, IEEE Commun. Surv. Tutor. 21 (2) 1744–1772 (2019)
19. Zhang K., Zhao F., Luo S., Xin Y., Zhu H., An intrusion action-based IDS alert correlation analysis and prediction framework, IEEE Access 7 150540–150551 (2019)
20. Sokol, P. and Zuzčák, M. and Sochor, T., Definition of attack in context of high level interaction honeypots, Advances in Intelligent Systems and Computing, vol. 349, pp. 155-164 (2015)
21. Sokol, P. and Zuzčák, M. and Sochor, T., Definition of attack in the context of low-level interaction server honeypots, Lecture Notes in Electrical Engineering, vol. 330, pp. 499-504 (2015)
22. Zuzčák M. and Bujok P., Causal analysis of attacks against honeypots based on properties of countries, IET Information Security (2019)
23. Box, G. and Jenkins, G., Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco (1970)
24. Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y.: Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J. Econom. 54(1–3), 159–178 (1992)
25. Holt, C. C., Forecasting seasonal and trends by eponentially weighted moving averages. Res. Mem. no. 52. Pittsburg: Carnagie Institute of Technology (1957)
26. Winters, P. R., Forecasting sales by exponentially weighted moving averages. Management Science, vol. 6, p. 324-342 (1960).
27. Engle, R. F., Autoregressive conditional heteroscedasticity with the estimates of the variance of United Kingdom inflations. Econometrica, vol. 50, p. 987-1007 (1982)

28. Efron B. Bootstrap Methods: Another Look at the Jackknife. In: Kotz S., Johnson N.L. (eds) Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY (1992)
29. Härdle W., Horowitz J., Kreiss J.-P., Bootstrap methods for time series, International Statistical Review 71, 435 – 459 (2003)
30. Chernick, M.R., LaBudde, R.A.: An Introduction to Bootstrap Methods With Applications to R. Wiley, Hoboken (2014)
31. Lahiri S. N., Resampling methods for dependent data, Springer-Verlag, New York (2003).

Matej Zuzčák earned his PhD in 2020 and he is currently working as a assistant professor and a researcher at the Department of Informatics and Computers of the Faculty of Science at University of Ostrava. His scientific research is focused mainly on honeypots, honeynets, network security, expert systems and data analysis. He has been the head of university CSIRT team - CSIRT OU since 2017. He is also a member of The Honeynet Project in Czech chapter.

Petr Bujok works as an Associate professor at the Department of Informatics and Computers of the Faculty of Science at University of Ostrava. His research area is mainly focused on the development and application of Evolutionary algorithms for global optimisation and applied statistics.

Received: July 15, 2020; Accepted: May 10, 2020.

Computer Science and Information Systems 18(4):1219–1238 https://doi.org/10.2298/CSIS210208036H

SimAndro-Plus: On Computing Similarity of Android Applications

Masoud Reyhani Hamedani[1] and Sang-Wook Kim[2,⋆]

[1] Department of Computer Science,
BK21PLUS Program for Advanced AI Research and Education,
Hanyang University, Seoul, Korea, 04763
[2] Department of Computer and Software, Hanyang University,
Seoul, Korea, 04763
{masoud,wook}@hanyang.ac.kr

Abstract. In this paper, we propose *SimAndro-Plus* as an *improved variant* of the state-of-the-art method, SimAndro, to compute the similarity of Android applications (apps) regarding their functionalities. SimAndro-Plus has two *major differences* with SimAndro: 1) it exploits *two* beneficial features to similarity computation, which are totally *disregarded* by SimAndro; 2) to compute the similarity score of an app-pair based on strings and package name features, SimAndro-Plus considers *not* only those terms co-appearing in both apps but *also* considers those terms appearing in one app while missing in the other one. The results of our extensive experiments with three *real-world* datasets and a dataset constructed by human experts demonstrate that 1) each of the two aforementioned differences is really *effective* to achieve better accuracy and 2) SimAndro-Plus *outperforms* SimAndro in similarity computation by 14% in average.

Keywords: android applications, apps data mining, feature extraction, API calls, manifest information, similarity computation

1. Introduction

Android applications (in short, apps) are rapidly growing in the number and variety [5] [17] distributed via the official Google Play Store[3] and other third-party stores such as Amazon App Store[4] and APKPure[5]. Google Play Store contains a *huge* number of apps divided into various categories such as game, communication, and business [11] [24]. As the number of apps in app stores increases dramatically, even if they are divided into various categories, smartphone users face a serious problem to find relevant apps providing their required functionalities [5] [13]. Therefore, there is an important *demand* for app search engines or recommender systems to alleviate this problem where employing an accurate *similarity method* is one of the most challenging issues [13] [15].

In the literature, some methods have been proposed for similarity computation of apps where we aim to find similar apps regarding their functionalities [4–7] [13] [22]. To do

⋆ Corresponding author

[3] http://play.google.com/apps

[4] https://www.amazon.com

[5] https://apkpure.com

this, proposed methods in references [4], [6], [22], SimApp [5], and DNADroid [7] extract the required features from the app stores; these feature might be *inaccurate*, *varied* in different app stores, *unavailable*, affected by *language barrier*, and *unappropriated* for similarity computation. Therefore, exploiting these features may lead us to *inaccurate* similarity computation [13]. On the contrary, SimAndro [13], an effective state-of-the-art method, exploits the features extracted (i.e., mined) from apps and the Android platform itself to compute the similarity of apps. The motivation behind SimAndro is that apps contain *helpful* and *unique* features for similarity computation such as API calls and manifest information that clearly capture the apps functionalities.

In this paper, we propose *SimAndro-Plus* as an *improved variant* of SimAndro to compute the similarity of apps *more* effectively. As SimAndro does, SimAndro-Plus performs feature extraction and similarity computation steps; however, it has *two major differences* with SimAndro in the *both* steps as follows. First, instead of API-method and manifest-complete, SimAndro-Plus exploits *two* new beneficial features to similarity computation named as *API-full-method* and *manifest-partial*, respectively. API-method *implicitly* capture the app's functionalities in some cases; for example, overloaded APIs are regarded as an *identical* feature, while they somehow perform different tasks. However, our API-full-method considers the API's signature (i.e., API's fully qualified name and its parameter list) as a feature, thereby capturing the app's functionalities *explicitly*. Manifest-complete considers the app components (i.e., extracted from the AndroidManifest.xml file) as part of the feature, which are *not* predefined entities in the Android platform and exploiting them in similarity computation may be *misleading*; for example, declaring an activity component by two different apps *cannot* imply the similarity between the two apps since these activity components can be implemented to perform totally different tasks in each of the two apps. However, SimAndro-Plus does *not* consider app components as part of the feature by exploiting the manifest-partial feature. Second, in computing the similarity score between two apps based on their corresponding strings and package name features, SimAndro considers *only* the terms co-appearing in both apps; however, it has been shown that in computing the similarity between two objects of terms (e.g., documents), the number of those terms appearing in one object while missing in the other one is important as well [16]. Therefore, by following [16], SimAndro-Plus considers those terms *appearing* in one app while *missing* in the other one along with those terms co-appearing in both apps. The results of our extensive experiments with three *real-world* datasets and a dataset constructed by *human experts* (i.e., authors) demonstrate that SimAndro-Plus *outperforms* SimAndro.

The contributions of this paper are as follows:

- We extract two new helpful features from apps.
- In computing the similarity based on strings and package name feature, we not only consider those terms appearing in both apps but also consider those terms appearing in only one of the apps.
- We verify the last two contributions help to improve the accuracy of original SimAndro.

The rest of the paper is organized as follows. In Section 2, we briefly explain the existing methods. In Section 3, we present our SimAndro-Plus and its two orthogonal steps. Section 4 explains our experimental setup and analyzes the results of our experiments. In Section 5, we conclude our paper.

2. Related Work

In this section, we explain SimAndro and other existing methods. However, since we mainly focus on SimAndro in this paper, the other methods are explained briefly; the complete explanations about them can be found at [13, Section 2].

Reference [4] proposes a method for app recommendation where the similarity score of an app-pair is computed based on their titles and user comments. Reference [22] proposes a method to invoke users for replacing an already installed app a with a new app b where the similarity score between a and b is computed by exploiting their descriptions. In SimApp [5], the similarity between two apps is computed individually based on multiple features such as description, rating, permissions, and size; then, the obtained individual similarity scores are combined into a single value as the final similarity score. In reference [6], a method is proposed for automatically tagging apps where the similarity score of an app-pair is computed as in SimApp. DNADroid [7] detects app cloning by computing the similarity between apps based on different features such as title, developer, and description. *All* of the aforementioned methods extract the required features from the *app stores*, which might incur the problems of being *inaccurate* (e.g., permission list), *varied* in different app stores (e.g., description and user comment), *unavailable* (e.g., user comment and rating), affected by *language barrier* (e.g., description and user comments), and *unappropriated* for similarity computation (e.g., size and rating); exploiting these features may lead us to inaccurate similarity computation. These methods highly depend on the human explanations and descriptions of apps and *neglect* the useful features that can be *mined* from apps themselves and the Android platform [13].

SimAndro [13] is an effective state-of-the-art method to compute the similarity of apps by exploiting features extracted (i.e., mined) from apps and the Android platform itself; it is an easy-to-understand and straightforward similarity method for apps that can be applied to a wide range of applications such as app search engines, app recommendation, and app clustering. The motivation behind SimAndro is that apps contain *helpful* and *unique* features for similarity computation that clearly capture the apps functionalities without depending on the human explanations or descriptions of apps. SimAndro performs the two orthogonal steps of feature extraction and similarity computation. In the former step, *API-methods*, *manifest-complete*, *strings*, and *package name* are extracted as four different features from the *classes.dex*, *AndroidManifest.xml*, *strings.xml*, and *AndroidManifest.xml* files, respectively. We note that a typical app is an archive file type called Android Package (APK); this file is easily extractable by any archiving software and contain different folders (e.g., assets, lib, and META-INF) and files (e.g., AndroidManifest.xml, classes.dex. and strings.xml) [9] [13]. In the latter step, four similarity scores of an app-pair (a, b) are calculated based on the aforementioned heterogeneous features *separately*. Then, by utilizing TreeRankSVM [1], the weighted linear combination of the above four scores is regarded as the *final* similarity score of (a, b).

3. Proposed Method

Figure 1 illustrates an overview of our SimAndro-Plus. The overall process in both feature extraction and similarity computation steps are *somehow* similar to the ones in SimAndro; however, in order to make the paper self-contained, we briefly explain the two steps in this section along with the two major differences between SimAndro-Plus and its predecessor.

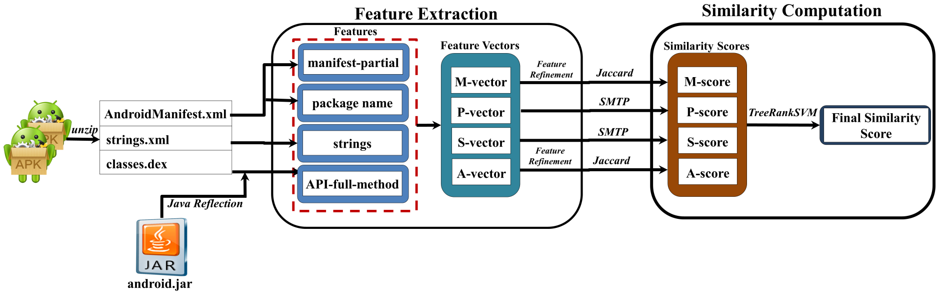


Fig. 1. An overview of SimAndro-Plus

3.1. Feature Extraction

In this step, we extract *API-full-method*, *manifest-partial*, *strings*, and *package name* from apps as four heterogeneous features where API-full-method and manifest-partial are two new features disregarded by SimAndro, while strings and package name are same as the ones exploited by SimAndro.

API-full-method Feature APIs in the Android platform are utilized by apps to interact with the underlying Android system and the device [8] [13]; for example, by calling the "android.os.Handler. removeMessages (int what)" API, an app removes pending messages with code "what" from the message queue. More specifically, API calls can clearly capture the app's behaviors and functionalities [12] [13] [23]. Therefore, *we extract the API calls as a feature to understand what operations an app executes*. SimAndro considers the API's fully qualified name as a feature called API-method (e.g., "android.os.Handler.removeMessages" for the above API). Let us consider the "android.os. Handler.removeMessages (int what, object obj)" API that removes pending messages with code "what" whose object is "obj" from the message queue. Although these two APIs are *different*, they are regarded *identical* by the API-method feature. To solve this problem, SimAndro-Plus exploits a new feature called *API-full-method* that considers the *API's signature* (i.e., API's fully qualified name and its parameter list) instead of only the fully qualified name. As an example, for the two aforementioned APIs, "android.os.Handler.removeMessages (I)"[6] and "android.os.Handler.removeMessages (I, L)" are considered as the API-full-method feature, respectively. API-full-method captures the apps functionalities *more* accurate that API-method since it considers the API's parameter list as the part of the feature. In Section 4.2, we show that API-full-method is *more* beneficial than API-method to similarity computation.

To extract the API-full-method feature, we utilize *both* APK file and Android platform as follows. First, we mine the DEX file via it's different sections such as the *header*, *method_ids*, *string_ids*, *type_ids*, *proto_ids*, and *data*. The method_ids section contains identifiers for all the app's methods; the string_ids section contains identifiers for all the strings (e.g., classes, methods, parameters, etc.) in the app; the type_ids section contains

[6] For simplicity, we use Dalvik symbols to represent parameters.

identifiers for all the types (classes and primitive types) defined by the app; proto_idx contains identifier for the return type and parameters of each method in the app; the header section defines the offset and the size of each of the aforementioned sections. Through the starting address of the method_ids section in the header, we read all entries in the section. Each entry in this section is a *data structure* that contains various kinds of information about a method including an index (*class_idx*) to an offset in the type_ids section, an index (*name_idx*) to an offset in the string_ids section, and an index (*proto_idx*) to an offset in the proto_ids section. The offset pointed by class_idx has an index to another offset in the string_ids section where we obtain the name of the method's owner class. We also extract the name of the method itself through name_idx. The offset pointed by proto_idx has an index to an offset in other list contains number of parameters and their types. For each entry in the method_ids section, we *concatenate* its class name, method name, and parameter list to construct a candidate API-full-method. Then, we apply the Java reflection to the "android.jar" file to obtain a list of all API descriptions in the Android platform; if a candidate API-full-method does *not* belong to this list, we *ignore* it.

Finally, based on the API-full-method feature, an app is represented as a *binary* vector, *A-vector*, where each dimension corresponds to a feature value and the content of a dimension indicates the presence (i.e., value as 1) or absence (i.e., value as 0) of its corresponding feature value in the app [19]. In order to clarify it, suppose that $\{a_1, a_2, ..., a_n\}$ is a set of n distinctive API-full-methods extracted from *all* the apps in a dataset; then, A-vector of app a is represented as $< v'_0, v'_1, ..., v'_{n-1} >$ with n dimensions where $v'_i = 1$ $(0 \leq i \leq n-1)$ if a contains the feature value a_{i+1}; otherwise $v'_i = 0$.

Manifest-partial Feature The AndroidManifest.xml file holds useful *meta* information (i.e., manifest information) about an app such as permissions, hardware/software components, app components (i.e., activity, service, broadcast receiver, and content provider), and intent filters (i.e., action and category); these information supports *both* installation and execution of the app [8] [13]. The permissions are *required* to perform critical tasks such as network access, the hardware/software components indicate either an essential or optional hardware (e.g., GPS) and software (e.g., VoIP) components that the app requires, the activity component implements a task with UI (user interface), the service component implements a background task without UI, the broadcast receiver component enables the app to receive events broadcast by the Android system or other apps, the content provider component supplies data access interface, and intent filters facilitates communication between the app's components and also between different apps. This information *can* capture the app's behaviors and functionalities as API calls do [2] [12] [13]; thus, we extract the manifest information as a feature for similarity computation.

SimAndro considers all the aforementioned information *including* the four app components as a feature called manifest-complete. An app component is defined by *developers* as a subclass of its specific standard class in the Android platform to implement the app's *specific* functionalities. For example, although two activity components c_1 and c_2 from two different apps *a* and *b* are both defined as subclasses of the "android.app.Activity" class, they are developed with their own arbitrary names and *under* specific functionalities of *a* and *b*, respectively; even if c_1 and c_2 are both activity components, they may *not* implement similar functionalities. More specifically, contrary to permissions, hardware/software components, action, and category, app components are *not* predefined entities in the Android platform and are developed independently for each app under the app's specifications, thereby considering them as a feature may provide us *inaccurate* similarity scores. To solve this problem, SimAndro-Plus exploits a new feature called *manifest-partial* where *only* permissions, hardware/software components, action, and category are considered.

Based on the manifest-partial feature, an app is represented as a binary vector, *M-vector*, which is similar to *A-vector*. In Section 4.2, we show that manifest-partial is *more* beneficial than manifest-complete to similarity computation.

Strings and Package Name Features Furthermore, we consider strings and package name as two other features as SimAndro does. The strings.xml file is a single reference for various strings in an app where each string has a *name attribute* as its unique identifier [13, Fig. 3]; we extract both the string and its name attribute since the name attribute also represents some semantic information about the app. As an example, the following line in the strings.xml file of "Weather Forecast", a free app for weather forecasting, defines a string:

$$< string\ name = "weather\_sunny" > Sunny < /string >$$

The package name located in the AndroidManifest.xml file is a *unique* identifier for the app and follows Java package naming convention. It is a combination of multiple terms (i.e., simple term or compound one) concatenated by dot and normally provides us abstract information about the app's functionalities; for example, the "weather.widget. weatherforecast" is the package name of the "Weather Forecast" app. For each of these two features, we remove non-alphabetical characters, split compound strings (e.g., weatherforecast), remove stop words, perform stemming, and calculate the TF-IDF score [19] for each term. Finally, based on strings and package name features, an app is represented as two *non-binary* vectors *S-vector* and *P-vector*, respectively, where each dimension corresponds to a term and the content of the dimension is the TF-IDF score of the term.

Feature Refinement In order to obtain better accuracy in similarity computation, we need to perform a feature refinement. The reason is that some of the feature values are widely used in a large number of apps *regardless* of their functionalities, thereby exploiting them in similarity computation leads to *inaccurate* similarity scores. An as examples, consider the two following cases; the "android.os.Message.sendToTarget()" API used by an app to send a message to a specific handler is invoked in more than 90% of apps in our datasets, and the "INTERNET" permission allowing the Internet access is requested by more than 95% of apps in our datasets. We apply a feature refinement similar to the one in SimAndro to our new features, API-full-method and manifest-partial, as follows.

To refine the API-full-method feature with a dataset, we consider a *threshold*, *T*, from 10% to 70% of the dataset size in step of 10% and a feature value is *neglected* if the number of apps in the dataset containing it is *higher* than *T*; in other words, we do *not* consider those feature values that are common among more than *T* of apps. Then, we compute the apps similarity based on *only* the API-full-method feature refined with each value of *T* and compare the accuracy of these seven different cases; the *T* value of the case providing us the better accuracy is selected as the best value of *T*. To refine the manifest-partial feature, we perform the same process.

3.2. Similarity Computation

As explained before, an app is represented by four different vectors as A-vector, M-vector, S-vector, and P-vector corresponding to its API-full-method, manifest-partial, strings, and package name features, respectively. As in SimAndro, to calculate the similarity score of an app-pair (a, b) based on the API-full-method feature, $A\text{-}score(a, b)$, and the manifest-partial feature, $M\text{-}score(a, b)$, we apply Jaccard Coefficient (Jaccard) [19] to corresponding A-vectors and M-vectors of a and b, respectively. In the case of $A\text{-}score(a, b)$, it is calculated as follows:

$$A\text{-}score(a,b) = \frac{\Sigma_i A_i^a \cdot A_i^b}{\Sigma_i A_i^a + \Sigma_i A_i^b - \Sigma_i A_i^a \cdot A_i^b} \tag{1}$$

where A_i^a and A_i^b denote the contents (i.e., 0 or 1) of the i^{th} dimensions in A-vector of *a* and A-vector of *b*, respectively.

We note that $M\text{-}score(a, b)$ is also calculated in the same way. We employed Jaccard to calculate these two scores since in the literature, it is a well-known similarity measure widely used to calculate the similarity of binary vectors (i.e., sets) in various topics such as image segmentation [10], document summarization [20], and similarity computation [14].

On contrary to SimAndro, to compute the similarity score of an app-pair (a, b) based on the strings feature, $S\text{-}score(a, b)$, and the package name feature, $P\text{-}score(a, b)$, we apply SMTP (similarity measure for text processing) [16] *instead of* Cosine [19] to corresponding S-vectors and P-vectors of a and b, respectively, for the following reasons. S-vector and P-vector are *non-binary* vectors where each dimension corresponds to a term (i.e., a feature value) and the content of the dimension is set as its weight (i.e., the TF-IDF score). To calculate the similarity between two non-binary vectors, *not* only the proximity between the weights of co-appearing terms in both vectors but *also* the number of those terms *appearing* in one vector while *missing* in the other one is important as well. More specifically, as have been shown in [16], 1) the presence or absence of a term is more important in similarity computation than the difference between the weights of a co-appearing term in both vectors; 2) the similarity score should increase when the difference between the weights of a co-appearing term decreases; 3) the similarity score should decrease when the number of terms appearing in one vector but missing in the other one increases. Let us consider three sample vectors *i*=$<2, 0, 3, 0>$, *j*=$<2, 1, 3, 1>$, and *k*=$< 2, 4, 2, 2 >$. Although there are two missing terms in *i*, the Cosine similarity score between *i* and *j* (i.e., 0.93) is *higher* than that between *j* and *k* (i.e., 0.78) where there is *not* any missing terms; Cosine does *not* acknowledge the third aforementioned case.

SMTP is an effective measure that considers *all* the three aforementioned cases in similarity computation. To calculate $S\text{-}score(a, b)$, we apply SMTP to the corresponding S-vectors of a and b as follows:

$$S\text{-}score(a,b) = \frac{\frac{\Sigma_i N_*(S_i^a, S_i^b)}{\Sigma_i N_\cup(S_i^a, S_i^b)} + \lambda}{1+\lambda}, \tag{2}$$

$$N_*(S_i^a, S_i^b) = \begin{cases} 0.5\cdot\left(1+exp\Big(-(\frac{S_i^a-S_i^b}{\sigma_i})^2\Big)\right), & S_i^a, S_i^b \neq 0, \sigma_i \neq 0 \\ 0.5, & S_i^a, S_i^b \neq 0, \sigma_i = 0 \\ 0, & S_i^a, S_i^b = 0 \\ -\lambda, & otherwise \end{cases}$$

$$N_\cup(S_i^a, S_i^b) = \begin{cases} 0, & S_i^a, S_i^b = 0 \\ 1, & otherwise \end{cases}$$

where S_i^a and S_i^b denote the weights of the i^{th} terms in S-vector of *a* and S-vector of *b*, respectively. λ denotes a constant and σ_i does the standard deviation of all non-zero weights of the i^{th} term in the dataset. Note that we regard an extra condition "$S_i^a, S_i^b \neq 0, \sigma_i = 0$", which is *not* considered in the SMTP original formulation; if $\sigma_i = 0$, the SMTP definition is incorrect and the similarity score is *undefined* since the division by zero happens. $N_\cup$ sums up the number of terms contributing in similarity computation.

The following three cases are considered through the four conditions in calculating N_*: 1) those terms co-appearing in *both* apps contribute *positively* to the similarity computation where the *amount* of contribution depends on the proximity of their corresponding weights in two apps and their standard deviations in the dataset (i.e., if $S_i^a, S_i^b \neq 0, \sigma_i \neq 0$). when the standard deviation is zero, the amount of contributions is less than the former case (i.e., if $S_i^a, S_i^b \neq 0, \sigma_i = 0$). 2) Those terms missing in *both* apps, do *not* contribute to the similarity computation (i.e., if $S_i^a, S_i^b = 0$). 3) Those terms *appearing* in one app but *missing* in the other one *adversely* affect the similarity score (i.e., fourth condition).

The similarity score of (a, b) based on their package name features, $P\text{-}score(a, b)$, is also calculated in the same way as $S\text{-}score(a, b)$. In Section 4.2, we show that SMTP is *more* beneficial than Cosine to similarity computation of apps. Finally, as in SimAndro, we apply a *weighted* linear combination to combine the four scores into a *single* value as the *final* similarity score of (a, b) as follows:

$$\begin{aligned} S(a, b) = w_1\cdot A\text{-}score(a, b) + w_2\cdot M\text{-}score(a, b) \\ +w_3\cdot P\text{-}score(a, b) + w_4\cdot S\text{-}score(a, b) \end{aligned} \tag{3}$$

where w_1, w_2, w_3, and w_4 are weights to control the degree of *importance* of each score in the combination. We automatically find the best value of these four weights by utilizing TreeRankSVM [1] as a machine learning technique; more details can be found in [13, Section 3.3].

It has been shown that instead of considering all the above scores equally significant and simply summing up them into a single value as the final similarity score, applying a weighted linear combination to combine them contributes to obtain better accuracy in similarity computation [13]. We note that it also could be an option to simply combine our four heterogeneous features into a single one (i.e., each app is represented by a single binary vector) and then compute apps similarity based on this single feature; however, it has been shown that considering each of the four heterogeneous features separately in similarity computation is beneficial to obtain better accuracy [13].

3.3. Overall Process: Review

In this section, we present a *simple* review of the overall process required to compute the similarity between two apps as follows.

Feature extraction and refinement First, we use an archiving software (e.g., ark[7]) to unzip all the apps in the dataset. Then, we extract the features for *each* app as follows. We mine the app's classes.dex file through its different sections to extract API-full-method (in the case of SimAndro-Plus) and API-method (in the case of SimAndro); the mining process of the classes.dex file is described in Section 3.1 and [13, Section 3.2.1] in detail. As an example, for the "WhatsApp Messenger" app, we extracted 5,398 feature values for API-full-method and 5,301 features values for API-method. Note that the API-full-method feature has *more* values since it considers the API's parameter list as part of the feature. As an example, "WhatsApp Messenger" calls both of the two following APIs: the "android.media.MediaCodec.releaseOutputBuffer (int index, boolean render)" API is called to return an unnecessary buffer to the codec or to render it on the output surface, while the "android.media.MediaCodec.releaseOutputBuffer (int index, long renderTimestampNs)" API is called to update the surface timestamp of an unnecessary buffer and return it to the codec to render it on the output surface; API-full-method considers *two various* feature values for the above two APIs as "android.media.MediaCodec.releaseOutputBuffer (I, Z)" and "android.media.MediaCodec.releaseOutputBuffer (I, J)", respectively; however, API-method considers an *identical* feature value for both cases as "android.-media.MediaCodec.releaseOutputBuffer". Next, we apply the feature refinement to both API-full-method and API-method features as explained in Section 2.1 where the best values of T are 30% (i.e., refer to Table 2) and 20% (i.e., refer to [13, Table 4]), respectively, with the google dataset. As an example, we have 1,907 and 1,561 feature values for API-full-method and API-method, respectively, with "WhatsApp Messenger" after refining them as its *final* features.

Now, we exploit the AndroidManifest.xml file to extract manifest-partial (in the case of SimAndro-Plus) and manifest-complete (in the case of SimAndro); since this file is in the XML format, the feature extraction is straightforward and not tedious on contrary to that of the classes.dex file. For example, for the "WhatsApp Messenger" app, we extracted 85 and 97 values for manifest-partial and manifest-complete features, respectively. Note that the manifest-partial feature has *less* number of values since it does *not* take into account the app components (i.e., activity, service, broadcast receiver, and content provider). Now, we apply the feature refinement to both manifest-partial and manifest-complete features where the best values of T are 30% (i.e., refer to Table 2) and 20% (i.e., refer to [13, Table 6]), respectively, with the google dataset. As an example, we have 64 and 74 values for manifest-partial and manifest-complete features, respectively, with "WhatsApp Messenger" after refining them as its *final* features.

Next, we extract the package name (e.g., "com.whatsapp" for our sample app), decompound it into its constituent terms (e.g., "com whats app" for above case) by utilizing the Levenshtein algorithms [3], remove non-alphabetical characters and stop words, perform stemming on the terms, and measure the TF-IDF score of each term to obtain the package name feature. Finally, we extract the name attributes and their unique identifiers from the strings.xml file, remove non-alphabetical characters, split the strings, remove stop words including the Java reserved keywords as well, perform stemming on the remaining terms, and calculate the TF-IDF score of each term to obtain the strings feature.

[7] `https://apps.kde.org/ark/`

Automatic weight tuning Each app is represented by four vectors as A-vector, M-vector, S-vector, and P-vector; in the case of SimAndro-Plus, these vectors correspond to the API-full-method, manifest-partial, strings, and package name features of the app, respectively, while in the case of SimAndro, they correspond to the API-method, manifest-complete, strings, and package name features, respectively. Now, we utilize TreeRankSVM to find the best values of w_1, w_2, w_3, and w_4 in Equation (3) automatically as follows; these values are later used to compute the similarity score of any app-pairs. We randomly choose 75% of the apps in the dataset to make a training set where each of the chosen apps is regarded as a *query* app. For *each possible* app-pair (a, q) regarding to a query app q, we make a *hyperplane vector* (see [13, Section 3.3] for more detail) as follows:

$$\{r, qid, A\text{-}score(a, q), M\text{-}score(a, q), P\text{-}score(a, q), S\text{-}score(a, q)\} \tag{4}$$

when r is set as 1 if a is relevant to q (i.e., a belongs to the same category of q), otherwise 0 and qid is a real number started from 1 denoting a query number. For both SimAndro-Plus and SimAndro, $A\text{-}score(a, q)$ and $M\text{-}score(a, q)$ are calculated by applying Jaccard to the corresponding A-vectors and M-vectors of a and q, respectively. For SimAndro-Plus, $P\text{-}score(a, q)$ and $S\text{-}score(a, q)$ are calculated by applying SMTP to the corresponding P-vectors and S-vectors of a and q, respectively; in these two cases, for SimAndro, Cosine is utilized instead of SMTP.

Similarity computation Let us consider two apps a as "WhatsApp Messenger" and b as "TalkU". To compute the similarity score of app-pair (a, b), SimAndro-Plus employs Jaccard to calculate $A\text{-}score(a, b)$ and $M\text{-}score(a, b)$, employs SMTP to calculate $P\text{-}score(a, b)$ and $S\text{-}score(a, b)$, and finally applies the best values of w_1, w_2, w_3, and w_4 (i.e., obtained in the previous step) to Equation (3) to compute $S(a, b)$. SimAndro performs the same process; however, 1) $A\text{-}score(a, b)$ and $M\text{-}score(a, b)$ are calculated based on API-method and manifest-complete, respectively; 2) $P\text{-}score(a, b)$ and $S\text{-}score(a, b)$ are calculated by employing Cosine; 3) consequently, the best values of w_1, w_2, w_3, and w_4 are also obtained by a separate automatic weight tuning than the one performed with SimAndro-Plus.

To compute the similarity score between a *new* app and the existing ones in the dataset, we utilize the already identified values of w_1, w_2, w_3, and w_4. To update these values, we can follow some strategies; for example, if the number of new apps added to the dataset is 25% of the *original* dataset size (i.e., *identical* to our test set for the *last* automatic weight tuning), we perform a *new* automatic weight tuning on the dataset.

4. Evaluation

In this section, we carefully evaluate the effectiveness of our two contributions (i.e., exploiting the two new features and applying SMTP instead of Cosine) and compare the accuracy of SimAndro-Plus with that of SimAndro.

4.1. Experimental Setup

In order to conduct a fair evaluation, we employed the *same* datasets with SimAndro; *google*, *apkpure*, and *amazon* are three *real-world* datasets constructed based on the data

Table 1. Statistics of our datasets

	google	apkpure	amazon	manual
# apps	8903	11068	20570	444
# categories	74	43	204	37

obtained by crawling Google Play Store, APKPure, and Amazon App Store, respectively. We constructed the *manual* dataset by selecting few apps from the three real-world datasets and carefully dividing them into various categories based on their functionalities. Table 1 shows the statistics of our datasets.

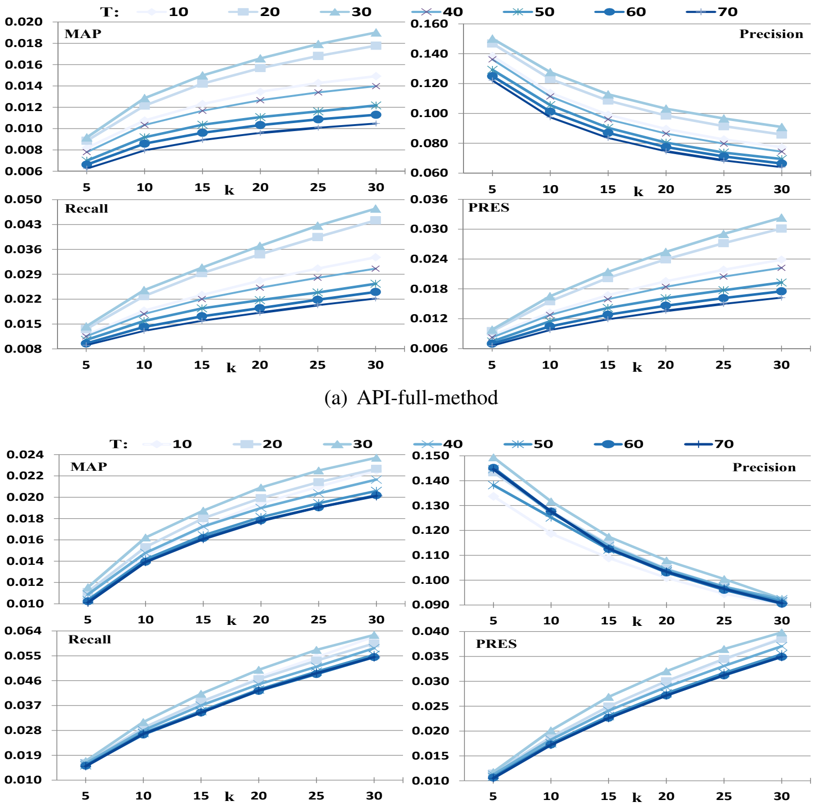
For the manual dataset, we can regard the categories as a ground truth set since the precise categorization is performed by humans expert (i.e., authors). For real-world datasets, their original categories are regarded as the ground truth sets since it is difficult and time-consuming to categorize them by humans expert (i.e., performing user studies); however, in order to conduct accurate and reliable evaluations, we consider a *fine-grained* categorization in our real-world datasets. For example, in our google dataset, the "Tools" category contains six sub-categories as "Alarm", "Flashlight", "Calculator", "Input", "Wi-Fi", and "Recommended"; instead of considering *all* the apps in these six sub-categories under a single category as "Tools", we consider these sub-categories as six *distinct main* categories as "Tools_Alarm", "Tools_Flashlight", "Tools_Calculator", "Tools_Input", "Tools_Wi-Fi", and "Tools_Recommended".

To evaluate the effectiveness, MAP, precision, recall [19], and PRES [18] are utilized as our evaluation metrics. In Equation (2), we set the value of λ as 1 by following [16].

4.2. Results and Analyses

In this section, we refine our new features, compare the effectiveness of applying API-full-method, manifest-partial, and SMTP to similarity computation with those of API-method, manifest-complete, and Cosine, respectively. Finally, we compare the accuracy of SimAndro-Plus with that of SimAndro.

Feature Refinement As explained in Section 3.1, we perform a feature refinement for API-full-method and manifest-partial features with our four datasets through the same process as in SimAndro. Figure 2(a) illustrates the result of our feature refinement for API-full-method with the google dataset on top *k* (k=5, 10, 15, 20, 25, 30) results; in the top of the figure, different values of T and their corresponding line patterns are shown (e.g., $T = 30$ and $T = 60$ are represented with triangle and circle marked lines, respectively). As shown in Figure 2(a), the best accuracy in terms of MAP, precision, recall, and PRES is observed when the value of T is set as 30% *regardless* of *k*; by setting T to smaller values than 30% (i.e., 10% and 20%) or to larger values than 30% (i.e., 40%, 50%, 60%, and 70%), we would get lower accuracy. Figure 2(b) illustrates the result of the feature refinement for the manifest-partial feature with the google dataset on top *k* results where the best accuracy is observed when the value of $T = 30\%$ regardless of *k*. Table 2 summarizes the complete results of the feature refinement with all datasets.



(a) API-full-method

(b) Manifest-partial

Fig. 2. Feature refinement with the google dataset.

Table 2. Results of feature refinement

	google	apkpure	amazon	manual
API-full-method	30%	30%	20%	20%
manifest-partial	30%	30%	40%	40%

Effectiveness Comparison of API-full-method and API-method As explained in Section 3.1, we exploit API-full-method instead of API-method, which is one of the major differences between SimAndro-Plus and SimAndro. Now, we compare the effectiveness of API-full-method with that of API-method in similarity computation with our four datasets as follows. For each dataset, we employ the best values of T for API-full-method from Table 2 and for API-method from [13, Table 4]; for example, with the google dataset, we set the best value of T as 30% and 20% for API-full-method and API-method, respectively. Then, with each dataset, we apply Jaccard to compute the similarity of apps by exploiting *only* API-full-method and API-method features *separately*; in other words, we do *not* consider the other three features in similarity computation. Finally, we compare the results of these two similarity computations for each dataset where for simplicity,

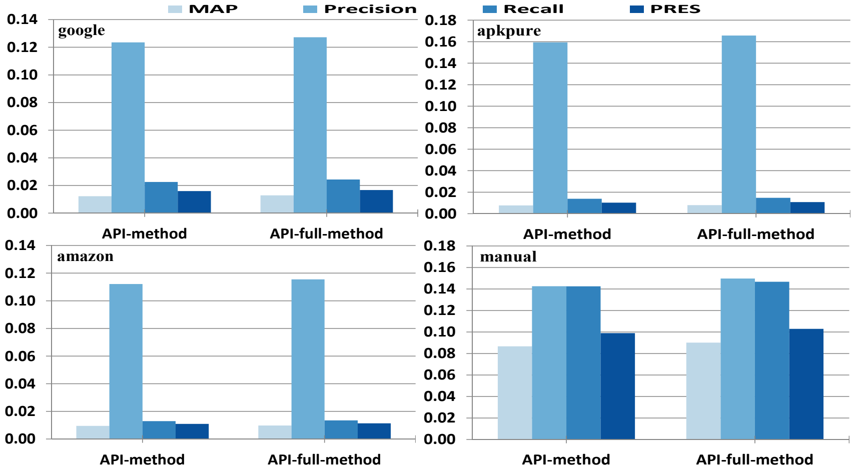


Fig. 3. Accuracy of API-full-method and API-method.

Table 3. Accuracy improvements(%) by API-full-method over API-method

	MAP	precision	recall	PRES
google	5	3	7	5
apkpure	4	4	6	7
amazon	3	3	4	4
manual	4	5	3	4

the effectiveness is considered as the *average* of MAP, precision, recall, and PRES on different values of k[8]. Figure 3 shows the results of this comparison; with *all* datasets, API-full-method shows *better* accuracy in terms of MAP, precision, recall, and PRES since it captures the apps functionalities *more* accurate than API-method by considering the API's signature, while API-method considers only the API's fully qualified name and neglects its parameter list. Table 3 represents the *percentage* of improvements in accuracy obtained by API-full-method over API-method with each dataset.

Effectiveness Comparison of Manifest-partial and Manifest-complete As explained in Section 3.1, SimAndro-Plus exploits the manifest-partial feature, while SimAndro exploits manifest-complete; this is another major difference between SimAndro-Plus and SimAndro. Now, we compare the effectiveness of these two features in similarity computation with our four datasets as follows. We employ the best values of T for manifest-partial from Table 2 and for manifest-complete from [13, Table 6] regarding to the target dataset; for example, with the apkpure dataset, we set the best value of T as 30% and 20% for manifest-partial and manifest-complete, respectively. Then, with each dataset, we apply Jaccard to compute the similarity of apps by exploiting *only* manifest-partial and manifest-complete features separately. Finally, we compare the results of these two

[8] As an example, we compute MAP for k=5, 10, 15, 20, 25, 30; then, the average of these six values is considered as MAP.

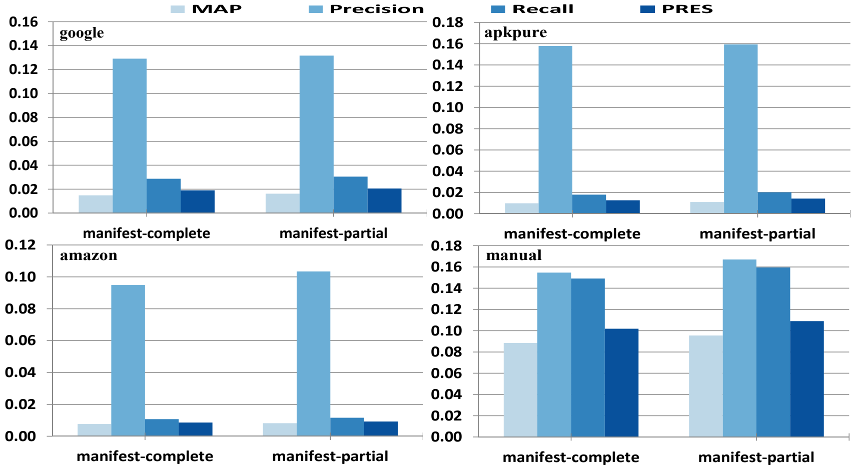


Fig. 4. Accuracy of manifest-partial and manifest-complete.

Table 4. Accuracy improvements(%) by manifest-partial over manifest-complete

	MAP	precision	recall	PRES
google	9	2	6	8
apkpure	11	1	12	12
amazon	7	9	8	7
manual	7	8	7	7

similarity computations for each dataset. Figure 4 illustrates the results of this comparison; with *all* datasets, manifest-partial shows *better* accuracy in similarity computation than manifest-complete in terms of all evaluation metrics. The reason is that manifest-complete considers app components in similarity computation; app components are *not* predefined entities and are developed independently in an app by developers under their own arbitrary names and functionalities, thereby considering them in similarity computation leads to inaccurate similarity scores. Table 4 represents the percentage of improvements in accuracy obtained by manifest-partial over manifest-complete with each dataset.

Effectiveness Comparison of SMTP and Cosine As explained in Section 3.2, SimAndro-Plus applies SMTP instead of Cosine to compute the similarity between two apps based on their strings and package name features. We compare the effectiveness of these two measures in similarity computation as follows. For each dataset, we compute the similarity of apps by applying Cosine and SMTP to *only* each of strings and package name features separately (i.e., four different cases); then for each feature, we compare the results of two similarity computations obtained by employing SMTP and Cosine. Figure 5 illustrates the results of this comparison; in the case of *both* strings and package name features, with *all* datasets, SMTP shows *better* accuracy than Cosine. The reason is that, on contrary to Cosine, SMTP considers not only the terms (i.e., feature values) co-appearing in both apps but also takes into account those terms appearing in one app while missing in the other

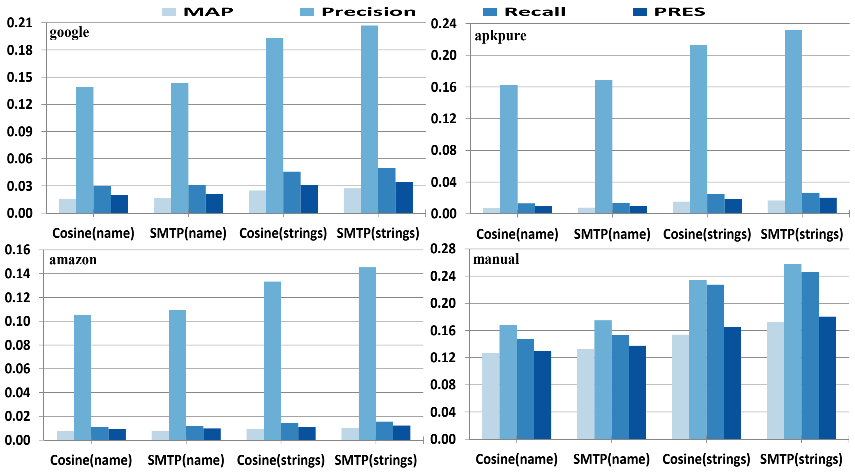


Fig. 5. Accuracy of SMTP and Cosine.

Table 5. Accuracy improvements(%) by SMTP over Cosine

	strings				package name			
	MAP	precision	recall	PRES	MAP	precision	recall	PRES
google	10	7	9	11	4	3	4	5
apkpure	10	9	7	10	4	4	5	3
amazon	9	9	8	9	3	4	4	5
manual	12	10	8	9	5	4	4	6

one. Table 5 represents the percentage of improvements in accuracy obtained by SMTP over Cosine with each dataset for both features; as observed in the table, SMTP shows higher improvements over Cosine with the strings feature than those with the package name feature since the latter feature for an app contains very few number of terms than the former one.

Effectiveness Comparison of SimAndro-Plus and SimAndro As shown in the last three sub-sections, considering API-full-method and manifest-partial as new features instead of API-method and manifest-complete, respectively, are effective; also, applying SMTP instead of Cosine to strings and package name features provides us better accuracy. These results imply that our *both* contributions are beneficial to similarity computation. As shown in reference [13], SimAndro outperforms existing methods proposed in references [5], [4], [22], [6], and [7]; therefore, here, we only compare the accuracy of SimAndro-Plus with that of SimAndro. More specifically, SimAndro-plus exploits API-full-method, manifest-partial, strings, and package name as four features, and applies Jaccard to the first two features and SMTP to the last two ones, while SimAndro exploits API-method, manifest-complete, strings, and package name as four features, and applies Jaccard to the first two features and Cosine to the last two ones. Figure 6 illustrates the results of this comparison with the four datasets; SimAndro-Plus *outperforms* SimAndro in terms of MAP, precision, recall, and PRES with *all* datasets. Table 6 represents the per-

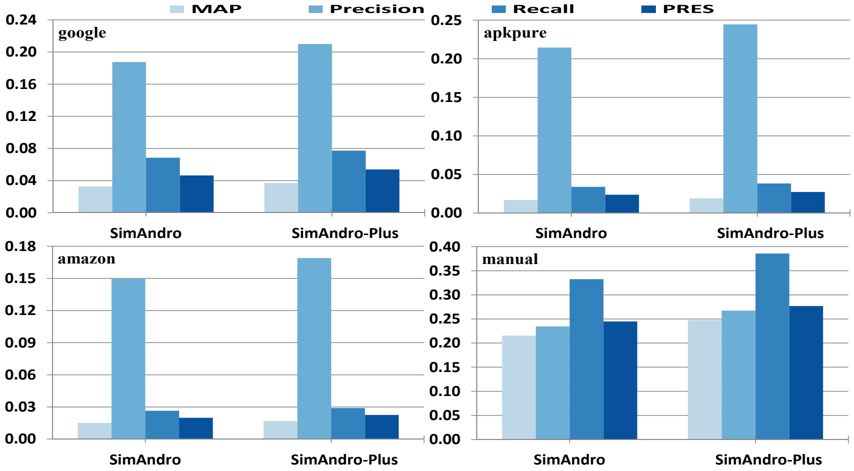


Fig. 6. Accuracy comparison between SimAndro-Plus and SimAndro.

Table 6. Accuracy improvements(%) by SimAndro-Plus over SimAndro

	MAP	precision	recall	PRES
google	14	12	13	16
apkpure	13	14	13	15
amazon	11	13	10	13
manual	15	14	16	13

centage of improvements in accuracy obtained by SimAndro-Plus over SimAndro with each dataset; in average over all datasets, SimAndro-Plus outperforms its predecessor by 14%.

As another evaluation, we perform the same queries in reference [13] by employing SimAndro-Plus and compare their results with those of SimAndro as follows. We consider two well-known apps in the google dataset as “WhatsApp Messenger” with the package name “com.whatsapp” and “Opera Browser” with the package name “com.opera.browser” from categories “Social_Messenger” and “Communication_WebBrowser”, respectively. Then, we find out the 10 most similar apps to each of these query apps (i.e., result sets) by applying SimAndro-Plus as the similarity method. Table 7 shows the results where the Relevant column contains ✓sign if the retrieved app is in the same category as the query; otherwise contains ✗ sign. Table 8 borrowed from reference [13] shows the results of the same queries with SimAndro. As shown in both tables, some apps are repeated under different signs in the Relevant column; the reason is that the google dataset assigns multiple categories to some apps. As an example, in the top result set of Table 7, the “Viber” app[9] with the package name “com.viber.voip” is repeated three times where it is marked as *relevant* in rank 8 since it belongs to the same category as the query, and it is marked as *irrelevant* in ranks 9 and 10 since it belongs to other categories than the query’s category.

[9] Viber is a cross-platform voice over IP and instant messaging software application provided by Japanese multinational company Rakuten.

Table 7. Result sets obtained by SimAndro-Plus for sample queries

	Rank	Package Name	Category	Relevant
WhatsApp Messenger	1	me.talkyou.app.im	Social_Messenger	✓
	2	me.talkyou.app.im	Communication_Message	✗
	3	kik.android	Social_Messenger	✓
	4	com.bbm	Social_Messenger	✓
	5	com.bbm	Communication_MovieChatting	✗
	6	me.dingtone.app.im	Communication_Message	✗
	7	me.dingtone.app.im	Social_Messenger	✓
	8	com.viber.voip	Social_Messenger	✓
	9	com.viber.voip	Communication_MovieChatting	✗
	10	com.viber.voip	Communication_Message	✗
Opera Browser	1	com.opera.mini.native	Communication_WebBrowser	✓
	2	com.apusapps.browser	Communication_WebBrowser	✓
	3	com.superapps.browser	Personalization	✗
	4	com.fsecure.ms.dc	Tool_Recommended	✗
	5	com.explore.web.browser	Social	✗
	6	com.explore.web.browser	Communication_WebBrowser	✓
	7	com.idotools.browser	Comics	✗
	8	mobi.mgeek.TunnyBrowser	Communication_WebBrowser	✓
	9	org.mozilla.firefox	Communication_WebBrowser	✓
	10	com.chrome.beta	Productivity	✗

As observed by comparing tables 7 and 8, SimAndro-Plus provides us *more* accurate results than SimAndro with the *both* queries. In the case of the first query (i.e., "WhatsApp Messenger") in Table 7, the "Kik Messenger" app[10] with the package name "kik.android" in rank 3 and the "Viber" app in ranks 8, 9, and 10 are both messenger apps as "WhatsApp Messenger", while they are absent in the result set obtained by SimAndro in Table 8. In the case of the second query (i.e., "Opera Browser") in Table 7, the "Firefox Browser" app with the package name "org.mozilla.firefox" in rank 9 is a web browser as "Opera Browser", while it is absent in the result set obtained by SimAndro in Table 8. More specifically, SimAndro-Plus fetches five similar apps for the both first and second query apps, while SimAndro does three and four similar apps for the first and second query apps, receptively.

5. Conclusions

In this paper, we proposed SimAndro-Plus to effectively compute the similarity of apps. SimAndro-Plus is an improved variant of SimAndro, the state-of-the-art method; however, it has two following major differences with SimAndro. First, SimAndro-Plus ex-

[10] Kik is a freeware instant messaging mobile app from the Canadian company Kik Interactive.

Table 8. Result sets obtained by SimAndro for sample queries

	Rank	Package Name	Category	Relevant
WhatsApp Messenger	1	net.mobileinnova.whatsmon	Tool_Recommended	✗
	2	me.talkyou.app.im	Social_Messenger	✓
	3	me.talkyou.app.im	Communication_Message	✗
	4	com.bbm	Social_Messenger	✓
	5	com.bbm	Communication_MovieChatting	✗
	6	me.dingtone.app.im	Communication_Message	✗
	7	me.dingtone.app.im	Social_Messenger	✓
	8	com.contapps.android	Communication_PhoneNumberBlocking	✗
	9	com.bsb.hike	Social	✗
	10	com.popularapp.fakecall	Productivity	✗
Opera Browser	1	com.opera.mini.native	Communication_WebBrowser	✓
	2	com.apusapps.browser	Communication_WebBrowser	✓
	3	com.fsecure.ms.dc	Tool_Recommended	✗
	4	com.superapps.browser	Personalization	✗
	5	com.explore.web.browser	Social	✗
	6	com.explore.web.browser	Communication_WebBrowser	✓
	7	com.idotools.browser	Comics	✗
	8	nh.smart.opensign	Finance	✗
	9	mobi.mgeek.TunnyBrowser	Communication_WebBrowser	✓
	10	com.chrome.beta	Productivity	✗

ploits two new features as API-full-method and manifest-partial, which are completely disregarded by SimAndro. Second, in similarity computation based on strings and package name features, SimAndro-Plus considers those terms appearing in one app but missing in the other one along with those terms appearing in both apps by employing the SMTP measure instead of Cosine. The results of our extensive experiments with four datasets of apps demonstrated that 1) the both new features are beneficial to similarity computation, 2) employing SMTP provides us better accuracy than Cosine, 3) SimAndro-Plus outperforms SimAndro.

As a future research direction, we plan to investigate the effectiveness of applying SimAndro-Plus to the app recommendation systems. SimAndro-Plus can be regarded as a reasonable solution to address the *item cold start problem* [21] in app recommendation where new released apps (i.e., items) with *no/few* related information in the app store cannot be recommended to the users. The reason is that SimAndro-Plus compute the similarity between apps only based on the features extracted from apps themselves.

Acknowledgments. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-01373, Artificial Intelligence Graduate School Program, Hanyang University), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1A2B5B03001960), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2018R1A5A7059549).

References

1. Airola, A., Pahikkala, T., Salakoski, T.: Training linear ranking svms in linearithmic time using redblack trees. Pattern Recognition Letters 32(9), 1328–1336 (Jul 2011)
2. Arp, D., Spreitzenbarth, M., Gascon, H., Rieck, K.: Drebin: Effective and explainable detection of android malware in your pocket. In: Proc. of NDSS, pp. 1–12 (2014)
3. Backurs, A., Indyk, P.: Edit distance cannot be computed in strongly subquadratic time (unless seth is false). In: Proceedings of the 47th Annual ACM Symposium on Theory of Computing, pp. 51–58 (2015)
4. Bhandari, U., Sugiyama, K., Datta, A., Jindal, R.: Serendipitous recommendation for mobile apps using item-item similarity graph. In: Proc. of AIRS, pp. 440–451 (2013)
5. Chen, N., Hoi, S., Li, S., Xiao, X.: Simapp: A framework for detecting similar mobile applications by online kernel learning. In: Proc. of ACM WSDM, pp. 305–314 (2015)
6. Chen, N., Hoi, S., Li, S., Xiao, X.: Mobile app tagging. In: Proc. of ACM WSDM, pp. 63–72 (2016)
7. Crussell, J., Gibler, C., Chen, H.: Attack of the clones: Detecting cloned applications on android markets. In: Proc. ESORICS, pp. 37–54 (2012)
8. Crussell, J., Gibler, C., Chen, H.: Andarwin: Scalable detection of android application clones based on semantics. IEEE TMC 14(10), 2007–2019 (Oct 2016)
9. Do, Q., Martini, B., Choo, K.K.: Exfiltrating data from android devices. Computers and Security 48(C), 74–91 (Feb 2015)
10. Dutta, B., Shinde, J.: Intuitionistic fuzzy clustering based segmentation of spine mr image. International Research Journal of Engineering and Technology 4(7), 790–794 (July 2017)
11. Faruki, P., Laxmi, V., Bharmal, A., Gaur, M., Ganmoor, V.: Androsimilar: Robust signature for detecting cariants of android malware. Information Security and Applications 22, 66–80 (2015)
12. Feizollah, A., Anuar, N.B., Salleh, R., Abdul Wahab, A.: A review on feature selection in mobile malware detection. Digital Investigation 13(C), 22–37 (Jun 2015)
13. Hamedani, M.R., Gyoosik, K., Seong-je, C.: Simandro: An effective method to compute similarity of android applications. Soft Computing pp. 1–22 (Jan 2019)
14. Hamedani, M.R., Kim, S.w.: Jacsim: An accurate and efficient link-based similarity measure in graphs. Information Sciences 414, 203–224 (November 2017)
15. Hamedani, M.R., Kim, S.W., Kim, D.J.: Simcc: A novel method to consider both content andcitations for computing similarity of scientific papers. Information Sciences 334-335(C), 273–292 (Mar 2016)
16. Lin, Y.S., Jiang, J.Y., Lee, S.J.: A similarity measure for text classification and clustering. IEEE TKDE 26(7), 1575–1589 (Jul 2014)
17. Ma, Z., Ge, H., Liu, Y., Zhao, M., Ma, J.: A combination method for android malware detection based on control flow graphs and machine learning algorithms. IEEE ACCESS 7, 21235–21245 (Feb 2019)
18. Magdy, W., Gareth, J.: Pres: A score metric for evaluating recall-oriented information retrieval applications. In: Proc. of ACM SIGIR, pp. 611–618 (2010)
19. Manning, C., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
20. Motta, J.M., Ladouceur, J.: A crf machine learning model reinforced by ontological knowledge for document summarization. In: Proceedings of the International Conference Artificial Intelligence, pp. 127–135 (2017)
21. Wei, J., He, J., Kai, C., Zhou, Y., Tang, Z.: Collaborative filtering and deep learning based recommendation system for cold start items. Expert Systems with Applications 69(1), 29–39 (March 2017)
22. Yin, P., Luo, P., Lee, W.C., Wang, M.: App recommendation: A contest between satisfaction and temptation. In: Proc. of ACM WSDM, pp. 395–404 (2013)

23. Zhang, M., Duan, Y., Yin, H., Zhao, Z.: Semantics-aware android malware classification using weighted contextual api dependency graphs. In: Proc. of ACM CCS, pp. 1105–1116 (2014)
24. Zhang, Y., Ren, W., Zhu, T., Ren, Y.: Saas: A situational awareness and analysis system for massive android malware detection. Future Generation Computer Systems 95, 548–559 (Jan 2019)

Masoud Reyhani Hamedani received the B.S. degree in computer science from Shahid Bahonar University, Kerman, Iran, in 2004, and the M.S. degree in software engineering from Payame Nour University, Tehran, Iran, in 2009, and the PhD degree in computer science from Hanyang University, Seoul, Korea in 2016. He worked as a postdoc researcher in Dankook University, Yongin, Korea until February 2018. In March 2018, he joint Hanyang University and currently is working as aresearch assistant professorin the Industry-University Cooperation Foundation, Program for Advanced AI Research and Education. His current research interests include data science, feature representation learning, similarity computation in social network, and deep learning.

Sang-Wook Kim received the B.S. degree in computer engineering from Seoul National University, in 1989, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST), in 1991 and 1994, respectively. In 2003, he joined Hanyang University, Seoul, Korea, where he currently is a professor at the Department of Computer Science and Engineering and the director of the Brain-Korea21-Plus research program. He is also leading a National Research Lab (NRL) Project funded by the National Research Foundation since 2015. From 2009 to 2010, he visited the Computer Science Department, Carnegie Mellon University, as a visiting professor. From 1999 to 2000, he worked with the IBM T. J. Watson Research Center, USA, as a postdoc. He also visited the Computer Science Department at Stanford University as a visiting researcher in 1991. He is an author of more than 200 papers in refereed international journals and international conference proceedings. His research interests include databases, data mining, multimedia information retrieval, social network analysis, recommendation, and web data analysis. He is a member of the ACM and the IEEE.

Received: February 08, 2021; Accepted: May 15, 2021.

Computer Science and Information Systems 18(4):1239–1252 https://doi.org/10.2298/CSIS210119023Z

Analysis of Entrepreneur Mental Model and Construction of its Portrait

Yongzhong Zhang[1], Yonghui Dai[2,⋆], and Haijian Chen[1]

[1] Institute of science and technology, Shanghai Open University,
Shanghai 200433, China
1502235429@qq.com, xochj@sou.edu.cn
[2] Management School,
Shanghai University of International Business and Economics,
Shanghai 201620, China
daiyonghui@suibe.edu.cn

Abstract. Previous studies have shown that the mental model of entrepreneurs has a significant impact on the growth of entrepreneurial enterprises. This paper explores a new method to analyze entrepreneur mental model and construct its portrait. Firstly, according to existing research results, this paper summarizes three key factors that affect entrepreneurial mental model: prior knowledge, personality characteristics and opportunity perception. Since then, the methods of entrepreneur mental portrait are introduced, which including cluster analysis method and fuzzy comprehensive evaluation method. Based on the investigation and analysis of 277 entrepreneurs, our study shows that the above construction method of mental model can accurately describe the entrepreneur mental model. The contribution of this paper is to explore the mental division of different types of entrepreneurs, and give the method of mental portrait of entrepreneurs, which provides a meaningful reference for promoting innovation and entrepreneurship education and training.

Keywords: Entrepreneur mental model, Mental portrait, Innovation and entrepreneurship, Data mining.

1. Introduction

In recent years, great changes in the global economic situation and the development of information technology have brought opportunities for the development of entrepreneurship. As a new engine of economic growth and enterprise transformation, innovation and entrepreneurship play an active role in promoting technological reform, enhancing product competitiveness and expanding employment. However, after the establishment of the enterprise, it needs to face many problems in the development process of the enterprise, such as the entrepreneurial environment, the entrepreneurial team, the integration of entrepreneurial resources, market competition etc., which makes the failure rate of entrepreneurship very high. Statistical data shows that the average life expectancy of Chinese private enterprises is 3.7 years, and the average life expectancy of SMEs is even shorter, it is only 2.5 years [10]. Previous studies have found that the mental model formed

⋆ Corresponding author

by entrepreneur personality characteristics, inherent temperament and experience has significant impact on the grasp of entrepreneurial opportunities, risk recognition, team building, social capital utilization and decision-making behaviors, which are the key factors for successful entrepreneurial practices. Therefore, the mental model of entrepreneurs has always been the focus of entrepreneurial education and training in recent years.

This paper is organized as follows. In section 2, literature review of mental model is introduced. In section 3, the analysis of entrepreneur mental model is shown, which including prior knowledge, personality characteristics and opportunity perception. In section 4, the construction of entrepreneur mental portrait is illustrated. In section 5, experiment and analysis are shown. In the end of section, the conclusion of this paper is described.

2. Literature Review

Mental model was first proposed in the field of psychology. Psychologists believe that the mental model is a combination of a person's inner psychology and the person's own intelligence. It explains the internal cognitive process constructed in people's brain that affects people's understanding, interpretation and facing the world [2][7]. Since then, the mental model has been improved by scholars from different disciplines and developed into a mental model theory [3]. In the definition of mental model, scholars understand from different professions and give their own definition of mental model. For example, some scholars think that mental model is a model that reflects the objective world constructed by individuals after observing the real world. It refers to the individual's interpretation of the objective world or reasoning and decision-making based on the objective world [15]. Mental model is closely related to the human brain. It is a temporary representation of the problem situation in the short-term memory of the brain when people recognize things in life, or the stable representation of the external world stored in long-term memory [4]. The formation of mental model can be seen as a kind of mapping of external things in the brain. It may not only come from the continuous accumulation of experience and knowledge of daily life, but also be formed by the instantaneous stimulation of external things to the brain [6][22]. Mental model is a 'small model' constructed by human brain itself corresponding to the real world after observing various things in the real world are stimulated. The model can be used for prediction, logical reasoning, or as the basis for explaining phenomena [12]. In the mental model, the mind represents the inner psychology of human beings, while the model refers to the external manifestation of human beings, which is related to individual characteristics. Therefore, mental model abstracts the internal characteristics of people's inner thoughts into an external form of expression [20]. Generally speaking, mental model can be regarded as the sum of the human brain's perception of what is seen and heard in real life, and the resultant psychological activities and thinking response ability. The generation process of mental model is shown in Fig. 1.

Entrepreneur mental is the psychological activity and thinking mode of entrepreneurs in the process of starting a business. It is also the sum of their thinking ability to make initial judgment and analysis on external things or events [21]. On the view of entrepreneurship mental model, scholars' research perspectives mainly include the content, influencing factors, cognition and function of entrepreneurship mind. For example, some scholars believe that entrepreneurs are the core factor that determines the success of entrepreneurship, and the minds of entrepreneurs play a key role in it [14]. Some scholars took entrepreneurs

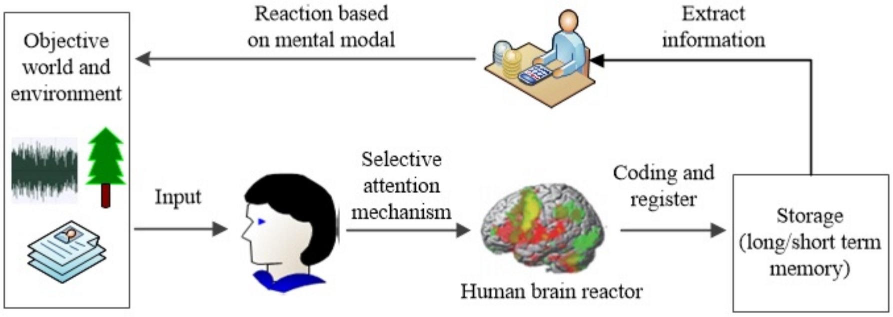


Fig. 1. The generation process of mental model

as the research object and analyzed their management behaviors and put forward the view that unique thinking, independent thinking, and risk-taking are typical qualities of entrepreneurs [5][1]. Some scholars have studied the minds of entrepreneurs from the two dimensions of prior knowledge and belief systems, and pointed out that the knowledge of entrepreneurs' innovation opportunities reflects the Entrepreneur prior knowledge, vigilance, beliefs, cognitive models and other cognitive elements Complex mental process [18][13]. The above research results provide a good reference for the study of entrepreneur portraits in our study.

3. Analysis of Entrepreneur Mental Model

In the research of entrepreneur mind, although scholars have studied the Entrepreneur mental model from different perspectives, the conclusions obtained are also different to some extent, but on the whole, the research on Entrepreneur mental model can be summarized into three dimensions: prior knowledge, personality characteristics and opportunity perception.

3.1. Prior Knowledge

Previous studies have shown that entrepreneurs with previous entrepreneurial experience will have an advantage over those without entrepreneurial experience. Entrepreneurs with entrepreneurial experience can quickly form judgments on the current situation and make optimal response plans in the face of sudden changes in the environment. Their accumulated experience and prior knowledge can prevent risks and seize opportunities through vigilance.

Entrepreneurial experience mainly refers to the practical experience and industry experience of entrepreneurs [19]. Under the guidance of previous experience, entrepreneurs can experience more entrepreneurial insights or capture more valuable information in the process of entrepreneurship, so that it is easier to identify the entrepreneurial opportunities in favor of their own enterprises. However, those entrepreneurs who haven¡¯t started a business before, such as college students who start a business for the first time. They haven't entrepreneurial experience and lack the knowledge reserves in this area, so their understanding of entrepreneurial opportunities and business information is weaker than those with entrepreneurial experience, which affects their entrepreneurial performance. Based on previous research results, we selected accumulated information, technical capabilities and entrepreneurial ideas as the measurement factors.

3.2. Personality Characteristics

Individual personality trait is a description of individual personality characteristics. It is a relatively stable personality characteristic, which has both innate part and acquired part. For entrepreneurs, entrepreneur characteristics are the description of the Entrepreneur internal psychological characteristics, which refers to the Entrepreneur own personality, cognitive bias, motivation and so on. It is the personality characteristics based on personal physiology, which will change under the influence of the surrounding environment and situation. It is the synthesis of the potential innovative thinking and behavior mode of the entrepreneur, such as the entrepreneurial intention of the entrepreneur Different characteristics of willingness and passion, determination and creativity compared with non-entrepreneurs. Some scholars have studied the relationship between personality characteristics and entrepreneurial performance. They found that the stability, extroversion and suitability of personality have a positive impact on team shared mind and team performance [16].

3.3. Opportunity Perception

Entrepreneurial opportunity perception refers to the business opportunities that are beneficial to start-up enterprises. Entrepreneurs can turn the opportunities into valuable products or services and provide them to customers so that they can get benefits. The identification of entrepreneurial opportunities is very important to start-up enterprises, which is an important factor for the smooth development of start-up enterprises. There are also different views on entrepreneurial opportunities due to different research perspectives. Some scholars believe that when supply and demand are unbalanced in the market, there will be entrepreneurial opportunities; some studies believe that the identification of entrepreneurial opportunities is a process of subjective psychological perception; some scholars regard entrepreneurial opportunities as the starting point of entrepreneurship. Entrepreneurial opportunity identification is the thinking process of acquiring and identifying things, and the process of entrepreneurs' perception and discovery of opportunities, and then starting new businesses and new enterprises.

In the dynamic process of entrepreneurship, if entrepreneurial enterprises cannot identify opportunities sensitively, they will easily fall into development difficulties. Some scholars believe that entrepreneurial opportunities are closely related to the ability of entrepreneurs to identify opportunities. Entrepreneurial opportunities change with the environment, market, customers and other factors. For start-up enterprises, the identification of entrepreneurial opportunities needs entrepreneurs to judge whether the opportunities are operable after examining their own qualifications, capabilities and partners. In terms of the source of entrepreneurial opportunities, some scholars believe that entrepreneurial opportunities exist in the objective world, only need entrepreneurs to identify them; entrepreneurial opportunities are created by entrepreneurs using their own talents;

entrepreneurs can identify suitable entrepreneurial opportunities by observing the environment, conditions, etc., using wisdom, technology and other means. Generally speaking, entrepreneurial opportunity identification is affected by the factors of opportunity discovery and opportunity identification.

4. Construction of Entrepreneur Mental Portrait

4.1. Framework of Entrepreneur Mental Portrait

Entrepreneur mental portrait, namely the labeling of Entrepreneur mind. It refers to a labeled Entrepreneur mental model abstracted from the Entrepreneur basic information and entrepreneurial activity characteristics. It is a means of using labels to depict the entrepreneur appearance. Our study combines the mental model measurement methods based on psychological response and physiological index response, and uses cluster analysis, association rule analysis and fuzzy synthetic evaluation to construct entrepreneur mental portrait. The framework of entrepreneur mental portrait is shown in Fig. 2.

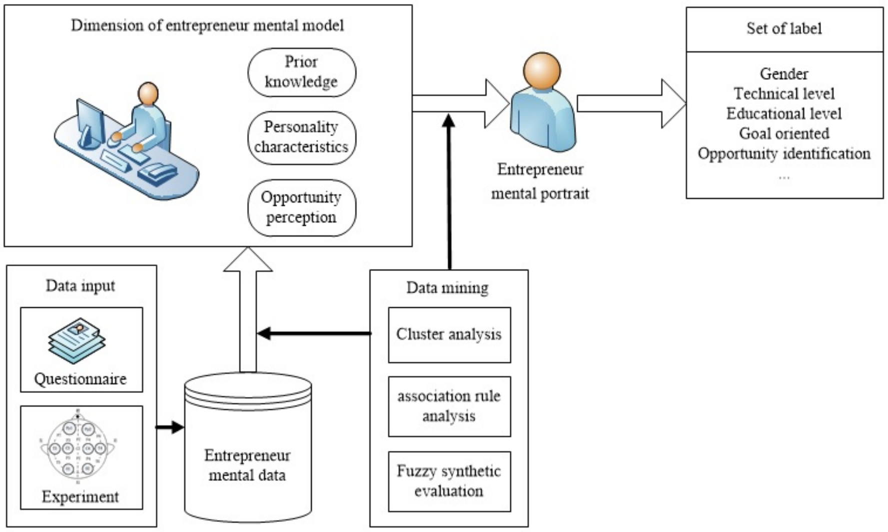


Fig. 2. Framework of entrepreneur mental portrait

It can be seen from Fig. 2 that the mental portrait of entrepreneurs mainly includes the following processes. Firstly, the Entrepreneur mental data are collected through interview survey, questionnaire survey, interview and experimental observation, and then data mining methods such as classification, cluster analysis and association rule analysis are used to draw the three dimensions of entrepreneur mental, namely, prior knowledge, personality characteristics and opportunity perception. After the above process, the set of entrepreneur mental label is given, then the entrepreneur mental portrait is complete. In order to get the ideal portrait, data mining is needed to process the data. The following is the introduction of clustering analysis algorithm and fuzzy synthetic evaluation method.

4.2. Clustering Analysis

Clustering analysis refers to the process of similar division of data sets by some rules and methods [17]. It originated from numerical taxonomy. In the past, people mainly rely on experience or professional knowledge to classify things. With the advent of information technology and big data era, only relying on experience and professional knowledge cannot meet the complex classification requirements. Therefore, the numerical classification based on mathematical tools is applied to the classification of things, and clustering analysis is produced. After years of development of clustering analysis, the current clustering analysis methods have formed many algorithms, such as k-means algorithm, Clara algorithm, PCM fuzzy clustering algorithm, SOM self-organizing neural network clustering algorithm, etc [8][9]. Among, the process of improved k-means algorithm is shown in Fig. 3.

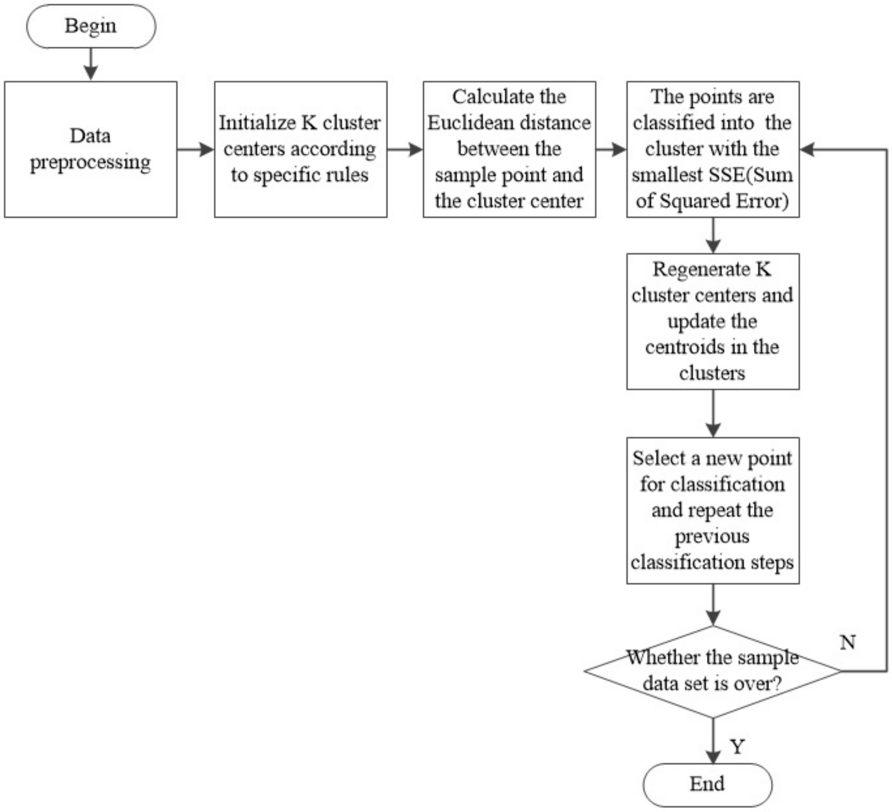

Begin
Data preprocessing
Initialize K cluster centers according to specific rules
Calculate the Euclidean distance between the sample point and the cluster center
The points are classified into the cluster with the smallest SSE(Sum of Squared Error)
Regenerate K cluster centers and update the centroids in the clusters
Select a new point for classification and repeat the previous classification steps
Whether the sample data set is over?
N
Y
End

Fig. 3. The improved k-means algorithm

The main steps included in the improved k-means algorithm are as follows:

Step 1: Initialize K cluster centers according to the principle that the distance between the initial cluster centers should be as far as possible.

Step 2: Assign the sample set data D to the nearest cluster according to the principle of shortest distance;

Step 3: Calculate the mean value of the sample center of each cluster, regenerate K cluster centers, and update the centroid within the cluster;

Step 4: Whether the cluster center is no longer changes or the maximum number of iterations n has been reached, if it is yes, then go to Step 5, otherwise repeat to Step 2;

Step 5: Output the final cluster center and k cluster divisions.

4.3. Fuzzy Synthetic Evaluation Method

Fuzzy synthetic evaluation method is a comprehensive evaluation method based on fuzzy mathematics membership degree theory [11]. This method uses fuzzy relations to quantify some qualitative problems, that is, those factors whose boundary is fuzzy and difficult to quantify are quantified by formula. The advantage of this method is that the results are displayed quantitatively, which is clear and easy to understand, and can solve those problems well Fuzzy and difficult to quantify qualitative problems. The establishment process of fuzzy comprehensive evaluation model mainly includes the establishment of fuzzy comprehensive evaluation matrix, single factor analysis, factor comprehensive evaluation and comprehensive evaluation value calculation.

(1) Establishment of fuzzy synthetic evaluation matrix

Suppose the set I=I$_1$,I$_2$,...,I$_n$ is the set of factor index, F=F$_1$,F$_2$,...,F$_m$ is the set of factor comments, where F$_j$ (j=1,2,...,m) is the evaluation grade of each factor, and the fuzzy evaluation of each factor is a fuzzy subset of the factor evaluation set S. Suppose that the single factor fuzzy evaluation of factor I is R$_i$=r$_{i1}$,r$_{i2}$,...,r$_{im}$ (i=1,2,...,n), and r$_{ij}$ is the membership degree of the i-th factor to the j-th comment. The fuzzy vector of R$_1$,R$_2$,...,R$_n$ forms the fuzzy relation from set I to set S, and the fuzzy comprehensive evaluation matrix is as follows.

$$R = \begin{bmatrix} R_1 \\ R_2 \\ ... \\ R_n \end{bmatrix} = \begin{bmatrix} r_{11}\ r_{12}\ ...\ r_{1m} \\ r_{21}\ r_{22}\ ...\ r_{2m} \\ ...\ ...\ ...\ ... \\ r_{n1}\ r_{n2}\ ...\ r_{nm} \end{bmatrix} \tag{1}$$

(2) Single factor analysis

Let the fuzzy vector V$_i$=(V$_{il}$,V$_{i2}$,...,V$_{in}$) The membership degree (k=1,2,...,m) on the set I represents the score of each factor in the single factor evaluation, and the single factor evaluation vector B$_i$ is as follows.

$$B_i = V_i * R_i = (b_{i1}, b_{i2}, ..., b_{im}), (i = 1, 2, ..., k) \tag{2}$$

(3) Construction of comprehensive evaluation vector Suppose that the weight vector of each subset is X=(X$_1$,X$_2$,...,X$_k$) The comprehensive evaluation matrix is R=(B$_1$,B$_2$,...,B$_k$) Therefore, the comprehensive evaluation vector is as follows.

$$B = X * R = (b_1, b_2, ..., b_m) \tag{3}$$

(4) Calculation of comprehensive evaluation value Each evaluation grade of the evaluation set is given a score, and the evaluation set is C=(c$_1$,c$_2$,...,c$_m$) The comprehensive evaluation score is as follows.

$$S = B * C^T \tag{4}$$

After obtaining the evaluation score of the factors, it can be found that the corresponding grade according to the score to know the level of the evaluated things.

5. Experiments and Results

According to the previous design of Entrepreneur mental portrait, this paper constructs the Entrepreneur mental portrait. First of all, some CEOs and executives of some start-up companies are interviewed about entrepreneurial mind research, and questionnaires are sent out to survey them, and then the wearable device experiment is conduct and construction of comprehensive evaluation vector is built. Then, the survey data are summarized and analyzed, and the entrepreneur mental tag is extracted and the entrepreneur mental portrait is completed.

5.1. Design of Questionnaire

In the process of entrepreneurship, entrepreneurs judge and make decisions on the competitive market based on their own experience, knowledge and ability, and identify and grasp business opportunities. In the above process, the prior knowledge, personality characteristics and opportunity perception ability of individuals play an important role, which are related to the minds of entrepreneurs. Therefore, based on the above content and the previous scholars' research on the Entrepreneur mind, the survey content of this study is summarized. The example of the questionnaire is shown in Table1.

The scale in the questionnaire is divided into three categories: prior knowledge, personality characteristics and opportunity perception. Among, the first category of ¡®prior knowledge¡¯ is divided into three categories: entrepreneurial experience, related knowledge and technical proficiency after referring to the previous studies. The results of survey data are shown as in Table2.

The second category is 'personality characteristics', which are divided into five categories, and the results of survey data are shown as in Table3.

The third category is 'opportunity perception', which is divided into two categories, and the results of survey data are shown as in Table4.

5.2. Reliability Test

In the reliability measurement of the questionnaire, the commonly used index is Cronbach alpha coefficient, which is between 0 and 1. The closer the coefficient is to 0, the lower the reliability of the questionnaire, on the contrary, it is closer to 1, the higher the reliability of questionnaire. Generally speaking, if the Cronbach alpha coefficient is greater than 0.8, it indicates that the reliability of the scale is in the ideal range. If the coefficient is less than 0.6, it indicates that the reliability of the scale does not meet the requirements, and it must be rebuilt. The reliability analysis of individual prior knowledge scale, personality characteristics and opportunity perception scale were performed in SPSS 22.0 software. The reliability test results are shown in Table5.

Table 1. Sample of questionnaire of entrepreneur mental

Dimension	Content	Example
Prior knowledge	It is mainly measured from the entrepreneur practical experience, entrepreneurial knowledge accumulation and entrepreneurial thinking.	How much do you know about entrepreneurship related knowledge and entrepreneurial environment? A. Very familiar with B. Relatively familiar C. General D. I don't understand E. I don't know
Personality characteristics	It is mainly measured from five aspects, namely, entrepreneur's values, extraversion, self-cognition, entrepreneurial achievement and stability	How much attention do you pay to new things or technologies? A. Very attentive B. more attentive C. General D. less attention E. Very inattentive
Opportunity perception	It is mainly measured by the ability of entrepreneur alertness and entrepreneur opportunity identification	Do you agree that you have a strong ability to discover opportunities, the products you discover are leading, and it is difficult to have substitutes in the short term? A. Very much agree B. Relatively agree C. General D. Disagree E. Very disagree.

Table 2. Results of prior knowledge

Questionnaire option		Number of answers to questions					Average
		1	2	3	4	5	
Related Knowledge	Question 1	2	7	37	133	98	4.15
	Question 2	1	5	46	112	113	
	Question 3	5	6	35	136	95	
Technical proficiency	Question 1	5	16	92	103	61	3.78
	Question 2	8	15	86	101	67	
	Question 3	7	10	67	115	78	
Related Knowledge	Question 1	14	37	91	86	49	3.38
	Question 2	21	46	72	91	47	
	Question 3	24	39	79	84	51	

Table 3. Results of personality characteristics

Questionnaire option		Number of answers to questions					Average
		1	2	3	4	5	
Extrovert	Question 1	3	5	43	99	127	4.13
	Question 2	5	6	65	94	107	
	Question 3	6	5	62	83	121	
Decisiveness	Question 1	4	15	56	97	105	3.98
	Question 2	3	19	57	102	96	
	Question 3	8	17	63	86	103	
Adventurousness	Question 1	9	18	79	88	83	3.84
	Question 2	8	19	68	94	88	
	Question 3	11	22	52	99	93	
Interest	Question 1	13	24	61	84	95	3.73
	Question 2	14	25	81	75	82	
	Question 3	12	21	77	81	86	
Logicality	Question 1	13	24	69	82	89	3.68
	Question 2	15	26	72	77	87	
	Question 3	16	29	88	71	73	

Table 4. Results of opportunity perception

Questionnaire option		Number of answers to questions					Average
		1	2	3	4	5	
Alertness	Question 1	2	8	84	98	85	3.88
	Question 2	3	10	93	92	79	
	Question 3	3	9	89	94	82	
Feasibility	Question 1	15	30	66	89	77	3.64
	Question 2	19	28	67	91	72	
	Question 3	21	24	65	86	81	

Table 5. Overall reliability and reliability test of each dimension

Test items of scale	Cronbach Alpha	Number of questions
Prior knowledge	0.854	9
Personality characteristics	0.954	15
Opportunity perception	0.915	6
All Questionnaire	0.968	30

It can be seen from Table5 that the alpha coefficients of the three dimensions of the questionnaire are all above 0.8, and the overall reliability coefficient of all questionnaire is 0.968, which indicating that the reliability of the questionnaire is high.

5.3. Validity Test

After the reliability of the scale has passed the feasibility test, the corresponding validity analysis is needed to verify the validity, and the verification results are shown in Table6.

Table 6. KMO and Bartlett's test of spherical

Kaiser-Meyer-Olkin	measure	0.923
Bartlett's test of spherical	chi square	12635.289
	df	435
	Sig	0.000

When the KMO value is higher than 0.8, the validity is very good. If the value is between 0.7 and 0.8, the validity is good; if the value is between 0.6 and 0.7, the validity of the questionnaire is acceptable; if the KMO value is lower than 0.6, the validity of the questionnaire is poor. The KMO value of the questionnaire for entrepreneur mental is 0.923, which indicating that the validity of the questionnaire is very good.

5.4. Descriptive Statistical Analysis

According to the descriptive statistical results of 277 valid questionnaires collected, there are 171 male entrepreneurs, accounting for 61.73%, and 106 female entrepreneurs, accounting for 38.27%. In our survey, male entrepreneurs are more than female entrepreneurs, which is consistent with the fact that there are more male entrepreneurs in China. From the perspective of age distribution of entrepreneurs, 22.02% of entrepreneurs are less than 25 years old, 44.77% are 26-35 years old, 18.05% are 36-45 years old, and 15.16% are over 45 years old. Generally speaking, the age of entrepreneurs is relatively young. The descriptive statistical information of entrepreneurs' basic information is shown in Table7.

According to the collected questionnaire data, the labels of three dimensions of entrepreneur mental can be established. At the same time, combined with the basic information of entrepreneur, the entrepreneur mental can be depicted as shown in Fig. 4.

6. Conclusions

The mental model of entrepreneurs is a research hot issue of entrepreneurial psychology, and it has been the concerned point in recent entrepreneurial education and training, which should be conducted according to the different entrepreneurs' mental models and profiles. However, traditional methods for entrepreneur mental model assessment usually have the limitations of inaccuracy and even leading to inconsistent results. The contribution of this paper is to analyze innate and acquired influencing factors of mental models, and give the

Table 7. Descriptive statistics of the basic information of the questionnaire

Item	Options	Percentage(%)
Gender	Male	61.73
	Female	38.27
Age	Less than 25 years old	22.02
	26-35 years	44.77
	36-45 years	18.05
	Over 45 years old	15.16
Position	CEO	33.94
	Business executives	66.06
Education level	High school and below	2.17
	Junior college	10.83
	Undergraduate	54.87
	Master’s degree and above	32.13
Industry	Biomedical industry	6.13
	New materials or new energy	12.64
	Training, education	15.16
	Information transmission and computer service	21.67
	Clothing industry	13.36
	Life service industries such as hotels and tourism	8.67
	Wholesale and retail	6.86
	Culture, sports, entertainment industry	6.13
	Agriculture, forestry, animal husbandry and fishery	5.05
	Other industry	4.33
Survival years	Less than 1 year	15.52
	1-3 year	48.74
	3-5 year	25.63
	More than 5 years	10.11
Number of enterprises scale	Less than 5 people	7.58
	6-10 people	17.33
	11-25 people	54.51
	26-50 people	11.55
	More than 50 people	9.03

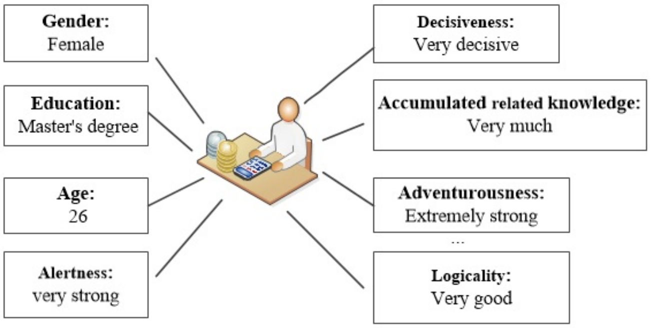


Fig. 4. Entrepreneur mental portraits

methods of mental model measurement and mental portrait construction of entrepreneurs, and it complete mental description and label portrait by combining research methods such as questionnaire, brain cognitive experiment, cluster analysis and fuzzy comprehensive evaluation, which provides a new idea for the study of entrepreneurial mental model.

The proposed method of this paper indicates as the effective way for testing and analyzing the entrepreneurs' mental models. However, as an exploratory research, the questionnaire design and sample size of the survey objects of this paper can be further improved. In addition, future research can introduce more diverse research methods, such as combining with wearable devices to conduct mental model experiments and analysis on entrepreneurs to make the results of the research more accurate.

Acknowledgments. This work is supported by the project of Shanghai Philosophy and Social Sciences Plan (No. 2018BGL023, No. 2016BGL004), Shanghai Higher Education Society (No.Y1-24) and National Natural Science Foundation of China (No. 71971066).

References

1. Berglund, H.: Between cognition and discourse: phenomenology and the study of entrepreneurship. International journal of entrepreneurial behavior and research 21(3), 472–488 (2015)
2. Craik, K.W. (ed.): The Nature of Explanation. Cambridge University Press (1943)
3. Defranco, J.F., Neill, C.J., Clariana, R.B.: A cognitive collaborative model to improve performance in engineering teams¡ªa study of team outcomes and mental model sharing. Systems Engineering 14(3), 267–278 (2011)
4. Dong, A., Kleinsmann, M.S., Deken, F.: Investigating design cognition in the construction and enactment of team mental model. Design Studies 34(1), 1–33 (2013)
5. Dutta, D.K., Gwebu, K.L., Wang, J.: Personal innovativeness in technology, related knowledge and experience, and entrepreneurial intentions in emerging technology industries: a process of causation or effectuation? International entrepreneurship and management journal 11(3), 529–555 (2015)
6. Filipowicz, A., Anderson, B., Danckert, J.: Adapting to change: The role of the right hemisphere in mental model building and updating. Canadian Journal of experimental psychology 70(3), 201–218 (2016)
7. Gadgil, S., Nokes-Malach, T.J., Chi, M.T.H.: Effectiveness of holistic mental model confrontation in driving conceptual change. Learning and Instruction 22(1), 47–61 (2009)
8. Ghaseminezhad, M.H., Karami, A.: A novel self-organizing map (som) neural network for discrete groups of data clustering. Applied Soft Computing 11(4), 3771–3778 (2011)
9. Grover, N.: A study of various fuzzy clustering algorithms. International Journal of Engineering Research 3(3), 177–181 (2014)
10. Jiang, J.Y.: Identification and avoidance of entrepreneurial risk of new enterprises. Chinese Market (45), 53–55 (2013)
11. Li, H., Liu, G., Yang, Z.: Improved gray water footprint calculation method based on a mass-balance model and on fuzzy synthetic evaluation. Journal of Cleaner Production 219, 377–390 (2019)
12. Li, H.T., Song, L.L.: Selection and application of mental model measurement methods for users using websites. Information Studies: Theory and Application 38(2), 11–16 (2015)
13. Obschonka, M., Hakkarainen, K., Lonka, K., Katariina, S.A.: Entrepreneurship as a twenty-first century skill: entrepreneurial alertness and intention in the transition to adulthood. Small Business Economics 48, 1–15 (2016)

14. Obschonka, M., Stuetzer, M.: Integrating psychological approaches to entrepreneurship: the entrepreneurial personality system. Small Business Economics 49(1), 203–231 (2017)
15. Pybus, L., Welk, A.K., Gillan, D.J.: Differences in mental model development among psychology and engineering students of a human factors course. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 60, pp. 361–365 (2016)
16. Rhee, J., Parent, D., Basu, A.: The influence of personality and ability on undergraduate teamwork and team performance. Springerplus 2(1), 1–14 (2013)
17. Sohrabi, B., Vanani, I.R., Abedin, E.: Human resources management and information systems trend analysis using text clustering. International Journal of Human Capital and Information Technology Professionals 9(3), 1–24 (2018)
18. Tang, J.T.: Environmental munificence for entrepreneurs: entrepreneurial alertness and commitment. International Journal of Entrepreneurial Behaviour & Research 14(3), 128–151 (2013)
19. Toft-Kehler, R., Wennberg, K., Kim, P.H.: Practice makes perfect: Entrepreneurial-experience curves and venture performance. Journal of Business Venturing 29(4), 453–470 (2014)
20. Wolfgang, S., Christian, K.: External and internal representations in the acquisition and use of knowledge: visualization effects on mental model construction. Instructional Science 36(3), 175–190 (2008)
21. Yu, F.L.T.: Entrepreneur interpretation, innovation and coordination in austrian subjectivist perspective. Global Business and Economics Review 9(2), 255–270 (2007)
22. Zhang, Y.Z., Dai, Y.H., Lu, S.Q., Li, S., Lin, Y.: Design of intelligent learning resources for mooc based on mental model. In: 3rd IEEE Information Technology, Networking, Electronic and Automation Control Conference. pp. 1028–1032 (2019)

Yongzhong Zhang is a professor at the Shanghai Engineering Research Center of Open Distance Education, Shanghai Open University, China. His current research interests include Educational technology and big data analysis. Contact him at 1502235429@qq.com.

Yonghui Dai is the corresponding author of this paper. He is currently a lecturer at the Management School, Shanghai University of International Business and Economics, China. He received his Ph.D. in Management Science and Engineering from Shanghai University of Finance and Economics, China in 2016. His current research interests include Entrepreneurship education and artificial intelligence. His works have appeared in international journals more than thirty papers. Contact him at daiyonghui@suibe.edu.cn.

Haijian Chen is a professor at the Shanghai Academic Credit Transfer and Accumulation Bank for Lifelong Education, China. He received his Ph.D. in Management Science and Engineering from Shanghai University of Finance and Economics, China in 2015. His current research interests include Educational technology and cloud computing. Contact him at xochj@sou.edu.cn.

Received: January 19, 2021; Accepted: March 28, 2021.

Computer Science and Information Systems 18(4):1253–1269 https://doi.org/10.2298/CSIS210119024L

Research on Influencing Factors of the Development of Cultural and Creative Industries Based on Grey Factor Analysis

Jianjun Li [1] and Jia Liao [2, *]

[1] Management School,
Shanghai University of International Business and Economics,
Shanghai 201620, China
lijianjun@suibe.edu.cn
[2] School of Business,
Shanghai University of International Business and Economics,
Shanghai 201620, China
liaojia@suibe.edu.cn

Abstract. In order to study the influencing factors of cultural and creative industries (CCIs), the Grey Factor Analysis and 30 different indexes are used to empirically analyze the correlation between the influencing factors and the added value of CCIs in Shanghai. At the same time, main environmental factors affecting the development of CCIs are explored. The result shows that technology research and development, policy and government financial support, human resources, social culture, cultural consumption environment, cultural industry basis and development status are important impacting factors on the development of CCIs in Shanghai. Based on the above research results, this paper puts forward some countermeasures and suggestions on the construction of a comprehensive environment to promote the sustainable and healthy development of CCIs.

Keywords: Influencing Factors; Grey Factor Analysis; Cultural and Creative Industries; Environmental factors.

1. Introduction

With the interactive and integrated development of the cultural industry and creative industry, a new industry format named CCI is formed by the elements comprising culture, creativity, technology, capital and manufacturing. The development of CCI has shown strong industrial functions in breaking through resource and environmental constraints, building industrial innovation capabilities, promoting industrial restructuring and upgrading and boosting overall economic growth (Fausto, 2018). In 2017, China’s added value of culture and related industries was 3472.2 billion-yuan, an increase of 12.8% over the previous year, accounting for 4.2% of the GDP, an increase of 0.06 percentage points over the previous year. Shanghai started to develop its CCIs comparatively earlier than other cities in China. As of February 2019, 137 municipal cultural creative industrial parks have been established. In 2017, the added value of

[*] Corresponding author

Shanghai’s CCIs reached 334.014 billion-yuan, accounting for 12.1% of the city’s GDP, contributing more than 23% to Shanghai’s economic growth. It can be seen that CCI is getting more and more importance in many countries and regions in the world to alleviate resource and environmental pressures, enhance regional comprehensive competitive advantages, promote industrial restructuring and upgrading, and form new economic growth points. However, difficulties and uncertainties still exist in the development of China’s CCIs. Therefore, it is necessary to conduct an in-depth analysis of the environmental factors, cracking on the obstacles to the development of CCIs, and construct a benign and comprehensive environment to promote the sustainable and healthy development of CCIs, so as to explore the development path and-model of this industry.

Previous studies mainly focused on particular influencing factors affecting CCIs from an empirical perspective, or focusing on individual factors affecting the development of creative industry. However, studies comprehensively analyzing the environmental factors affecting CCIs in the context of social and economic development are still rare. This paper aims to fill this gap.

The rest of this paper is arranged as follows. Section 2 reviews the related literature. Section 3 introduces indicators and research methodology. Section 4 describes data and empirical results. Section 5 gives conclusion and suggestions.

2. Literature Review

There is no consensus towards the definition of CCIs. The earliest attempt to define the term “cultural and creative industries” is made by the Department of Culture, Media and Sport of UK in 1998, which identified 13 sectors as constituting creative industries in the UK [3]. After that, two influential definitions were given by European Commission [6], who defines CCIs as industries using culture as an input, and UNCTAD [27], who stresses more on the creative aspect and describes CCIs as a set of creative economic activities. Inspired by the CCI concept, governments began to attach great importance to create better cultural and creative environment, so as to benefit urban development and economic growth. Overall, CCIs are increasingly important in the economic development of various countries [14].

Scholars have conducted in-depth analysis of the factors affecting the development of CCIs from different perspectives [1, 18]. It is believed that five essential conditions are required for the development of CCIs, known as the “four Ts” plus one, namely, technology, talent, tolerance, territorial assets and experimentation of constantly introducing new ideas, products and processes [9, 19, 26]. Florida [10] maintains that the most important development resource in the emerging economy of the 21st century is creative talent. He further points out the driving force of social progress lies in the rise of human creative activities, and believes that the region can continue to develop creative industry when it has the three conditions of talent, technology and tolerance. Xu [33] from Hong Kong maintains that talent, society, cultural resources and infrastructure are key elements for a region to develop CCIs. Chen and Ge [2] believe that institutions, environment, talents and culture are important factors influencing the location choice of CCIs, among which culture is the foundation, institution is the guarantee, talent is the key and the environment is the support. Zhang et al. in [35] analyze factors affecting Beijing’s CCIs and find that government policies and cultural environment are very important to Beijing’s CCIs development. Wen and Hu in [32] deem that technologic-al factors, tolerance and talents are the main factors affecting the development of China’s provincial CCIs, while the influence of infrastructure and government policy is relatively small.

Another strand of literature analyzes the influencing factors of CCIs from the perspective of industrial agglomeration. Rumpel et al. (2010) maintains that urban areas usually attract CCI enterprises, forming specialized clusters. MIT (2011) regards CCI clusters as one of the variables that influence the location decisions of CCI enterprises. Wang [28] points out that “agglomeration” is a general development modeling of CCIs at the initial stage, and it is a multidimensional dynamic evolution process, which can be divided into a starting stage of factor agglomeration, a take off stage of fusion penetration, and a mature stage of radiation linkages. Wang [28] believes that CCIs have an obvious trend of regional agglomeration, which is an important mode of creative industry agglomeration. The supplier market, labor market and accompanying knowledge spill over process formed in the process of agglomeration will enhance the competitiveness of CCIs. Wang [29] finds that there are regional differences in the effects of industrial structure, human capital and industrial policies on the concentration of urban cultural industries in different regions. Wang et al. [31] identified main factors affecting the development of Macao’s cultural industry using grey relational method, which include demand capacity for cultural industry, government support, talents and related industries. Meng et al. [21] conducted similar study from the perspective of supply and demand in the cultural market.

In summary, most of the research focuses on individual factors of creative industry or analyzing particular factors of CCIs from an empirical perspective, seldom do these literatures put the CCIs in the context of social and economic development to comprehensively analyze the environmental factors affecting the development of CCIs, and try to construct a good and comprehensive environment conducive to the development of these industries. In this paper, grey relational analysis and grey factor analysis are used to select relevant indicators that are closely related to the development of CCIs. By systematic classification, the impact degree of environmental factors on CCIs is measured, and the environmental factors that affect CCIs are identified. After that, counter measures for developing CCIs are proposed.

Shanghai is a region with a relatively high level of development of CCIs in China. In 2015, Shanghai’s comprehensive index of cultural industry development exceeded Beijing for the first time, ranking first in the country. Differ from other international metropolises which started the CCIs in the post-industrial era, such as London, New York and Tokyo, Shanghai is developing CCIs under the background of rapid industrialization, which has its unique development path and characteristics. Therefore, taking Shanghai as an example to study the influencing factors of the development of CCIs is of great importance and has reference value for promoting the sustainable and healthy development of CCIs in developing countries. However, relevant literatures concerning the influencing factors of CCIs development in Shanghai is very rare, only Chu [4] analyzes the rules in the spatial agglomeration of CCIs in Shanghai, and Chu and Huang [5] taking Shanghai as a case study to analyze the geographical location factors that shape the In-city location of CCI parks. Both of them only focus on the geographical distribution of Shanghai’s CCIs while ignoring the influencing factors and their impacts.

The contribution of this paper is as follows: first, as CCIs are highly valued worldwide and taking Shanghai, a newly developed center of CCIs in an emerging country, as an example, this article gives the environmental factors that affect the development of CCIs, which has reference value for promoting the sustainable and healthy development of CCIs elsewhere. Second, compared with the existing research on the development of Shanghai's CCIs, this article puts forward some innovative views and suggestions. It is pointed out that "innovative R&D" indicators such as the amount of city patent grants, the total amount of enterprise R&D investment in science and technology, and the total amount of government R&D investment in science and technology are highly correlated with the added value of the CCIs, and the government's financial support and investment in the CCIs play an important role. Third, this paper is also innovative in the research method, gray factor analysis can better find out the factors that affect the development of Shanghai's CCIs and analyze its relevance and relative importance.

3. Indicator Selection and Methodology

3.1. Selection of Indicators

Selection of indicator is an important step in the analysis of impact factors, which is directly related to the reliability and scientific nature of the research results. Based on relevant research of scholars at home and abroad, following the principles of comprehensiveness, authenticity, comparability and availability, after investigation and consultation with industry experts in the cultural industry, this paper selects 30 indicators closely related to CCIs development to analyze, which is shown in Table 1 (taking Shanghai as an example). The comprehensive principle in this paper refers to the fact that the entire indicator system should reflect the development status of the CCIs and the predictable development capabilities in the future. The principle of authenticity means that the selected indicators should truly reflect the development of the CCIs, try to eliminate the individual's subjective preference for indicators and choose an objective and fair indicator system. The principle of comparability means that each element in the set of indicators must be consistent in terms of calculation caliber, measurement time and measurement unit; while the principle of availability refers to whether the relevant data of the indicator is available when the indicator is selected.

3.2. Selection of Methods

Grey Relational Analysis. With the complexity of social and economic systems, the structural relationship between them constantly changes and the data of various indicators are characterized by incompleteness and uncertainty. As a new industrial form of social and economic development, indicators for CCIs also possess characteristics of incompleteness and uncertainty. Therefore, this paper uses the Grey Relational Analysis (GRA) to empirically analyze the impacting factors of Shanghai CCIs. GRA is a multi-factor statistical analysis approach (Yin, 2018), by analyzing the sample data of each factor, the relational degree between the factors such as strength, weakness, size and order is measured by the grey relational degree according to the similarity between the developmental trends of the factors. If the trend of the two factors reflected by the sample data is basically the same, the degree of relation between the factors is relatively large; conversely, the relational degree is small.

Table 1. Main Indicators Affecting CCIs

Code	Impact indicator	Code	Impact indicator	Code	Impact indicator
X_1	The city's art performances	X_{11}	Infrastructure construction investment in the city	X_{21}	Number of employees in the city's CCIs
X_2	Variety of books published in culture, education, science and sports	X_{12}	The average number of mobile phones per 100 households in the city	X_{22}	Number of students in regular HEIs in the city
X_3	Number of international exhibitions held in the city	X_{13}	Urban per capita parks and green areas	X_{23}	Number of ordinary colleges and universities in the city
X_4	Share of registered population to permanent resident population	X_{14}	The main business income of the city's cultural and entertainment institutions	X_{24}	Number of international students in the city
X_5	Per capita disposable income of urban residents in the city	X_{15}	Labor productivity of the tertiary industry in the city	X_{25}	Number of scientific research personnel in the city
X_6	Per capita annual cultural and entertainment service expenditure	X_{16}	Import and export of cultural products and services in Shanghai	X_{26}	The total amount of enterprise technology R&D in the city
X_7	Consumer price index of entertainment, education, cultural goods and services for city residents	X_{17}	Government investment in technology R&D	X_{27}	Number of cultural industry research institutions in the city
X_8	Percentage of family culture consumption in the total consumption in the city	X_{18}	Government spending on environmental protection	X_{28}	Total technology contract value in the city
X_9	Per capita GDP of the city	X_{19}	Government expenditure on cultural sports and media	X_{29}	The number of patent grants in the city
X_{10}	Number of cultural creative industrial parks in the city	X_{20}	The actual amount of foreign capital utilized in the city	X_{30}	The output value of new products of large and medium-sized industrial enterprises in the city

Since the "added value of CCIs" can measure the development level of Shanghai's CCIs to a large extent, this sequence of variables is used as a reference series of model, and is recorded asX_0. The sequence represented by the 30 indicators is selected as the comparison series of the model, and each series is listed as $X_i = (i = 1, t...30)$.We use GRA to analyze the correlation between reference series and comparison series. First, we average the values of each series, and then calculate the absolute value difference between the two series in the same period. Let $H0i(t)$ be the relational degree between variable i and X_0 (the added value of CCIs) at period t.

$$H_{0i}(t)=\frac{\Delta\min+\rho\Delta\max}{\Delta oi(t)+\rho\Delta\max} \tag{1}$$

$\Delta oi(t)$ is the absolute value difference between variable i series and reference series, while $\Delta min$and Δmaxrepresent the minimum and maximum absolute value difference of each period separately, and ρ is the resolution coefficient.

By calculating the relational degree between the reference series and variables in the comparison series, the relational degree and ranking between Shanghai's CCIs and the related indicators are explored. After that, the impact factors and their rankings that affect the development of Shanghai's CCIs are analyzed

Grey Factor Analysis. Based on the GRA of the added value of Shanghai CCIs and related indicators, the Grey Factor Analysis method is used to analyze the factors affecting the added value of Shanghai CCIs, and to find out the grey common factors and main influencing factors. Assume series$X_i = [x_i(1), x_i(2)\dots x_i(n)]^T$, $(i = 1,2,\dots,p)$,and $X = (X_1, X_2, \dots X_n)^T$, let $V = (\varepsilon_{ij})_{pxp}$be the grey absolute relational matrix of X, then:

$$V = \begin{bmatrix} 1 & \varepsilon_{12} & \varepsilon_{13}\dots\varepsilon_{1p} \\ \varepsilon_{21} & 1 & \varepsilon_{23}\dots\varepsilon_{2p} \\ \dots\dots\dots\dots\dots\dots \\ \varepsilon_{p1} & \varepsilon_{p2} & \varepsilon_{p3}\dots 1 \end{bmatrix} \tag{2}$$

Set $X = (X_1, X_2, \dots X_n)^T$be a random vector of P measurable indicators, and then the mathematical model of Grey Factor Analysis can be expressed as:

$$X_{px1} = A_{pxm} F_{mx1} + \varepsilon_{px1} \tag{3}$$

Where$p \leq m$;F= $(\mathrm{F}_1, F_2 \cdots F_m)^T$ and $\varepsilon =$ $(\varepsilon_1, \varepsilon_2 \cdots \varepsilon_p)^T$are both random vectors. $A = (a_{ij})pxm$ is a constant matrix. $F$is the common factor of X, $\varepsilon$is the special factor of$X$, a_{ij} refer to the factor load, while matrix A is the factor load matrix.

4. Empirical Analysis

4.1. Data Processing

Data Selection. This paper selects data of 30 indicators closely related to the development of CCIs in Shanghai from 2013 to 2017, and analyzes the environmental factors affecting the development of Shanghai's CCIs through empirical analysis. All data are from documents or reports of the Shanghai Municipal Bureau of Statistics and relevant government sectors. Due to statistical limitations, data from certain years are lost, but this has little effect on the results of statistical analysis and can be ignored. The specific data is shown in Table 2.

Table 2. Relevant Indicator Data of Shanghai CCIs from 2013-2017

Year Item		**2013**	**2014**	**2015**	**2016**	**2017**
X_0	Value added in cultural and creative industries（billion Yuan）	250.00	282.00	302.00	330.00	339.50
X_1	The city’s art performances (times)	33910	27970	28730	22930	26140
X_2	Variety of books published in culture, education, science and sports	24969	24676	25954	27462	27772
X_3	Number of international exhibitions held in the city (times)	247	258	292	287	293
X_4	Share of registered population to permanent resident population	0.5930	0.5931	0.5974	0.5992	0.6017
X_5	Per capita disposable income of urban residents in the city (Yuan)	43851	47710	52962	57692	62596
X_6	Per capita annual cultural and entertainment service expenditure (Yuan)	4122	4931	3718	4174	4686
X_7	Consumer price index of entertainment, education, cultural goods and services for city residents	100.1	101.8	100.3	102.7	100.9
X_8	Percentage of family culture consumption in the total consumption in the city（%）	14.6	16.2	15.0	15.8	16.4
X_9	Per capita GDP of the city (Yuan)	90993	97370	106009	116582	126634
X_{10}	Number of CCI parks in the city	139	158	165	174	183
X_{11}	Infrastructure construction investment in the city（billion Yuan）	104.331	105.725	142.508	155.187	170.522
X_{12}	The average number of mobile phones per 100 households in the city	279	298	292	301	303
X_{13}	Urban per capita parks and green areas (square meters)	13.38	13.79	7.6	7.8	8.1
X_{14}	The main business income of the city’s cultural and entertainment institutions（billion Yuan）	49.222	50.585	27.329	193.415	65.140
X_{15}	Labor productivity of the tertiary industry in the city（Yuan/person）	200121	213541	22043	22969	235500
X_{16}	Import and export of cultural products and services in Shanghai（billion Yuan）	26.96	27.94	30.28	32.13	33.74
X_{17}	Government investment in technology R&D（billion Yuan）	77.678	86.195	93.614	104.932	120.521
X_{18}	Government spending on environmental protection (billion Yuan)	60.788	69.989	70.883	82.357	92.353
X_{19}	Government spending on cultural sports and media (billion Yuan)	3.13	3.22	3.53	3.85	4.13
X_{20}	The actual amount of foreign capital utilized in the city(billion US$)	16.780	18.166	13.013	1.867	1.876
X_{21}	Number of employees in the city’s CCIs (10,000)	130.00	135	257758	248322	222667

X_{22}	Number of students in regular higher education institutions in the city (10,000)	50.48	50.66	51.16	51.47	51.49
X_{23}	Number of ordinary colleges and universities in the city	68	68	67	64	64
X_{24}	Number of international students in the city	18970	23702	29242	31416	31941
X_{25}	Number of scientific research personnel in the city	590	595	602	623	636
X_{26}	The total amount of enterprise technology R&D in the city(billion Yuan)	40.478	44.922	47.424	49.008	54
X_{27}	Number of cultural industry research institutions in the city	10	12	12	14	15
X_{28}	Total technology contract value in the city (billion Yuan)	62.087	66.799	70.799	82.286	86.753
X_{29}	The number of patent grants in the city	48680	50488	60623	64230	72806
X_{30}	The output value of new products of large and medium-sized industrial enterprises in the city (billion Yuan)	681.102	740.799	731.224	779.449	915.991

Source: Shanghai Statistical Yearbook (2013-2018) and CCIs Statistics Bulletin.

4.2. Empirical Analysis

CCIs rely on human skills, creativity and wisdom to process and create cultural resources through high tech means, and use intellectual property rights to protect cultural creative products to meet people's cultural needs. The development of CCIs is related to various factors. Through the analysis of the grey relational degree between the added value of Shanghai CCIs and various related indicators, the relational degree and ranking of each relevant indicator are obtained, as shown in Table 3.

It can be seen from Table 3 that if the GRA analysis is only performed on the added value of CCIs and related indicators, main factors affecting Shanghai's CCIs cannot be clearly seen. Therefore, this paper uses SAS 9.2 software to further analyze the grey factors of various environmental indicators affecting CCIs. The cumulative contribution rate of the first six principal components has reached 84.775%, and the interpretation effect is good. In order to further grasp the economic significance of each grey factor, orthogonal rotation processing of the factor load matrix is performed to obtain a rotating load matrix, we get 6 grey common factors that affect Shanghai's CCIs, and name them according to their intrinsic characteristics. The details are shown as in Table 4.

Table 3. Ranking of Relational Degree between Shanghai CCIs and Related Indicators

Rank	Impact indicator	Relational degree	Rank	Impact indicator	Relational degree
1	The number of patent grants in the city (X_{29})	0.6689	16	Total technology contract value in the city (X_{28})	0.4187
2	The total amount of enterprise technology R&D in the city (X_{26})	0.6467	17	Variety of books published in culture, education, science and sports (X_2)	0.3998
3	Government investment in technology R&D (X_{17})	0.5895	18	Urban per capita parks and green areas (X_{13})	0.3975
4	The main business income of the city's cultural and entertainment institutions (X_{14})	0.5113	19	The average number of mobile phones per 100 households in the city (X_{12})	0.3888
5	The output value of new products of large and medium-sized industrial enterprises in the city (X_{30})	0.5098	20	Number of ordinary colleges and universities in the city (X_{23})	0.3831
6	Number of employees in the city's CCIs (X_{21})	0.4938	21	Per capita GDP of the city (X_9)	0.3765
7	Per capita disposable income of urban residents in the city (X_5)	0.4879	22	Percentage of family culture consumption in the total consumption in the city (X_8)	0.3789
8	The city's art performances (X_1)	0.4789	23	Consumer price index of entertainment, education, cultural goods and services for city residents (X_7)	0.3787
9	The actual amount of foreign capital utilized in the city (X_{20})	0.4786	24	Share of registered population to permanent resident population (X_4)	0.3661
10	Number of cultural creative industrial parks in the city (X_{10})	0.4687	25	Number of cultural industry research institutions in the city (X_{27})	0.3657
11	Government spending on environmental protection (X_{18})	0.4536	26	Number of scientific research personnel in the city (X_{25})	0.3635
12	Number of international students in the city (X_{24})	0.4501	27	Number of students in regular higher education institutions in the city (X_{22})	0.3543
13	Labor productivity of the tertiary industry in the city (X_{15})	0.4468	28	Number of international exhibitions held in the city (X_3)	0.3345
14	Per capita annual cultural and entertainment service expenditure (X_6)	0.4446	29	Infrastructure construction investment in the city (X_{11})	0.2885
15	Import and export of cultural products and services in Shanghai (X_{16})	0.4256	30	Government expenditure on cultural sports and media (X_{19})	0.2395

Source: Authors' calculation

Table 4. Classification of Each Indicator and Naming of Grey Factors

Factor name	Indicators	Factor load
Cultural industry science and technology environment factor	Number of scientific research personnel in the city (X_{25})	0.738
	The total amount of enterprise technology R&D in the city (X_{26})	0.887
	Number of cultural industry research institutions in the city (X_{27})	0.746
	Total technology contract value in the city (X_{28})	0.789
	The number of patent grants in the city (X_{29})	0.897
	The output value of new products of large and medium-sized industrial enterprises in the city (X_{30})	0.793
Cultural industry infrastructure environment factor	Number of CCI parks in the city (X_{10})	0.874
	Infrastructure construction investment in the city (X_{11})	0.879
	The average number of mobile phones per 100 households in the city (X_{12})	0.715
	Urban per capita parks and green areas (X_{13})	0.779
	The main business income of the city’s cultural and entertainment institutions (X_{14})	0.798
	Labor productivity of the tertiary industry in the city (X_{15})	0.778
	Import and export of cultural products and services in Shanghai (X_{16})	0.867
Funding and policy environment	Government investment in technology R&D (X_{17})	0.935

Factor name	Indicators	Factor load
Cultural consumption environment factor	Share of registered population to permanent resident population (X_4)	0.732
	Per capita disposable income of urban residents in the city (X_5)	0.878
	Per capita annual cultural and entertainment service expenditure (X_6)	0.879
	Consumer price index of entertainment, education, cultural goods and services for city residents (X_7)	0.923
	Percentage of family culture consumption in the total consumption in the city (X_8)	0.912
	Per capita GDP of the city (X_9)	0.815
Cultural industry human resource environment factor	Number of employees in the city’s CCIs (X_{21})	0.815
	Number of students in regular higher education institutions in the city (X_{22})	0.832
	Number of ordinary colleges and universities in the city (X_{23})	0.782
	Number of international students in the city (X_{24})	0.756
Socio-cultural environment factor	The city’s art performances (X_1)	0.867
	Variety of books published in culture, education, science and	0.756

factor				sports (X_2)	
	Government spending on environmental protection (X_{18})	0.878		Number of international exhibitions held in the city (X_3)	0.817
	Government expenditure on cultural sports and media (X_{19})	0.847			
	The actual amount of foreign capital utilized in the city (X_{20})	0.765			

4.3. Result Analysis

Analysis of Various Environmental Factors of High Relevance. As can be seen from the analysis results, factors that have a high relational degree with Shanghai's CCIs development are X_{29},X_{26},X_{17},X_{14} and X_{30}, with the relational coefficient being 0.6689, 0.6467, 0.5859, 0.5113 and 0.5098 separately, indicating that these five indicators have a greater impact on the development of Shanghai's CCIs. The "innovative R&D" indicators, such as (The number of patent grants in the city), (The total amount of enterprise technology R&D in the city), (Government investment in technology R&D), (The output value of new products of large and medium-sized industrial enterprises in the city) have a high relational degree with the added value of Shanghai's CCIs, indicating that the government is very important in providing financial support and investment to Shanghai's CCIs development. This shows that the government's support policies, science and technology R&D, cultural industry foundation and cultural environment are closely related to the development of CCIs, which is also an important focus for enhancing the competitiveness of Shanghai's CCIs.

Analysis of Environmental Factors of Medium and Low Relevance. As is shown in Table 3, X_{25} (Total technology contract value in the city),X_{22} (Number of students in regular higher education institutions in the city),X_3 (Number of international exhibitions held in the city),X_{11} (Infrastructure construction investment in the city) andX_{19} (Government expenditure on cultural sports and media) rank the last five in relational degree, with the relational coefficient being 0.3635, 0.3543, 0.3345, 0.2885 and 0.2395 separately, indicating that these five indicators do not contribute much to the development of Shanghai's CCIs. Among them, X_{19} has the lowest relational degree with Shanghai's CCIs, only 0.2395, signifying that the government's investment in CCIs is not balanced, and that government investment is generally insufficient relative to industry demand. Enterprises in the CCIs need more financing channels and a more relaxed financing environment. The government should try its best to meet the financing requirements of SMEs in the industry through policy guidance and mechanism design in the process of supporting the development of CCIs. At the same time, the relational degree of X_{11} (Infrastructure construction investment in the city) and X_{10} (Number of CCI parks in the city) are both not very high, being 0.2885 and 0.4687 separately, demonstrating that Shanghai's good cultural industry infrastructure construction and the existence of numerous creative industrial parks do not play an important role in the development of Shanghai's CCIs. After careful studying, we found that although the industrial agglomeration of the park has begun to take shape and the infrastructure in the park is good, due to problems in the management system and operation mechanism, the clustering effect of the park has not been fully exerted. Furthermore, the development of the parks is uneven, the function of the park is not accurate, with a lack of rational overall strategic planning and refined management, which may explain medium and low relevance of X_{10} and X_{11}.

The relational coefficient of X_8 (Percentage of family culture consumption in the total consumption in the city) and X_7 (Consumer price index of entertainment, education, cultural goods and services for city residents) are 0.3789 and 0.3787 respectively, representing a low degree of relevance, indicating that there is still much room for Shanghai urban residents to improve in cultural consumption. The relational coefficient of X_{27} (Number of cultural industry research institutions in the city) and X_{25} (Number of scientific research personnel in the city) are 0.3657 and 0.3635 separately, representing a relatively low degree. Despite that the output value of the CCIs in Shanghai has increased year by year, Shanghai's research strength and investment in the industry have not kept up with the actual needs of it, therefore, more investment and support should be encouraged. Meanwhile, relational coefficients of X_2 (Variety of books published in culture, education, science and sports),X_3 (Number of international exhibitions held in the city) andX_4 (Share of registered population to permanent resident population) are all below 0.4, showing that the socio-cultural environment and cultural consumption environment for the development of Shanghai's CCIs need to be further ameliorated

Main environmental factors affecting Shanghai's CCIs development. Table 4 shows that factors affecting Shanghai's CCIs development can be attributed to six grey common factors, namely: Cultural industry science and technology environment factor, Cultural industry infrastructure environment factor, Cultural consumption environment factor, Cultural industry human resource environment factor, Funding and policy environment factor and Socio-cultural environment factor. Among them, cultural industry science and technology R&D, government policy and financial support, cultural industry human resources and cultural consumption environment are important factors for Shanghai's CCIs development. For further promoting the sustainable and healthy development of CCIs, efforts must be made to improve the socio-cultural environment, build a good cultural consumption environment, vigorously cultivate the innovative and professional talents of the cultural industry, and provide human resources support for Shanghai CCIs. Moreover, policy guidance and financial support for CCIs also need to be strengthened.

5. Conclusions and Discussion

CCIs play a vital role in promoting coordinated economic and social development, stimulating economic restructuring and upgrading, and strengthening urban soft power. To enhance the core competitiveness of the CCIs, make full use of it in the construction of urban culture, promote the regional economy of "innovation-driven and transformational development", this paper proposes countermeasures for the environmental construction that promotes the sustainable and healthy development of CCIs based on the above research results.

First, a good cultural creative environment should be cultivated and a vibrant cultural consumption atmosphere should be built. It can be seen from the above research that among the impact factors of Shanghai's CCIs, the "cultural consumption environment factor" is an important one. Moreover, the various components of the cultural consumption environment factor (see Tables 3 and 4) are highly related to Shanghai CCIs. Since its inception, Shanghai has been the forefront of the collision between Chinese and Western cultures, forming a Shanghai culture with innovative ideas and inclusive characteristics. In addition to that, Shanghai also has a wealth of revolutionary cultural resources and historical cultural resources. Consequently, to promote the development of CCIs, Shanghai needs to comprehensively utilize and integrate cultural resources, creatively promote the transformation of cultural resources into cultural products and their derivatives, actively introduce excellent cultural products from abroad to meet people's demand, at the same time, to construct a vigorous cultural consumption atmosphere to constantly enrich the cultural life of the masses, and improve people's quality of life and living standards.

Second, emphasis should be put on the cultivation and introduction of creative talents to build a talent highland for CCIs. As the CCI is based on human creativity, the quality of talents determines the development level of a region's CCIs. This study shows that cultural creative talent resource index is not highly related to the development of Shanghai's CCIs (relational degree is only 0.4896), and the supply of creative industry talent cannot keep up with the rapid development of Shanghai's CCIs. The lack of talents, especially the complex and professional senior management talents, and the incomplete talent cultivation mechanism have become the bottleneck restricting the development of Shanghai's CCIs. Therefore, nurturing and introducing creative talent is the key. On the one hand, advanced experience of cultural creative talents training should be learned, the new training mode of "production-study-research" should be used to strengthen the school-enterprise cooperation and establish a talent practice base to cultivate more professional talents who are familiar with cultural attributes and operational rules, as well as cultural enterprises management. On the other hand, the government and enterprises should broaden their horizons, provide relevant preferential policies and treatments, strengthen cooperation internationally, and vigorously introduce talents with outstanding cultural industry creativity and cultural enterprise management, so as to build a CCI highland and promote the industry develop healthily and rapidly.

Third, the main body of the cultural market should be cultivated and diversified. Cultural creative enterprises are the main force and carrier to promote the development of CCIs. Only by vigorously cultivating CCI enterprises with distinctive characteristics and strong strength, and promoting the diversification of the main body of CCIs, can we more effectively promote the rapid development of CCIs. The research shows that the number of cultural creative enterprises in Shanghai is very large, but there are not many strong cultural creative enterprises. Therefore, a fairer and more open market competition mechanism should be established to encourage the healthy competition and interdependence of all kinds of market subjects, and promote the diversification of cultural industry subjects. At the same time, the government should give cultural enterprises more supportive policies in various aspects, such as finance, taxation, land, etc., and cultivate a number of large cultural enterprise groups with international competitiveness and influence to play a leading role in the development of CCIs.

Fourth, financing channels of creative industry should be broadened and a diversified investment and financing pattern should be constructed. Strong financial support and multi-channel sources of funds are important basis for the development of CCIs. The results of this study also show that financial support is closely related to the development of CCIs in Shanghai, and the government financial support is an important driving force for CCI development. However, relying mainly on the input and support of the government, not widening the financing channels and building a diversified investment and financing situation is bound to affect the future development of Shanghai's CCIs. Moreover, the CCI in most places consists mainly of SMEs, which have limited financing channels, and relatively high financing threshold. Therefore, to promote the development of CCIs, focus should be put on promoting the construction of financial markets related to CCIs, encouraging financial institutions to develop financing products suitable for SMEs and broadening the financing channels for SMEs. Meanwhile, the government should improve its financial support planning, strengthen its financial support for small and medium-sized cultural enterprises with potential and innovation, and form a number of investment and financing platforms for the development of cultural industries led by the government to absorb social capital into CCIs.

Fifth, creative industrial park resources should be integrated and the agglomeration advantage of industrial park should be given full play. There are many problems in cultural creative parks in China, such as "emphasizing form, neglecting business form", low level of professional services in the region, serious homogeneous competition among parks, unsound management standards in the region, and uneven development of various parks. This study also shows that the relational degree between the infrastructure construction index of Shanghai's cultural industry and the development of Shanghai's CCIs is low (0.2885), which is one of the five indicators with the weakest correlation. While the relational degree of the number of cultural creative industrial parks and industrial agglomeration area of Shanghai is 0.4687, which is in the middle of the ranking table of the overall index relational degree. This is not in line with the good infrastructure construction of Shanghai's cultural industry and the fact that there are many creative industrial parks in Shanghai. Therefore, adopting various effective measures to integrate various resources of existing creative industrial parks and bring into play the clustering effect of industrial parks is an important path for the development of CCIs. First of all, efforts should be made to expand and strengthen a number of enterprises with international influence, form a driving role, and improve the overall competitiveness of the industrial park. Secondly, improve the infrastructure construction in the existing park, avoid repeated construction and integrate current clusters with the same functional orientation, transform the development model of the cluster, strengthen the individualized management of enterprises in the park, and stimulate the innovation ability of enterprises. Thirdly, clarify characteristics of the park under construction, try the third-party audit system for the park's efficiency, pay attention to fostering the symbiotic relationship between the enterprises in the park, focus on the integration and development of the CCIs and other industries, and take full advantage of the gathering function of the creative industry park.

Finally, efforts should be made to constantly improve the policy and regulation system, and build a good business environment. As a new industry, CCI needs government's policy support and guidance. It can be seen from this study that the rapid development of Shanghai's CCIs cannot be realized without the government's vigorous promotion or even direct leadership; and the government has served an important role in guiding and promoting the construction of cultural industry infrastructure and social and cultural environment. However, the government should respect the market position of enterprises, work hard to improve the market environment for CCIs, strengthen the infrastructure construction for the development of CCIs, formulate and improve the regulatory system for cultural market and optimize the ecological environment for CCIs. Relevant government departments should, in light of the actual situation of CCIs in the region, formulate various regulatory systems and policy regulations that promote the development of CCIs with local regional characteristics. Meanwhile, it is necessary to strengthen the intellectual property rights protection, encourage the export of cultural creative products, and promote CCI enterprises to go global. Actively attract and utilize foreign capital to invest in CCIs in the region, strengthen exchanges and cooperation between international advanced cultural creative enterprises and local cultural creative enterprises, and build a large market environment that is open, transparent, efficient and fair in favor of the development of CCIs.

Acknowledgment. This work is supported by the Youth foundation for Humanities and Social Sciences of Ministry of Education of the People's Republic of China (No. 19YJCGJW007) and the project of Shanghai Philosophy and Social Sciences Plan (No. 2018BGL023).

References

1. Caves, R.: Creative Industries-Contracts between Art and Commerce, Harvard University Press, Cambridge, MA. (2000)
2. Chen, J., Ge, B.: Analysis on the Agglomeration Effect and Influence Factors of Cultural and Creative Industries, Contemporary Economy & Management, Vol.30, No.9, 71-75. (2008). Available from: http://doi.org/10.4135/9781848608443.n27.
3. Cho, R, L.T., Liu, J. S., Ho, M. H. S.: What Are the Concerns? Looking Back on 15 Years of Research in Cultural and Creative Industries. International Journal of Cultural Policy, Vol.24, No. 1, 25-44. (2018)
4. Chu, J.F.: A Study on Spatial Difference of the Creative Industrial Zones in Shanghai. Human Geography, Vol.24, No. 2, 23-28. (2009)
5. Chu, L.X., Huang, L.: Research on the Geographical Location Factors that Shape the In-city Location of Cultural and Creative Industrial Parks: a Case Study of Shanghai. Modern Urban Research, No.1, 37-41. (2019)
6. EC, European Commission: Green Paper. Unlocking the Potential of Cultural and Creative Industries, COM183 final, April (2010)
7. Fausto, C.: Reviewing the Cultural Industry: from Creative Industries to Digital Platforms. Communication & Society, Vol.31, No.4, 135-146. (2018)
8. Florida, R.: Cities and the Creative Class. City & Community, Vol.2, No.1, 3-19. (2010)
9. Florida, R.: The Rise of the Creative Class: and How it’s Transforming Work, Leisure, Community and Everyday Life. Canadian Public Policy. (2003)
10. Florida, R.: “The Experiential Life”, in Hartley, J.(ed.), Creative Industries, Blackwell Publishing, Malden, MA and Oxford, 133-145.(2005)
11. Gibson, C., Klocker. N.: Academic Publishing as ‘Creative’ Industry, and Recent Discourses of ‘Creative Economies’: Some Critical Reflections. Area, Vol.36, No.4, 423-434.(2004)
12. Gibson, C., & Kong, L.: Cultural Economy: A Critical Review. Progress in Human Geography, Vol.29, No.5, 541-561. (2005)

13. Gibson, C., Luckman, S., Willoughby-Smith, J.: Creativity Without Borders? Rethinking Remoteness and Proximity. Australian Geographer, Vol.41, No.1, 25-38. (2010)
14. Gundolf, K., Jaouen, A., Gast, J.: Motives for Strategic Alliances in Cultural and Creative Industries[J]. Creativity and Innovation Management, No.3:1-13. (2017)
15. Harvey, D.C., Hawkins, H., Thomas, N.J.: Thinking Creative Clusters Beyond the City: People, Places and Networks. Geoforum, Vol.43, No.3, 529-539. (2012)
16. Howkins, J.: The Mayor’s Commission on the Creative Industries, in Hartley, J.(ed.), Creative Industries, Blackwell Publishing, Malden, MA and Oxford, 117-125. (2005)
17. Huggins, R., Clifton, N.: Competitiveness, Creativity, and Place-based Development. Environment and Planning A, Vol.43, No.6, 1341-1362. (2011)
18. John Howkins,: The Creative Economy: How People Make Money From Ideas, Penguin UK. (2001)
19. Justin, O.: Intermediaries and Imaginaries in the Cultural and Creative Industries. Regional Studies, Vol.49, No.3:374-387. (2015)
20. Kong, L.: Improbable Art: The Creative Economy and Sustainable Cluster Development in A Hong Kong Industrial District. Eurasian Geography and Economics, Vol.53, No.2, 182-196. (2012)
21. Meng, S., Lei, Y.: A Study on the Influencing Factors of the Development of Chinese Cultural Industry, Statistics & Decision, Vol.35, No.7,100-104. (2019)
22. O’Connor, J.: The Cultural and Creative Industries: A Review of the Literature, a Report for Creative Partnerships. London: Arts Council England. (2007)
23. Polese, M.: The Arts and Local Economic Development: Can A Strong Arts Presence Uplift Local Economies? A study of 135 Canadian Cities. Urban Studies, Vol.49, No.8, 1811-1835. (2012)
24. Scott, A.J.: A New Map of Hollywood: the Production and Distribution of American Motion Pictures. Regional Studies, Vol.36, No.9, 957-975. (2002)
25. Smit, A.J.: The Influence of District Visual Quality on Location Decisions of Creative Entrepreneurs. Journal of the American Planning Association, Vol.77, No.2, 167-184.(2011)
26. Sophia, L.: The Smart Economy: Cultural and Creative Industries in Greece. Can They Be A Way Out of the Crisis? Economic Bulletin, No.39, 73-103.(2014)
27. UNCTAD Creative Economy Report 2010. A Feasible Development Option, United Nations. (2010)
28. Wang, H.: Research on the Development of Shanghai Creative Industries Clusters, Journal of Social Sciences, No.7, 31-39. (2012)
29. Wang, M., Wang, Y.: Research on Influencing Factors of Urban Cultural Industry Agglomeration: Evidence from 35 Large and Medium Cities, Journal of Jiangxi University of Finance and Economics, Vol. 97, No.1, 13-20. (2015)
30. Wang, Z.: The Agglomeration Effect of Cultural and Creative Industries and the Problems It Faces, Economic Review Journal, No.8, 76-78. (2012)
31. Wang, Z., Song, S.: A Study on the Evaluation and Development Path of the Influencing Factors of Macao’s Cultural Industry, Commentary on Cultural Industry in China, Vol.26, No.1, 267-283.(2018)
32. Wen, H., Hu, B.: A Spatial Econometric Study on the Influence Factors of Provincial Cultural and Creative Industries Development in China, Economic Geography, Vol.34, No.2, 101-107. (2014)
33. Xu, C.: Cultural and Creative Industries in Hong Kong: A New Perspective and Strategy, Exploration and Free Views, No.8, 30-31. (2007)
34. Yin, K. D., Xu, Y., Li, X. M., Jin, X.: Sectoral Relationship Analysis on China's Marine-land Economy Based on a Novel Grey Periodic Relational Model. Journal of Cleaner Production, No.197, 815-826. (2018)
35. Zhang, W., Yao, H.: The Empirical Analysis on the Influential Factors of Beijing Cultural and Creative Industries, Social Science of Beijing, No.3, 20-25. (2011)

Jianjun Li is an associate professor at the Management School, Shanghai University of International Business and Economics, China. He received his Ph.D. in School of business administration at Shanghai University of Finance and Economics, China. His research interests include Cultural and creative industries. Contact him at lijianjun@suibe.edu.cn.

Jia Liao is an associate professor at the School of Business, Shanghai University of International Business and Economics. She received her Ph.D. in School of Economics at Fudan University, China. Her research interests include International Trade and investment. Contact her at liaojia@suibe.edu.cn.

Received: January 19, 2021; Accepted: March 28, 2021.

Computer Science and Information Systems 18(4):1271–1286 https://doi.org/10.2298/CSIS200731020C

Conversational Agent for Supporting Learners on a MOOC on Programming with Java

Cristina Catalán Aguirre, Nuria González-Castro, Carlos Delgado Kloos,
Carlos Alario-Hoyos, and Pedro J. Muñoz-Merino

Universidad Carlos III de Madrid,
Av. Universidad, 30, 28911 Leganés, Madrid
{crcatalan@pa, nurigonz@db, cdk@it, calario@it, pedmume@it}.uc3m.es

Abstract. One important problem in MOOCs is the lack of personalized support from teachers. Conversational agents arise as one possible solution to assist MOOC learners and help them to study. For example, conversational agents can help review key concepts of the MOOC by asking questions to the learners and providing examples. JavaPAL, a voice-based conversational agent for supporting learners on a MOOC on programming with Java offered on edX. This paper evaluates JavaPAL from different perspectives. First, the usability of JavaPAL is analyzed, obtaining a score of 74.41 according to a System Usability Scale (SUS). Second, learners' performance is compared when answering questions directly through JavaPAL and through the equivalent web interface on edX, getting similar results in terms of performance. Finally, interviews with JavaPAL users reveal that this conversational agent can be helpful as a complementary tool for the MOOC due to its portability and flexibility compared to accessing the MOOC contents through the web interface.

Keywords: Conversational Agent, Computer Science, MOOC, Programming, Java.

1. Introduction

MOOCs (Massive Open Online Courses) are still one of the most important trends in online learning, allowing the interaction among learners with different backgrounds and from different origins. MOOCs have contributed to expand the access to quality content, especially in engineering and computer sciences as a good number of MOOCs are framed within these two areas of knowledge [1]. Moreover, some of the MOOCs are instructed by renowned teachers giving learners the opportunity to satisfy their appetite for learning even if they are not enrolled in any traditional university.

Nevertheless, MOOCs present important issues. For example, the fact that the courses are open, with no enrolment fees, goes hand in hand with a low commitment from a good number of participants [2]. Completion rates are usually below 10% of the enrolled learners [3]. Although this is not always a problem, since there are learners who take MOOCs just to explore the content they are interested in, there are some other learners who drop out because their level (or background) is not sufficient or because they are not able to take an online course autonomously. When it comes to engineering and computer sciences MOOCs, apart from the above-mentioned issues, learners have to struggle with the intrinsic difficulty of the contents.

Numerous studies have focused on analyzing dropout rates, especially in MOOCs on topics related to engineering and computer sciences, with the ultimate aim to enhance the course design or to introduce interventions that can serve to reduce these dropout rates [4][5]. For example, it is possible to detect if a learner is going to leave a MOOC earlier through the identification of behavioral patterns [6][7] and react by giving feedback or specific advice to that learner. For example, authors in [8] discuss the use of a mentoring program to motivate learners to not giving up before the end of the course, thus overcoming the limitation related to the lack of teacher support in MOOCs; however, this solution faces again the problem of scalability in MOOCs with a very large numbers of learners. In contrast, authors in [9] propose the use of conversational agents, eliminating the human component in order to better scale up, as a support to MOOC learners. The idea behind the use of conversational agents is to improve the learning impact through dialog (dialog learning) since the participation in a conversation gives learners a more active role in the learning activity they are doing [10]. However, the use of conversational agents in the learning process in MOOCs has not been deeply examined [11]. In consequence, the effectiveness and usability of conversational agents in MOOCs has yet to be evaluated.

A pioneering example of conversational agent to support learners enrolled in MOOCs is JavaPAL, for which a first prototype was briefly introduced in a "computers in education" conference [9]. JavaPAL is a voice-based conversational agent designed to support learners who are taking the MOOC on "Introduction to Programming with Java" deployed on the edX platform. The objective of JavaPAL is to facilitate the study and revision of concepts related to Java programming using dialog, providing definitions on key concepts and asking some related questions to the learners. Although JavaPAL operates independently from the MOOC (as a standalone conversational agent), all the concepts for which JavaPAL provides definitions, and all the questions asked by JavaPAL are taken (and adapted in the case of some questions) directly from the above-mentioned MOOC. JavaPAL is aimed at accompanying learners through the learning process and at serving as reinforcement of the basic concepts that learners need to grasp. Nonetheless, the usability of a conversational agent such as JavaPAL and its effect in the learning experience is still unknown [11]. Moreover, the differences between learners' interaction through a conversational agent and the traditional MOOC web interface have yet to be analyzed when it comes to reviewing key concepts and answering questions.

Thus, this paper aims to shed some light on the usability of question-driven conversational agents to support learners enrolled in MOOCs, and the comparison of learners' performance when answering questions through web and conversational agent interfaces, all this using the JavaPAL as an example case. Therefore, the research questions of this paper are:

RQ1: Can a question-driven conversational agent fulfil learners' expectations in terms of usability?

RQ2: Does the use of a conversational agent affect learners' performance when answering questions in MOOCs?

RQ3: What are the differences learners and teachers find between the use of a MOOC web interface and a conversational agent interface when reviewing concepts and answering questions?

The structure of the paper is as follows. Section 2 provides the background on conversational agents both text-based (chatbots) and voice-based, focusing on their application in education. Section 3 describes the methodology used to conduct this study, including a description of the conversational agent used, the process of collecting and analyzing the data and the tools used. The results of the analysis are presented in Section 4, answering the research questions. Section 5 draws the conclusions of this research work.

2. Related Work

The best way to express interest, wishes or queries directly and naturally to computers is by speaking (voice-based dialog) or typing (text-based dialog) since these kinds of communication facilitate Human Computer Interaction [12]. Conversational agents are dialog systems that not only conduct natural language processing but also respond automatically using human language [13]. Chatbots are defined as computer programs that mimic the human interaction by using a text-based conversational agent to provide the interaction [14]. Consequently, both text-based and voice-based conversational agents arise as the most appropriate technology to use to fulfil the extension of human communication with computers.

The increase in the presence of chatbots in society is such that there is an estimation of 80% of enterprises using them, and around 40% of companies having used one or more assistants or AI-based chatbots over mobile devices in 2019 [15]. However, these programs are not new at all: Eliza, the first chatbot, was created in 1964. Eliza was a textual chatbot that used simple keyword matching techniques to match user input: it was designed to simulate a psychotherapist [16]. Another example of chatbot is A.L.I.C.E., which was developed using Artificial Intelligence Markup Language (AIML). AIML consists of objects that contain topics and categories. Each category represents a rule to match a user input to generate an output. This match is based on the internal templates from A.L.I.C.E. [17].

It is important to notice that, in their early stages of development, chatbots were not intelligent systems since they provided some pre-programmed questions and gave specific and predetermined responses. Jia [18] highlights the idea of users being upset with the responses provided by basics chatbots, since their pattern-matching system can be considered insufficient to be used in a real conversation. However, with the development of artificial intelligence, conversational agents have the potential to learn and assume the role of humans in some areas, including education [11]. Thus, systems that can learn from their previous experiences using AI, like Edwing.ai [19], which can elaborate more personalized responses, can have a wider adoption.

The use of chatbots in education, for example to raise questions that can be answered by students, could help teachers detect weaknesses in their students, as well as identify concepts and topics which pose a greater challenge [20]. This idea can be particularly useful in engineering and computers sciences due to the usual complexity of the key concepts in these areas of knowledge. Furthermore, chatbots can give support to each individual student since they are in position to acknowledge strengths, abilities or interests, and encourage learners to be more independent and engaged. In addition, the development of new Machine Learning techniques has led to an important rise of a new generation of conversational agents for education, such as the one presented in [21], which relies on a Naïve Bayesian classifier to answer questions posted by students as if it were a virtual teacher. With the development of Machine Learning techniques and the improvement of Natural Language Understanding, conversational agents are expected to enhance their characteristics simplifying the communication between user (learner) and (conversational) agent.

Design strategies to build up conversational agents for education are diverse. Particularly, in e-learning conversational agents range from simple text-based forms to voice-based. All of them should share the same objective: acting as a partner for the students and taking a non-authoritative role in a social learning environment. Each design strategy reflects one of the possible conversational systems: 1) simple text-based forms in which users type their responses; 2) embodied conversational agents capable of displaying emotions and gestures; and 3) voice input and output systems able to synthesize text to speech and vice versa [22].

Intelligent tutoring systems (ITSs) have been widely studied for natural dialog in education, with some ITSs being framed under the definition of conversational agents. One example of ITS for education that can also be defined as a conversational agent is Adele, a pedagogical text-based agent to support learners in web-based educational simulations, such as medical simulations [23]. Another example ITS which acts as a text-based chatbot is TutorBot [24]. TutorBot follows a question/answer schema in which the learner can retrieve information from a knowledge source using natural language. Finally, one voice-based example of ITS is AutoTutor [25], which is able to keep conversations with humans in natural language and incorporates strategies of human tutors previously identified in human tutoring protocols [25].

Although the existence of conversational agents is far from being new, their application within MOOCs is still uncertain. Authors in [26] present a first approach on the use of conversational agents in MOOCs, although these authors only present the basis for the development of a conversational agent, but they do not actually implement it. Another example of conversational agent developed to support MOOC learners through dialog is Bazaar [27], a text-based chat tool that provides synchronous interaction to learners within MOOCs. Another example of the use of conversational agent in MOOCs is QuickHelper, whose aim is to help learners to reduce their reluctance to ask questions and increase the number of questions posted in the forums [11]. These pioneering conversational agents base their conversation with the MOOC learner on text rather than on voice so there is still a research gap on voice-based conversational agents to support MOOC learners.

Apart from the previous examples, conversational agents are becoming more important in online education and blended learning since the participation of conversational agents during a peer communication reinforces the knowledge about the topic by activating relevant cognitive activity [26]. Consequently, conversational agents in e-learning can also help to improve peer to peer interaction [26]. Furthermore, the integration of conversational agents in MOOCs may trigger productive interaction in group discussion, increase the engagement and the commitment of MOOC learners and, therefore, reduce the overall dropout rates [11]. In this line, authors in [9] introduced a first prototype of a voice-based conversational agent called JavaPAL, which enables the possibility of reinforcing the knowledge of the contents of a MOOC on “Introduction to

Programming with Java" using voice dialog. Since Java programming is one of the basics subjects of many engineering or computer science degrees, JavaPAL is meant to help MOOC learners in the process of learning the key concepts from the above-mentioned MOOC by reviewing these concepts and posing multiple choice questions to the learner.

Thanks to the development and the application of Natural Language Understanding, conversational agents are expected to improve their characteristics and functionalities in the next years. This can lead to an improvement in online education, especially within MOOCs, since conversational agents may allow learners to get support at any place by using their mobile phones, tablets or devices such as Google Home or Alexa. Particularly, MOOCs with more complex content, such as those related to engineering or computer sciences, can benefit from the advantages of conversational agents to improve the lack of the support they face. Consequently, some important problems MOOCs are currently facing can be addressed and blended education can be open to a new paradigm of opportunities in which conversational agents can play a major role.

3. Methodology

A mixed-methods design [28] was applied in this research work. More specifically, the mixed-methods sequential explanatory design was followed. First, quantitative data was collected through the SUS (System Usability Scale) questionnaire [29] and logs were obtained from the use of the conversational agent; these quantitative data served to answer the first two research questions. Then, qualitative data was collected through interviews with users of the conversational agent; these qualitative data served to answer the third research question. A controlled group of 39 users was selected to evaluate JavaPAL. The members of this group did not correspond to actual learners enrolled in the MOOC JavaPAL supports, but to a number of pre-selected users with several backgrounds (students, teachers, researchers) whose mission was to evaluate JavaPAL. All these users shared previous knowledge on the topic of the MOOC. More specifically, 39 users tested JavaPAL and subsequently filled in the SUS regarding the usability of the conversational agent (RQ1), and 15 of them participated later in a quasi-experimental design aimed at comparing JavaPAL and the MOOC web interface when answering questions, which resulted in the collection of evidences through logs (RQ2) and interviews (RQ3).

Next, there is an overview of JavaPAL, the conversational agent used in this research. The instruments for data collection and data sources used are further explained right after.

3.1. JavaPAL

JavaPAL is a voice-based conversational agent developed at Universidad Carlos III de Madrid with the aim of supporting learners enrolled in the MOOC on "Introduction to Programming with Java", which is deployed in the edX platform. JavaPAL operates as a standalone conversational agent and was implemented using the natural language understanding platform DialogFlow and runs on Google Assistant, so it can work on any device that supports Google Assistant (e.g., smartphone, Google Home, etc.) (more technical details can be found in [9]). In addition to supporting voice-based interaction as the main mode of operation, JavaPAL also supports text-based interaction if the user wants it. JavaPAL supports three operation modes (quiz, review, and status) and to enter any of these three operation modes the learner must start the conversation with JavaPAL and indicate the desired operation mode.

Quiz mode. This mode asks questions to the learner. These questions are taken from the MOOC on "Introduction to Programming with Java" and are arranged according to the modules of the MOOC (five modules). When the learner accesses the quiz mode, s/he indicates the module from which s/he wants to receive questions and JavaPAL provides random questions from that module. The total number of consecutive questions asked by JavaPAL is predefined to three, but this can be configured by the learner in the status mode. This means that the learner receives three questions of the same module before s/he can change to another module. In this operation mode, the conversational agent takes the initiative of the communication since it asks questions and the learner answers them. The questions are designed in a way that the learner only needs to give a word (options from "a" up to "e" for multiple choice questions or true/false) to make it easier to remember each option.

Review mode. This mode provides learners with the definition of the key concepts addressed in the MOOC. In this operation mode, the learner takes the initiative of the communication since s/he indicates JavaPAL the concept s/he wants to revise (e.g., "class", "object", "method", etc.) and then JavaPAL provides the corresponding definition. If the definition of the concept is not available, then JavaPAL offers the definition of another related concept. Moreover, besides offering the definition of the concept indicated by the learner, JavaPAL can suggest other related concepts to provide their definitions based on an ontology that relates concepts of Java programming [30].

Status mode. In this mode, the learner can check his/her performance during the quiz mode and change some settings, such as nickname or number of consecutive questions.

JavaPAL has been developed using an iterative approach. Prototypes have been developed and tested with real users, and aspects to be improved have been detected through surveys and interviews with the users, and subsequently implemented in the following prototypes.

3.2. Instruments and Data Sources

System Usability Scale (SUS). Standardized usability questionnaires were analyzed in a first phase. After this analysis, SUS (System Usability Scale) was chosen as it is more reliable and detects differences at smaller sample sizes than other questionnaires, and the number of items to be assessed is not very large (10 items, 5 options per item), which facilitates data collection [29]. The measurement of the validity has been calculated using Cronbach's alpha [31]. **39 answers** to the SUS questionnaire were obtained from users of JavaPAL. These answers came from users with different knowledge about Java, mixing students, teachers and researchers. SUS allowed to gather the information needed to answer *RQ1*.

Logs. JavaPAL has been designed to collect logs on users' interaction, users' performance (understood as the correct/incorrect answers in the quiz mode) and the concepts learners ask for (in the review mode). **15 users** participated in a quasi-experimental design aimed at collecting data through logs. These users had previous experience with Java programming and were divided into two groups. The first group was asked to use the conversational agent first and then the MOOC web interface, in both cases to answer questions related to the MOOC contents. The second group was asked to use the MOOC web interface first and then the conversational agent, again to answer questions related to the MOOC contents. The objective was trying to detect differences between both types of interfaces from the learner's performance. It is worth to mention that some questions had to be adapted for the conversational agent to be concise enough and easy to remember by the learners using voice interaction. Some example of these adaptations can be seen in Appendix A. JavaPAL logs allowed to gather the information needed to answer *RQ2.*

Interviews. An interview was conducted with the **15 users** after they completed the interaction with the two types of interfaces. The process followed to collect and analyze the data from the interviews was: 1) the interviews were recorded; 2) a transcription of the recordings was generated; 3) an analysis of the content of each interview was done using a codification technique. It is worth to mention that all conversations took place in the participants' native language (Spanish). Therefore, the content has been translated trying not to lose the significance of the words and expressions used. The questions asked during the interviews can be seen in Appendix A. The interviews allowed to gather the information needed to answer *RQ3.*

4. Results

This section answers the three research questions (RQs) of this work based on the data collected from: 1) SUS (RQ1); 2) Logs (RQ2); and 3) Interviews (RQ3).

4.1. Can a Question-driven Conversational Agent Fulfil Learners' Expectations in Terms of Usability? (RQ1)

The first step to follow while calculating the scores for SUS is to figure out the contribution of each question, whose range has to be between 1 (strongly disagree) and 5 (strongly agree). The *SUS Score* is calculated as *(X+Y)*2.5*, where

- *X* = Sum of the points for all odd-numbered questions – 5
- *Y* = 25 – Sum of the points for all even-numbered questions

The final score, *SUS Score*, is a number between 0 and 100. Once the *SUS Score* was computed for the 39 cases of the sample, the final scores ranged from 35 to 97.5 points out of 100 as it can be seen in the boxplot in Figure 1. Half the users scored JavaPAL within the range from 65 to 85. The median of the all scores is 77.5. The boxplot shows that the sample is positively skewed, as it is in the high part of the graph. The mean of the scores is 74.71 and the standard deviation is 16.479.

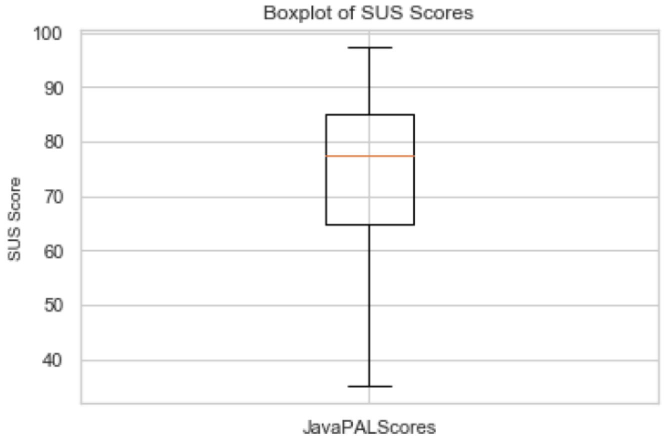


Fig. 1. SUS Global Scores boxplot for JavaPAL (median 77.5, mean 74.71).

SUS Score can also be converted into percentile ranks as indicated by Lewis and Sauro [32]. This percentile rank gives an idea of the usability of JavaPAL in relation to other products in a database. A score of 74.71 corresponds to a 70% of percentile rank, meaning that JavaPAL can be considered more usable than 70% of the products in the database [33] and less usable than 30 % of them. By definition, if a product has a percentile higher than 50% is considered to be above average.

Considering the scale by Bangor [29] (see Figure 2), it is possible to convert the SUS Score into grades, getting JavaPAL a C in this case. Moreover, in the Acceptability Ranges JavaPAL obtains “Acceptable”, while in the Adjective Ratings JavaPAL obtains “Good”. Bangor concludes that these results should be used together to create clearer pictures of products related to their overall usability [29].

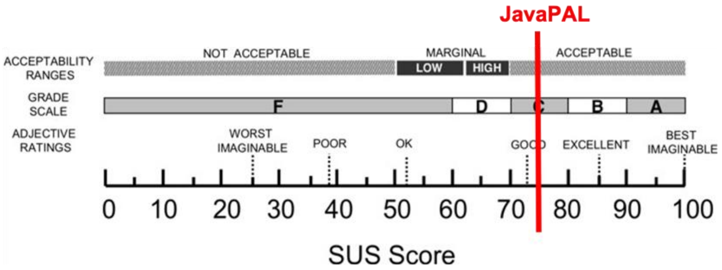


Fig. 2. SUS Bangor scale [29] and SUS Score for JavaPAL (mean value).

Another interpretation of SUS suggests dividing the items in two subscales: 1) “Learnability” (items 4 and 10); and 2) “Usability” (items 1-3 and 5-9) [33]. In order to compare the two subscales, the 2-item subscale should add the value of its items and multiply the results by 12.5, while the 8-item subscale should add the value of its items and multiply the result by 3.125. Figure 3 shows the results for “Learnability”, while Figure 4 shows the results for “Usability”. In the case of SUS Learnability Score the mean value is 83 and the median 87.5 (there is an outlier in 0); the results are positively skewed. In the case of SUS Usability Score the mean value is 72.51 and the median 75; Q1 and Q3 are 57.81 and 87.5, respectively, indicating that the results are also positively skewed.

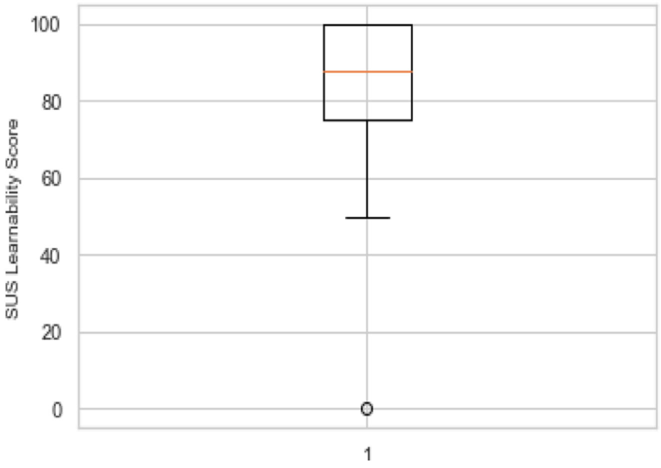


Fig. 3. SUS Learnability Score boxplot for JavaPAL (median 87.5, mean 83).

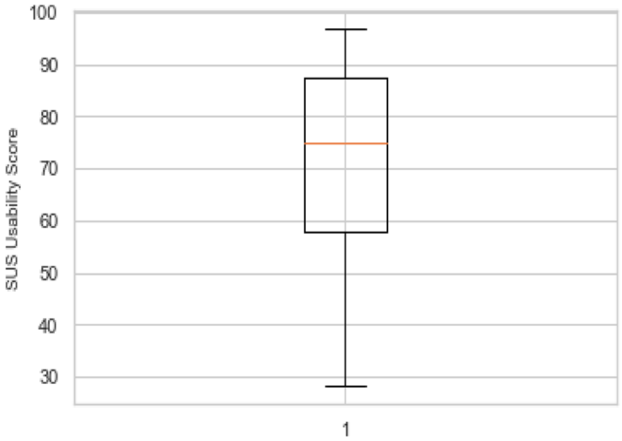


Fig. 4. SUS Usability Score boxplot for JavaPAL (median 75, mean 72.51).

All in all, regarding RQ1 it is possible to state that a conversational agent such as JavaPAL can provide a good usability to learners, including from the perspectives of both learnability and usability.

4.2. Does the use of a conversational agent affect learners’ performance when answering questions in MOOCs? (RQ2)

Logs were used to analyze the number of correct and incorrect answers provided by the participants (N=15) in the quasi-experiment. The number of correct and incorrect answers is the indicator used to measure learners' performance. Group 1 used the conversational agent first and then the MOOC web interface. Group 2 used the MOOC web interface first and then the conversational agent. Table 1 summarizes the results obtained

Table 1. Correct answers (mean and standard deviation) in Groups 1 and 2.

	Conversational agent	MOOC web interface
Group 1		
Mean correct answers	82.54%	74.6%
Standard deviation	15.83	8.32
Group 2		
Mean correct answers	89.76%	74.31%
Standard deviation	9.03	13.6

Group 1 members obtained a mean value of 74.6% of correct answers (SD=8.32) using the MOOC web interface, while Group 2 members obtained a mean value of 74.31% of correct answers (SD=13.6). In the case of the conversational agent, Group 1 obtained 82.54% of correct answers (SD=15.83), while Group 2 members obtained a mean value of 89.76% of correct answers (SD=9.03). In both cases the learners using the conversational agent obtained a higher percentage of correct answers on average.

A Mann-Whitney test of the difference between Group 1 and Group 2 was also applied. The confidence interval (-22.22, 5.5) was obtained with 95% confidence. This means that, in the worst-case scenario, and with a 95% of confidence, the use of conversational agent would decrease the learners’ performance by 5.5% (in terms of correct answers) when comparing with learner’s performance using the MOOC web interface.

All in all, regarding RQ2, it is possible to state that the use of the conversational agent like JavaPAL does not mean a worse learner’s performance, measured through the number of correct answers. Nevertheless, it is important to be cautious about these result as in some cases the questions presented to learners in the conversational agent had to be simplified (question and/or its answers) in order to make the question/answers easier to remember.

4.3. What are the differences learners and teachers find between the use of a MOOC web interface and a conversational agent interface when reviewing concepts and answering questions? (RQ3)

Interviews with 15 users of JavaPAL served to collect qualitative data with the aim to gain insights on the differences between the traditional MOOC web interface and the conversational agent interface when reviewing concepts and answering questions related to the MOOC.

One of the main advantages of JavaPAL highlighted by interviewees is the *immediate feedback*. This aspect was pointed out by *8 interviewees*. For example, Users 5 and 12 indicated that immediate feedback makes the conversational agent “more engaging”. User 6 added that thanks to the immediate feedback students could learn more because they would be more aware of their mistakes. Nevertheless, immediate feedback is a feature that can also be obtained through the MOOC web interface if the questions are configured properly. Nonetheless, it is easier (and faster) to ask the conversational agent to define a specific concept (in the review mode) than to search for the concept definition using the MOOC web interface.

Another positive aspect of JavaPAL highlighted by interviewees is the possibility of interacting with the conversational agent directly from the *mobile phone*. This aspect was pointed out by *6 interviewees*. For example, User 2 indicated that answering questions using the “mobile phone is faster than the mouse” in the web interface as it is the case when comparing voice-based interaction with text-based (or click-based) interaction. This same fact was also supported by three additional users. Users 1 and 5 also pointed out the benefits of using the conversational agent in the mobile phone “while travelling or commuting”. In addition, Users 3 and 7 also argued that the use of the conversational agent in the “mobile phone improves accessibility”. In contrast, User 1, for example, believed that the “web interface allows you to have a more general vision of the questions” you must answer unlike in the case of the conversational agent, and that the web interface allows for “more complex questions”. User 4 also indicated that, in general, people are “more used to using the web interface”, and that the web interface is “easier to use than the conversational agent”, which requires a certain learning curve. This idea was also reinforced by Users 1, 8 and 13. In contrast, User 3 claimed the opposite saying that more people interact through the mobile phone than through web interfaces.

The limitations of the conversational agent according to the interviewees are diverse. For example, User 4 believed that the conversational agent “cannot substitute the use of MOOCs” and it has to be seen as “a complementary tool”. User 13, for example, stated that it is “more complicated to type with the mobile phone than using the web interface”. However, four users believed that were no disadvantages in the case of the conversational agent. When it comes to the web interface, Users 7 and 10 identified as a drawback that there is no conversational interface. Users 8 and 9 both agreed on the fact that the web interface is “more monotonous” than the conversational agent and, in general, they all believe that the web interface is “more rigid” than the conversational agent.

Regarding preferences, User 8 indicated that he would always use the conversational agent. Users 1, 2 and 5 claimed that they would use the conversational agent while traveling instead of the web interface. The remaining users believed that the conversational agent would be preferable in case there was no access to the web interface, or when the content of the MOOC is highly theoretical. In contrast, Users 3 and 8 expressed a preference for the web interface. User 12 indicated a preference for the web interface only when having access to a computer, while User 15 indicated this preference for the web interface when accessing the main material to study.

Finally, in terms of learning, 6 interviewees believed that they would learn more with the conversational agent than with the web interface. For example, Users 8 and 11 believed that the conversational agent is “more engaging” so it would be easier to learn with it. However, 4 interviewees believed that they would obtain the same outcome when using the conversational agent and the web interface. On the contrary, 3 interviewees believed that the MOOC web interface would be better to learn because it has the videos and the questions interwoven and this is much better for those students with zero knowledge about Java programming.

Comments from interviewees were quite polarized regarding the preference between the use of a conversational agent like JavaPAL and the MOOC web interface. Regarding RQ3, and after assessing the advantages and disadvantages mentioned by the interviewees, it is possible to conclude that a conversational agent like JavaPAL can be a good complement to the MOOC, especially for some types of learners, mainly those who are more accustomed to the use of mobile devices or for which the learning curve to use a conversational agent is not very high.

5. Conclusions

An important limitation in MOOCs is the lack of support to learners. This limitation is particularly critical in the case of MOOCs on engineering and computer sciences due to the intrinsic difficulty of the contents. Conversational agents may alleviate this problem of lack of support to learners, becoming learners’ study partners. JavaPAL is a pioneering work on the use of voice-based conversational agents to support MOOC learners offering a quiz mode (JavaPAL asks the learner questions selected from the MOOC) and a review mode (JavaPAL provides definitions of the key concepts addressed in the MOOC as requested by the learner). This article has shed some light on the use of conversational agents through the example of JavaPAL, concluding that: 1) a conversational agent such as JavaPAL can provide a good usability to learners; 2) a conversational agent such as JavaPAL does not mean a worse learner’s performance in terms of answering correctly questions from the MOOC; and 3) a conversational agent like JavaPAL can be a good complement to the MOOC for learners who are more used to using mobile devices.

Although the results obtained are encouraging, this research work is not without limitations, which should be addressed as future research. First, JavaPAL has been designed to support learners enrolled in a specific MOOC. More research should be done adapting this conversational agent to the contents (key concepts and questions) extracted from other MOOCs (not necessarily in the areas of engineering or computer sciences). Second, the number of users from which data was collected is 39 (for RQ1) and 15 (for RQ2 and RQ3). More research should be done with a higher sample of JavaPAL users, and particularly with a sample that contains learners who are indeed taking the MOOC for which JavaPAL provides support. After the prototyping phase the conversational agent is now ready to be offered to a large number of learners taking the MOOC. Third, the comparison between the conversational agent interface and the MOOC web interface was designed to be fair, although some existing questions from the MOOC had to be adapted (in the case of RQ2) to be used in JavaPAL (as show in Appendix A). It would be interesting to do the opposite and design questions in a MOOC directly to be used in a conversational agent and then transfer these same questions to the MOOC web interface.

Acknowledgment. This work was supported in part by the FEDER/Ministerio de Ciencia, Innovación y Universidades-Agencia Estatal de Investigación, through the Smartlet and H2O Learning projects under Grant TIN2017-85179-C3-1-R and PID2020-112584RB-C31, and in part by the Madrid Regional Government through the e-Madrid-CM Project under Grant S2018/TCS-4307 and under the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M21), and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation), a project which is co-funded by the European Structural Funds (FSE and FEDER). Partial support has also been received from the European Commission through Erasmus+ Capacity Building in the Field of Higher Education projects, more specifically through projects LALA, InnovaT, and PROF-XXI (586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP), (598758-EPP-1-2018-1-AT-EPPKA2-CBHE-JP), (609767-EPP-1-2019-1-ES-EPPKA2-CBHE-JP). This publication reflects the views only of the authors and funders cannot be held responsible for any use which may be made of the information contained therein.

References

1. Shah, D., By the Numbers: MOOCs in 2019. (2019) [Online]. Available: https://www.classcentral.com/report/mooc-stats-2019 (current July 2020)
2. Cook, S., Bingham, P., Reid, S., Wang, X.: Going massive: Learner engagement in a MOOC environment. In Proceedings of the THETA 2015 - Create, Connect, Consume-Innovating today for tomorrow. Gold Coast, Queensland, Australia. (2015)
3. Alario-Hoyos, C., Pérez-Sanagustín, M., Delgado Kloos, C., Muñoz-Organero, M., Rodríguez-de-las-Heras, A.: Analysing the impact of built-in and external social tools in a MOOC on educational technologies. In Proceedings of the 8th European Conference on Technology Enhanced Learning (EC-TEL). Springer, Berlin, Heidelberg, Paphos, Cyprus, 5-18. (2013)
4. Yukselturk, E., Ozekes, S., Türel, Y. K.: Predicting dropout student: an application of data mining methods in an online education program. European Journal of Open, Distance and e-learning, Vol. 17, No. 1, 118-133. (2014)
5. Chyung, S. Y.: Systematic and systemic approaches to reducing attrition rates in online higher education. American Journal of Distance Education, Vol. 15, No. 3, 36-49. (2001)
6. Vitiello, M., Walk, S., Hernández, R., Helic, D., Gütl, C.: Classifying students to improve MOOC dropout rates. In Proceedings of the 2016 European MOOCs Stakeholders Summit (EMOOCs) Research Track. Graz, Austria, 501-508. (2016)
7. Vitiello, M., Walk, S., Helic, D., Chang, V., Guetl, C.: User Behavioral Patterns and Early Dropouts Detection: Improved Users Profiling through Analysis of Successive Offering of MOOC. Journal of Universal Computer Science, Vol. 24, No. 8, 1131-1150. (2018)
8. Dhorne, L., Deflandre, J. P., Bernaert, O., Bianchi, S., Thirouard, M.: Mentoring learners in MOOCs: A new way to improve completion rates. In Proceedings of the 2017 European MOOCs Stakeholders Summit (EMOOCs). Springer Cham, Madrid, Spain, 29-37. (2017)
9. Aguirre, C. C., Delgado-Kloos, C., Alario-Hoyos, C., Muñoz-Merino, P. J.: Supporting a MOOC through a Conversational Agent. Design of a First Prototype. In Proceedings of the 2018 International Symposium on Computers in Education (SIIE). IEEE, Jerez de la Frontera, Spain, 1-6. (2018)
10. Flecha, R.: Sharing Words: Theory and Practice of Dialogic Learning. Rowman & Littlefield, Lanham (2000)
11. Caballé, S., Conesa, J.: Conversational Agents in Support for Collaborative Learning in MOOCs: An Analytical Review. In International Conference on Intelligent Networking and Collaborative Systems (pp. 384-394). Springer, Cham. (2018)

12. Zadrozny, W., Budzikowska, M., Chai, J., Kambhatla, N., Levesque, S., Nicolov, N.: Natural language dialogue for personalized interaction. Communications of the ACM, Vol. 43, No. 8, 116-120. (2000)
13. Deepai.org. [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/conversational-agent (current July 2020)
14. Song, D., Oh, E. Y., Rice, M.: Interacting with a conversational agent system for educational purposes in online courses. In Proceedings of the 2017 10th International Conference on Human System Interactions (HSI). IEEE, Ulsan, South Korea, 78-82. (2017).
15. Chatbots magazine. [Online]. Available: https://chatbotsmagazine.com/chatbot-report-2019-global-trends-and-analysis-a487afec05b (current July 2020)
16. Weizenbaum, J.: ELIZA - A computer program for the study of natural language communication between man and machine. Communications of the ACM, Vol. 9, No. 1, 36-45. (1966)
17. AbuShawar, B., Atwell, E.: ALICE chatbot: Trials and outputs. Computación y Sistemas, Vol. 19, No. 4, 625-632. (2015)
18. Jia, J.: The study of the application of a keywords-based chatbot system on the teaching of foreign languages. (2003). [Online]. Available: https://arxiv.org/pdf/cs/0310018.pdf (current July 2020)
19. Edwin.ai. [Online]. Available: https://edwin.ai (current July 2020)
20. Knill, O., Carlsson, J., Chi, A., Lezama, M.: An artificial intelligence experiment in college math education. (2004). [Online]. Available: http://people.math.harvard.edu/~knill/preprints/sofia.pdf (current July 2020)
21. Niranjan, M., Saipreethy, M. S., Kumar, T. G.: An intelligent question answering conversational agent using Naïve Bayesian classifier. In Proceedings of the 2012 IEEE International Conference on Technology Enhanced Education (ICTEE). IEEE, Kerala, India, 1-5. (2012)
22. Kerry, A., Ellis, R., Bull, S.: Conversational agents in E-Learning. In Proceedings of the International Conference on Innovative Techniques and Applications of Artificial Intelligence. Springer, Cambridge, UK, 169-182. (2008)
23. Shaw, E., Johnson, W. L., Ganeshan, R.: Pedagogical agents on the web. In Proceedings of the third annual conference on Autonomous Agents (AGENTS'99), ACM, Seattle Washington USA, 283-290. (1999)
24. De Pietro, O., & Frontera, G.: Tutorbot: An Application AIML-Based for Web-Learning. Advanced Technology for Learning, Vol. 2, No. 1, 29-34. (2005)
25. Graesser, A. C., Person, N., Harter, D., and the Tutoring Research Group: Teaching tactics and Dialog in AutoTutor. International Journal of Artificial Intelligence in Education, Vol. 11, 1020-1029. (2000)
26. Demetriadis, S., Karakostas, A., Tsiatsos, T., Caballé, S., Dimitriadis, Y., Weinberger, A., ... Hodges, M.: Towards integrating conversational agents and learning analytics in MOOCs. In Proceedings of the International Conference on Emerging Internetworking, Data & Web Technologies. Springer, Cham, Tirana, Albania, 1061-1072. (2018)
27. Tomar, G. S., Sankaranarayanan, S., Rosé, C. P.: Intelligent conversational agents as facilitators and coordinators for group work in distributed learning environments (MOOCs). In Proceedings of the 2016 AAAI Spring Symposium Series, AAAI Press. (2016).
28. Creswell, J. W.: Educational research: Planning, conducting, and evaluating quantitative. Upper Saddle River, NJ: Prentice Hall. (2002)
29. Bangor, A., Kortum, P., Miller, J.: Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of usability studies, Vol. 4, No. 3, 114-123. (2009)
30. Delgado Kloos, C. Alario-Hoyos, C., Muñoz-Merino, P. J., Catalán-Aguirre, C. & González-Castro, N.: Principles for the Design of an Educational Voice Assistant for Learning Java. In Proceedings of the International Conference on Sustainable ICT, Education, and Learning. Springer, Zanzibar, Tanzania, 99-106. (2019)

31. Landauer, T. K.: Behavioral research methods in human-computer interaction. In Handbook of human-computer interaction. North-Holland, 203-227. (1997)
32. Lewis, J. R., & Sauro, J.: The factor structure of the system usability scale. In International conference on human centered design. Springer, Berlin, Heidelberg, San Diego, California, USA, 94-103. (2009)
33. Sauro, J.: A practical guide to the system usability scale: Background, benchmarks & best practices. Measuring Usability LLC. (2011)

Appendix A

Questions asked during the interviews

1. What are the differences between using the conversational agent and the traditional web interface to answer multiple choice questions?
2. What are the advantages of using a conversational agent?
3. What are the disadvantages of using the conversational agent?
4. What are the advantages of using the web interface?
5. What are the disadvantages of using the web interface?
6. In which situations would you use the conversational agent instead of using the web interface?
7. In which situation would you use the web interface instead of using the conversational agent?
8. Can you compare the conversational agent and the web interface in terms of learnability?
9. Can you compare the conversational agent and the web interface in terms of usability?
10. Can you compare the conversational agent and the web interface in terms of utility?

Example questions from the MOOC and transformed for the conversational agent

Questions from MOOCs	**Questions from JavaPAL**
Users have to write the result in a box given a value of “n”. System.out.println(n%6)	What is the result of the operation 7 percentage 2?
Users have to write the result of the operation, taking into account the precedence of the operators.	What operation is computed first? Multiplication / Addition / Division
Select True or False: An array can be extended after it has been initialized.	Can an array be extended after it has been created? True/False
The term “application” is similar to… Program Algorithm	The term application in Java is similar to: Program Algorithm
It is possible to run a program multiple times simultaneously? False True	Is it possible to run a program several times simultaneously? True False
The processing unit (select the correct answer out of 4 possible answers): - Is the module that runs the programs	The processing unit is the module that executes the program: - True / False

Cristina Catalán Aguirre is currently Product Line Maintenance Manager in Cloud Core Data-Storage Manager design organization at Ericsson. She received a double master's degree in Telecommunication Engineering and Telematics Engineering from Universidad Carlos III de Madrid in 2019 and a bachelor's degree on Telecommunication Technologies Engineering from the Public University of Navarra in 2017. She worked as a research assistant at the Telematics Engineering Department of Universidad Carlos III de Madrid from 2017 to 2019.

Nuria González-Castro is currently Cloud Engineer in Openbank. She received a double master's degree in Telecommunication Engineering and Telematics Engineering from Universidad Carlos III de Madrid in 2020, and a bachelor's degree on Telecommunication Technologies Engineering from this same university in 2018. She worked as a research assistant at the Telematics Engineering Department of Universidad Carlos III de Madrid from 2017 to 2020.

Carlos Delgado Kloos received the PhD degree in Computer Science from the Technische Universitat Munchen and in Telecommunications Engineering from the Polytechnical University, Madrid. He is full professor of Telematics Engineering at the Universidad Carlos III de Madrid, where he is the director of the GAST research group, director of the UNESCO Chair on "Scalable Digital Education for All", and Vice President for Strategy and Digital Education. He is also the Coordinator of the eMadrid research network on Educational Technology in the Region of Madrid.

Carlos Alario-Hoyos is Visiting Associate Professorin the Department of Telematics Engineering at the Universidad Carlos III de Madrid. He received M.S. and PhD degrees in Information and Communication Technologies from the Universidad of Valladolid, Spain, in 2007 and 2012, respectively. His skills and experience include research and development in MOOCs, social networks, collaborative learning, or evaluation of learning experiences.

Pedro J. Muñoz-Merino is Associate Professor at the Universidad Carlos III de Madrid, where he got his PhD in Telematics Engineering in 2009. Pedro has received several awards for his research. He is author of more than 100 scientific publications and has participated in more than 30 research projects. His skills and experience include research and development in learning analytics, educational data mining, evaluation of learning experiences, and gamification, among others.

Received: July 31, 2020; Accepted: February 28, 2021.

Computer Science and Information Systems 18(4):1287–1309 https://doi.org/10.2298/CSIS201201035N

Assessing Learning Styles through Eye Tracking for E-Learning Applications

Nahumi Nugrahaningsih[1], Marco Porta[2], and Aleksandra Klašnja-Milićević[3]

[1] University of Palangkaraya, Department of Informatics,
Kampus Unpar Tunjung Nyaho, Jl. Yos Sudarso , Palangkaraya 73112, Indonesia
nahumi@it.upr.ac.id
[2] University of Pavia, Department of Electrical, Computer and Biomedical Engineering,
Via A. Ferrata 5, 27100, Pavia, Italy
marco.porta@unipv.it
[3] University of Novi Sad, Department of Mathematics and Informatics,
Trg Dositeja Obradovica 3 – 21 000 Novi Sad, Serbia
akm@dmi.uns.ac.rs

Abstract. Adapting the presentation of learning material to the specific student's characteristics is useful to improve the overall learning experience and *learning styles* can play an important role to this purpose. In this paper, we investigate the possibility to distinguish between Visual and Verbal learning styles from gaze data. In an experiment involving first year students of an engineering faculty, content regarding the basics of programming was presented in both text and graphic form, and participants' gaze data was recorded by means of an eye tracker. Three metrics were selected to characterize the user's gaze behavior, namely, percentage of fixation duration, percentage of fixations, and average fixation duration. Percentages were calculated on ten intervals into which each participant's interaction time was subdivided, and this allowed us to perform time-based assessments. The obtained results showed a significant relation between gaze data and Visual/Verbal learning styles for an information arrangement where the same concept is presented in graphical format on the left and in text format on the right. We think that this study can provide a useful contribution to learning styles research carried out exploiting eye tracking technology, as it is characterized by unique traits that cannot be found in similar investigations.

Keywords: e-learning, learning models, learning styles, eye tracking, gaze behavior.

1. Introduction

The recent years, and especially the recent months, have seen a significant increase of e-learning solutions. However, in most cases the teaching material is the same for all students, without any distinction based on their specific "needs". *Learning styles* are a way to potentially identify how people learn best. Assessing learning styles to present the right material in proper ways to the user/learner can be of paramount importance in e-learning [1], [2].

It is a fact that there are different ways to learn, and different students will favor the learning modalities that are more suitable for them. Several investigations have highlighted the existence of bipolar learning styles, depending on whether, for example, a person prefers to learn by seeing or hearing, reflecting, or acting, reasoning in a logical or intuitive manner, visualizing or building mathematical models [3][31]. In general, researchers agree on the fact that learning materials should reflect students' learning styles. The huge amount of teaching resources currently offered in electronic form should therefore be adapted to the specific skills of the individual learner, in order to maximize the learning experience and improve learner achievements [4][32].

The most common way to assess learning styles is by means of questionnaires, through which students are asked to answer some questions aimed at discovering their preferred ways of learning. However, this kind of explicit assessment has some drawbacks ─ for instance, it may be considered long and boring, which causes careless responding, and the provided answers may not be sufficiently reliable. Thus, is it possible to automatically evaluate a person's learning style from the way he or she looks at the learning material? The aim of the paper is to identify the possibilities of *Eye tracking* technology to provide this kind of information, making learning style assessment a seamless procedure integrated into e-learning platforms ─ for example, through the analysis of the user's gaze behavior in the very initial stages of an e-learning course.

Incorporating eye tracking into adaptive e-learning systems by using data about pupil and gaze to indicate attentional focus and cognitive load levels can be useful in a process of adaptation to the requirements and needs of the learner. Personalization of an e-learning program based on the learner's cognitive load levels and learning styles calculated from eye-tracking data will impart the advantage of having a personal tutoring system into a wideband environment, with successful training by increasing information transfer and maintenance.

This is now realistic, as recent technological advances have enabled the development of affordable, robust, and mainstream eye-tracking solutions. Eye tracking is the process through which devices called *eye trackers* can detect the user's gaze direction [5]. In other words, an eye tracker identifies where a person is looking at (typically on a screen) and records the related gaze coordinates. Eye movements are characterized by very fast *saccades*, generally lasting less than 100 ms, interspersed with relatively steady periods of *fixations*, normally lasting between 100 and 600 ms. The main purpose of eye movements is to reallocate the gaze on the specific target, so that it can be clearly sensed on the fovea, the most sensitive area of the retina.

Eye trackers are becoming increasingly widespread nowadays, thanks to the availability of cheap devices. Current eye trackers are also non-invasive tools that do not constrain the user and allow to gather meaningful information in relatively simple ways. In the study presented in this paper, we exploited eye tracking technology to assess the user's Visual/Verbal learning styles from the way some slides presenting basic computer science topics (on the notion of *variable*, the concept of *algorithm*, and the *sequence*, *selection*, and *iteration* basic imperative programming constructs) were read/observed by 90 first-year engineering students.

Our study has two research objectives, one primary and one secondary. The primary research objective is: *Is it possible to distinguish Visual and Verbal learners from their gaze data recorded by an eye tracker?* Three metrics were selected to characterize the user's gaze behavior, namely, percentage of fixation duration, percentage of fixations, and average fixation duration. Our experiments were designed around this main purpose, through the presentation of basic computer science concepts in both textual and graphical form. However, as a secondary research objective, we also considered the possibility to *recognize the Active/Reflective, Sensible/Intuitive,* and *Sequential/Global* bipolar styles from learners' gaze behavior.

Gaze data were coupled with the outcomes of the Index of Learning Styles (ILS) questionnaire [6], one of the most widespread methods to evaluate users' learning styles. Even if a clear connection between the Visual/Verbal learning style could be found only for a specific information layout, we believe that our investigation can provide a constructive contribution to the field of e-learning in general, and to the area of automatic learning style assessment specifically. Exploiting eye tracking in this field is of paramount importance because it can potentially enable "intelligent" e-learning systems in which learning styles are assessed in a seamless way.

The paper is structured as follows. Section 2 presents a short summary of works that have exploited eye tracking technology for learning style assessment. Section 3 explains the main research questions at the basis of our study. Section 4 describes the methodology used for our investigation. Section 5 illustrates the performed analysis and the obtained results, which are then discussed in Section 6. Lastly, Section 7 outlines the conclusion and future work on the presented topic.

2. Background

This section provides an overview of eye tracking studies aimed at detecting learning and cognitive styles. A summary of the collected works is shown in Table 1. As can be seen, most investigations are focused on the Visual/Verbal learning styles. The order of the presented works is chronological.

Hughes et. al. [7] conducted an eye tracking study with 12 participants to investigate the difference between Verbalizer/Visualizer learners. The learning style was measured using the Verbalizer Visualizer Questionnaire (VVQ) by Kirby et al. [8]. Gaze data was recorded with an ASL 504 eye tracker. Stimuli were organized into ten "screens", each containing 20-25 video segments. The task for participants was to find a video that matched with a given topic. To avoid learning effects, the positions of text and visual components in the slides were alternately on the left and on the right. Since the VVQ results showed that no participants were in the verbalize group, the comparison was made only on visualizer and balanced learners. The eye features used were the average duration of fixations on slides, average fixation count, and average fixation duration. The statistical analysis revealed that there was a significant difference between the two groups. In particular, it indicated that balanced learners spent more time in the text area than visualizer learners. Despite variations in the layout, the statistical analysis revealed that participants' first fixations tended to be on the left side of the slide, regardless of the specific content in that area.

Tsianos at al. [9] tried to distinguish subjects into the wholist/analyst and verbal/imagery groups according to the Riding and Cheema's Cognitive Style Analysis (CSA) [10]. Twenty-one participants were involved in the experiment. The stimuli were web pages containing basic programming theories. The employed eye features were the ratio of fixation duration (i.e., the ratio between the times spent within image and text areas), the number of fixations on the page menu, and the experiment duration. The results of the statistical analysis on fixation ratios showed that imagers focused more on images, verbalizers more on text, and intermediates on both kinds of stimuli. The analysis on the number of fixations on the menu indicated that there was no difference among groups. Regarding session duration, imagers and intermediates devoted about the same time to read the whole content, while verbalizers spent considerably less time.

Al-Wabil et al. [11] observed the difference between visual and verbal learners according to Felder and Silverman’s learning style [3]. Eight participants were involved in the study. The stimuli were six slides containing an introduction to statistics. The eye features employed in the study were total fixation duration, mean of fixation duration, and number of fixations. Eye features were compared without performing a statistical significance test. Acquired data showed that visual learners looked more at the multimedia area, while verbal learners looked more at the text area. Regarding the mean of fixation duration, there was no difference between visual and verbal learners. The comparison of the number of fixations showed that there was no difference, as all participants tended to have more fixations on text.

Mehigan et al. [12] tried to distinguish between visual and verbal learners, who were evaluated with an online survey [6] implementing the Felder and Silverman’s learning style model. Several candidate participants were analyzed until a minimum of five visual learners and five verbal learners were found. The stimuli were composed of two slides: the first contained material about server-side programming, while the second contained a multiple-choice question to test the participants’ comprehension level. The first slide was divided into two equal areas containing an image and text. Fixation count, total fixation time on the text area, and total fixation time on the image area were analyzed. The data showed that visual learners made more fixations on the graphic slide area than verbal learners. However, no statistical significance assessment was conducted to confirm this result. Regarding the fixation time in image and text areas, a visual inspection on correlation distribution revealed that students with longer fixation duration on visual content tended to be more visual in their learning style, while learners with longer fixation duration on textual content tended to be verbal.

Cao and Nishihara [13] conducted an eye tracking experiment with 38 participants. The main focus of their research was to find the difference between Visual/Verbal and sequential/global learners according to the Felder and Silverman learning styles. The stimuli were 11 slides through which the participants could freely navigate. To distinguish visual and verbal participants, fixation time was employed as a feature. The obtained results showed that even though visual participants spent more time on picture areas than verbal participants, the difference was not significant. The same trend also appeared in the text area. To discriminate between sequential and global learners, the features employed were fixation duration, saccadic length, and saccadic orientation (i.e., the angle between the horizontal line and the saccade direction line). Results showed that global learners tended to have shorter fixation durations and moved the eyes faster and with larger degrees. However, differences were not significant in this case either.

Alyahya [14] examined the different performance between verbal/visual students when they were observing a historical map. The experiment involved 62 female students and learning styles of participants were self-assessed with the Verbal-Visual Learning

Style Rating instrument [15]. Stimuli of the eye tracking experiment were derived from Minard’s map, (i.e., a graph that combines a geographic map with a bar graph, time series, and text to present the journey of Napoleon’s march from France to Moscow in 1812). After presenting content slides to participants, their comprehension was tested with 20 textual multiple-choice questions (which the author called a *verbal test*) and a *visual test* to verify how much maps and cities were recalled. The results of an ANOVA analysis showed that there was a significant variability in the results of the visual test among the participants with different learning style. The posthoc analysis results indicated that the visual group performed better than the *mostly visual* group. However, there was no significant difference for the verbal test. From a visual inspection of the accumulative heatmap of each learning style group, it was found that both groups spent about the same time on the text area. However, the difference was especially evident in the map area where visual learners watched more than verbal learners.

Nisiforou and Laghos [16] investigated the relation between eye movements and cognitive style. A total of 54 students participated in the experiment. The cognitive style of participants was evaluated with a paper-based Hidden Figures Test [17]. Based on the results of the evaluation, participants were grouped into Field Dependent (FD), Field Independent (FI), and Field Neutral (FN) participants. In an eye tracking experiment, participants were asked to answer four questions that were inspired by the Hidden Figure Test. Participants had to click on the shapes that were hidden in pictures. The results obtained from a visual inspection of the gazeplot indicated that FD participants had more random gazeplot compared to FI participants, who had more oriented and organized gazeplots. Moreover, one-way ANOVA analyses carried out on fixation count and saccade count showed that there were significant differences among groups.

Goswami et al. [18] observed the gaze behavior of 13 participants with different learning styles when they tried to identify errors in a project document. Learning styles were assessed using the Felder and Silverman Index of Learning Style. Stimuli were 14 pages containing 14 errors, which were marked as Areas of Interest (AOI). It is to be stressed that the purpose of this work was not to recognize the user’s learning styles (like in our study), but, instead, to compare the user’s performance according to learning styles. An evaluation was conducted to recognize effective (those who found more faults) and efficient (those who found faults faster) participants. The considered eye features were total fixation time per page, duration per page, linear saccade per page, total fixation per AOI, and duration per AOI. The results of multiple regression analyses indicated that total fixation, total fixation per AOI, and duration per AOI were factors that significantly contributed to achieve a high effectiveness. High effectiveness was shown by participants with sensible and sequential styles. As for efficiency, there were no factors that were positively significant; however intuitive and global participants tended to have an eye behavior that influenced efficiency in a negative way. The same negative tendency on effectiveness was also found on participants with a combination of verbal and linear styles.

Koc-Januchta et al. [19] carried out an investigation to explore the differences between visualizers and verbalizers according to how they look at pictures and text during the learning process. Through questionnaires, students were categorized based on their visual or verbal cognitive styles. Two different topics were used. The results showed that visualizers spent more time on images than verbalizers, and verbalizers spent more time reading text. Also, verbalizers observed non-informative picture areas earlier than visualizers. A similar study was carried out by Höffler et al. [20] (from the same research group) to validate the Object-Spatial Imagery and Verbal Questionnaire (OSIVQ) – which assumes a three-dimensional cognitive style model discriminating between object imagery, spatial imagery, and verbal dimensions. They found substantially different correlations of the different cognitive style scales with gaze behavior and visual-spatial ability. Participants scoring high on the object scale and/or the spatial scale of OSIVQ relied more heavily on pictures than on texts (indicated by high positive correlations with a joint gaze behavior score), while participants scoring high on the verbal scale tended to rely on texts (indicated by a negative, non-significant correlation). Additionally, only participants scoring high on the spatial scale tended to additionally have a high visuo-spatial ability, as indicated by a significant positive correlation.

Raptis et al. [21] presented two studies based on a multifactorial model. In both, participants carried out visual tasks with different characteristics, and eye tracking analysis discovered significant differences among participants characterized by different cognitive styles. In particular, the authors considered the Field Dependence-Independence (FD-I) cognitive style theory: while field-dependent users tend to prefer holistic ways for processing visual information, field-independent users tend to favor more analytical information processing approaches. The study revealed that the first category of users followed a more disoriented approach when performing visual search tasks, while the second category adopted a more organized visual strategy. Such differences suggested classification experiments in which different classifiers were trained with eye tracking data to infer the category a user belongs to.

Alhasan et al. [22] conducted a preliminary eye tracking study to analyze the pattern of learner behavior in order to obtain their learning style as a personalization aspect in an e-learning system. The electroencephalography (EEG) Emotive Epoc device was used to disclose learners with more accurate data. A method was developed to determine whether the verbal and visual learning styles reflect actual preferences in an e-learning environment based on the Felder and Silverman Learning Style Model. “Emotions” were exploited to exclude the periods of time when the learner was not focusing on learning. The primary experiment designed to test the combination of eye tracking and EEG confirmed operability and efficiency of this approach for studying and analyzing learning styles.

The studies that mainly guided the methodological choices of our investigation were [7], [9], [11], [12], [13], [14], and [19]. All of them have, as a main purpose, the recognition of learning styles from users’ gaze behavior. They also include the Visual/Verbal styles, which are the primary research objective of this study. Moreover, our investigation introduces novel elements compared to these previous works, such as the fact that participants had no time limits. This choice allowed us to carry out an experiment closer to real learning scenarios, without sacrificing a time-dependent analysis. As will be illustrated in Section 5.2 (*Statistical Procedures*), such analysis was implemented through the subdivision of the single participants’ interaction times with the learning stimuli into intervals, which is also an original aspect of our work.

Table 1. Studies using eye tracking for learning style detection.

Studies	Number of Participants	Learning Style Instruments	Learning Styles	Eye Features
[28]	22	Felder and Soloman	Visual/Verbal	Gaze paths, fixation count, fixation duration and average time for each fixation
[29]	28	Felder and Silverman Learning Style Model (FSLSM)	Visual/Verbal	Fixation duration, fixation count and the average time on each fixation
[30]	7	Felder-Silverman Index of Learning Styles	Visual/Verbal	The time that the participants gazed at text-based or graphic-based learning objects
[22]	48	Felder and Silverman Learning Style Model [3]	Visual/Verbal	Fixation count, average fixation duration
[21]	36	Field Dependence-Independence theory	Field dependent/field independent	Fixation count, fixation duration, saccade length, combined metrics
[20]	32	Object-Spatial Imagery and Verbal Questionnaire (OSIVQ, [23])	Object visualizers, spatial visualizers, and verbalizers	Dwell time (sum of durations from all fixations and saccades that hit the AOI in seconds) and revisits (number of returns to the AOI after the first visit)
[19]	32	Santa Barbara Learning Style Questionnaire (SBCSQ, [15]), Individual Differences Questionnaire [24], Vividness of Visual Imagery Questionnaire (VVIQ, [25]), Verbalizer – Visualizer Questionnaire [26],	Visual/Verbal	First gaze time (duration from start of the trial to the first hit of the AOI), dwell time (sum of durations of all fixations and saccades that hit the AOI), and transitions (movements from one AOI to another)

		Object-Spatial Imagery and Verbal Questionnaire (OSIVQ, shortened version, [23])		
[18]	13	Felder and Silverman Index of Learning Style [3]	Active/reflective, Sensible/intuitive, Visual/Verbal, Sequence/global	Total fixation time per page, duration per page, linear saccade per page, total fixation per AOI, duration per AOI
[16]	54	Hidden Figures Test [17]	Field dependent / field neutral/field independent	Fixation count, saccade count, average fixation duration, average saccade duration
[14]	62	Verbal-Visual Learning Style Rating [15]	Visual/Verbal	Total fixation duration on the text and map
[13]	38	Felder and Silverman Index of Learning Style [3]	Visual/Verbal, Sequential/global	Fixation duration, saccadic length, and the saccadic orientation (i.e. the angle between the horizontal line and the saccade line)
[12] *	10	Felder-Solomon Index of Learning Styles [6]	Visual/Verbal	Fixation count, total fixation time on text area, total fixation time on image area
[11] *	8	Felder and Silverman Index of Learning Style [3]	Visual/Verbal	Total fixation duration, average fixation duration, and fixation count
[9]	21	Riding and Cheema’s Cognitive Style Analysis [10]	Wholist/analyst Verbal/imagery	Ratio of fixation duration (i.e. ratio between time spent in image and text areas), fixation count on the menu, and duration of the sessions
[7]	12	Kirby, Moore and Schofield's Verbalizer Visualizer Questionnaire [8]	Verbalizer/visualizer	Average duration of slide, average fixation count, average fixation duration

* Direct comparisons were performed without a statistical significance test

3. Research Questions

As already stated in the Introduction, the present study was based on a primary (RQ1) and on a secondary (RQ2) research questions. The reason for such a distinction is because our experiments were mainly designed to answer RQ1. Nevertheless, we wanted to verify whether, using the same data gathered for RQ1, it was also possible to answer RQ2.

The two research questions were:

RQ1. Is it possible to distinguish Visual and Verbal learners from the features of their gaze behavior (percentage of fixation duration, percentage of fixations, and average fixation duration) recorded by an eye tracker?

RQ2. Is it possible to recognize Active/Reflective, Sensible/Intuitive, and Sequential/Global learners from the features of their gaze behavior (percentage of fixation duration, percentage of fixations, and average fixation duration recorded by an eye tracker)?

4. Methodology

4.1. Participants

In total, 90 volunteer students participated in the experiment (57 males and 33 females, 18 years old on average). All of them were freshman Computer Engineering students of the Informatics Department of the University of Palangkaraya and had not attended any computer programming course yet. The recruitment occurred through announcements in bulletin boards in the department. All the participants, generally curious about eye tracking technology, were fully informed about the experiment procedures before starting them. No personal data were stored, as all the participants in the experiment were anonymously identified through numbers (only needed to match questionnaire data with eye tracking data). The participants did not get any academic credits for participating in the experiments, but they simply received their "gazeplots" (graphical representations indicating the visual scanpaths of their gaze) as "souvenirs”.

4.2. Materials

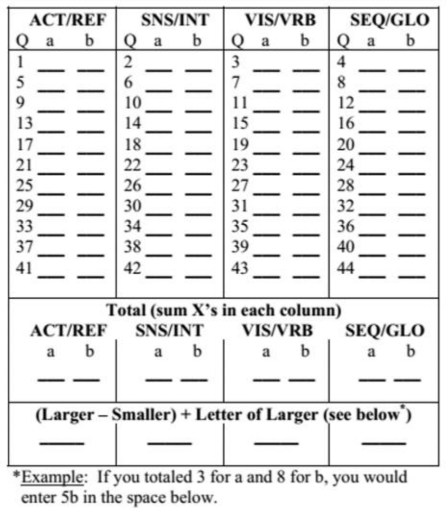
To record gaze data, we employed the low-cost Eye Tribe ET-1000 eye tracker [27], with 60 Hz data sampling rate. Stimuli were displayed on a 21.5" monitor.

4.3. Procedure

The experiments were subdivided into two phases, namely *Experimental Phase 1* and *Experimental Phase 2*.

Experimental Phase 1. To preliminarily investigate their learning styles through a “traditional” approach, the participants were initially asked to complete the Index of Learning Styles (ILS) questionnaire [6]. The ILS questionnaire is an instrument composed of 44 multiple-choice questions which aims to distinguish four bipolar styles, namely Active/Reflective (AR), Sensible/Intuitive (SI), Visual/Verbal (VV), and Sequential/Global (SG). There are two answers (*a* and *b*) for each question. In our study, the original questionnaire was translated into Indonesian.

The Index of Learning Styles of each participant was calculated using the scoring sheet shown in Figure 1.



ACT/REF			SNS/INT			VIS/VRB			SEQ/GLO		
Q	a	b	Q	a	b	Q	a	b	Q	a	b
1	___	___	2	___	___	3	___	___	4	___	___
5	___	___	6	___	___	7	___	___	8	___	___
9	___	___	10	___	___	11	___	___	12	___	___
13	___	___	14	___	___	15	___	___	16	___	___
17	___	___	18	___	___	19	___	___	20	___	___
21	___	___	22	___	___	23	___	___	24	___	___
25	___	___	26	___	___	27	___	___	28	___	___
29	___	___	30	___	___	31	___	___	32	___	___
33	___	___	34	___	___	35	___	___	36	___	___
37	___	___	38	___	___	39	___	___	40	___	___
41	___	___	42	___	___	43	___	___	44	___	___

Total (sum X’s in each column)

ACT/REF		SNS/INT		VIS/VRB		SEQ/GLO	
a	b	a	b	a	b	a	b
___	___	___	___	___	___	___	___

(Larger – Smaller) + Letter of Larger (see below*)

____	____	____	____

*Example: If you totaled 3 for a and 8 for b, you would enter 5b in the space below.

Fig. 1. ILS Scoring sheet

The result score is an odd number between 1 and 11, whose interpretation, according to Felder and Soloman, is as follows:

- If the score is 1 or 3: the respondent is fairly well balanced on the two dimensions of that scale.
- If the score is 5 or 7: the respondent has a moderate preference for one dimension of the scale and will learn more easily in a teaching environment which favours that dimension.
- If the score is 9 or 11, the respondent has a very strong preference for one dimension of the scale and may have real difficulties when learning in an environment which does not support that preference.

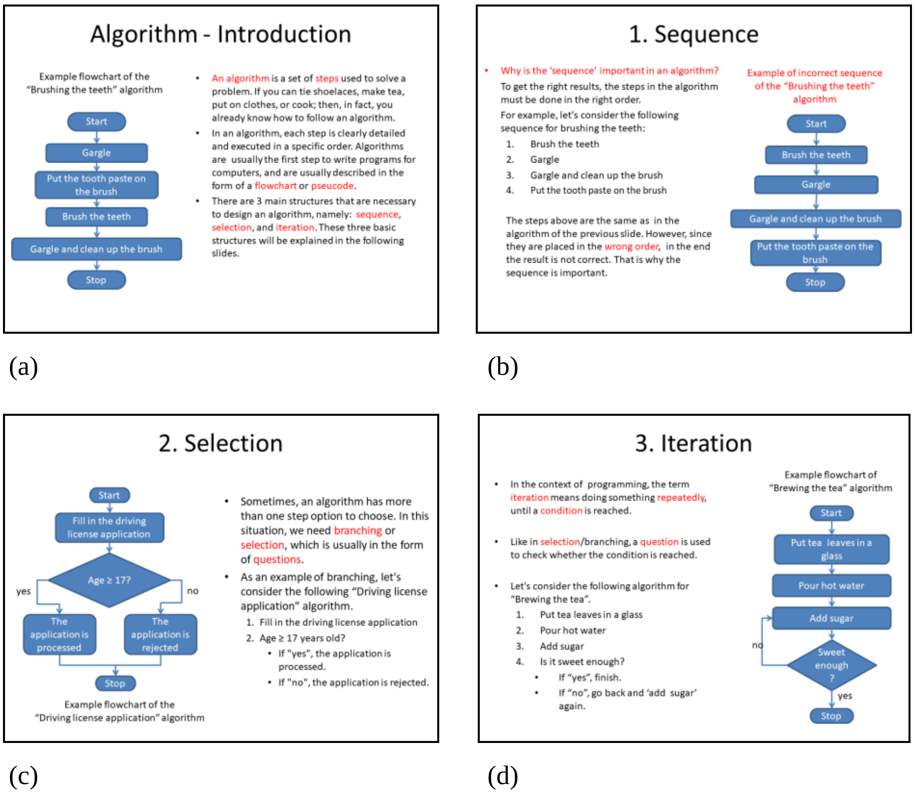
Experimental Phase 2. Subsequently, after three days from the Experimental Phase 1, the participants also attended an eye tracking experiment. The participants were not informed that this trial was connected with the questionnaire they had answered in Phase

1. A within-subjects experimental design was used, in which participants tried all the available conditions.

The eye tracking experiment was conducted in a quiet room, with artificial illumination from the ceiling. The participant in the test was seated at about 55 cm from the monitor. The task was to read and try to understand the topics presented in a group of slides. No time limit was set for each slide, so that the participants could learn at their own pace (a new slide was loaded by pressing the space bar).

In total, there were seven slides. The first one contained a description of the task; the second consisted of a graphical overview of the topics; the third explained the basic notion of *variable*; the fourth presented the concept of *algorithm*; and the fifth, sixth and seventh slides, respectively, covered the three basic imperative programming constructs, namely *sequence*, *selection*, and *iteration*.

In this study, we focused on slides from the 4th to the 7th in the above list, that, in the following, we will identify as slides *a*, *b*, *c*, and *d*. Figures 2a-2d show the translation of the original slides (written in Indonesian) into English.



Algorithm - Introduction

Example flowchart of the "Brushing the teeth" algorithm

Start
Gargle
Put the tooth paste on the brush
Brush the teeth
Gargle and clean up the brush
Stop

- An algorithm is a set of steps used to solve a problem. If you can tie shoelaces, make tea, put on clothes, or cook; then, in fact, you already know how to follow an algorithm.
- In an algorithm, each step is clearly detailed and executed in a specific order. Algorithms are usually the first step to write programs for computers, and are usually described in the form of a flowchart or pseucode.
- There are 3 main structures that are necessary to design an algorithm, namely: sequence, selection, and iteration. These three basic structures will be explained in the following slides.

(a)

1. Sequence

- Why is the 'sequence' important in an algorithm?

To get the right results, the steps in the algorithm must be done in the right order.

For example, let's consider the following sequence for brushing the teeth:

1. Brush the teeth
2. Gargle
3. Gargle and clean up the brush
4. Put the tooth paste on the brush

The steps above are the same as in the algorithm of the previous slide. However, since they are placed in the wrong order, in the end the result is not correct. That is why the sequence is important.

Example of incorrect sequence of the "Brushing the teeth" algorithm

Start
Brush the teeth
Gargle
Gargle and clean up the brush
Put the tooth paste on the brush
Stop

(b)

2. Selection

Start
Fill in the driving license application
Age ≥ 17?
yes
no
The application is processed
The application is rejected
Stop

Example flowchart of the "Driving license application" algorithm

- Sometimes, an algorithm has more than one step option to choose. In this situation, we need branching or selection, which is usually in the form of questions.
- As an example of branching, let's consider the following "Driving license application" algorithm.
  1. Fill in the driving license application
  2. Age ≥ 17 years old?
     - If "yes", the application is processed.
     - If "no", the application is rejected.

(c)

3. Iteration

- In the context of programming, the term iteration means doing something repeatedly, until a condition is reached.
- Like in selection/branching, a question is used to check whether the condition is reached.
- Let's consider the following algorithm for "Brewing the tea".
  1. Put tea leaves in a glass
  2. Pour hot water
  3. Add sugar
  4. Is it sweet enough?
     - If "yes", finish.
     - If "no", go back and 'add sugar' again.

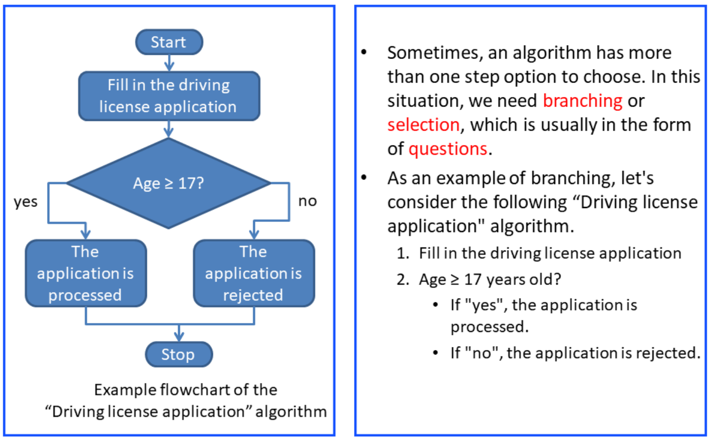
Example flowchart of "Brewing the tea" algorithm

(d)

Fig. 2. English version of the slides used as stimuli in the eye tracking experiment

5. Analysis of Eye Tracking Data and Results

In each slide, we defined two AOIs (Areas Of Interest): one for the text section and another for the picture region (Figure 3). As can be seen from Figure 2, text and pictures were alternately on the left and on the right within slides.


Start
Fill in the driving license application
Age ≥ 17?
yes
no
The application is processed
The application is rejected
Stop

Example flowchart of the "Driving license application" algorithm

- Sometimes, an algorithm has more than one step option to choose. In this situation, we need branching or selection, which is usually in the form of questions.
- As an example of branching, let's consider the following "Driving license application" algorithm.
  1. Fill in the driving license application
  2. Age ≥ 17 years old?
     - If "yes", the application is processed.
     - If "no", the application is rejected.

Fig. 3. Example of AOIs in a slide

The *independent* variables of the eye tracking study were the position of the picture and of the text areas on the slides (left-right or right-left).

The *controlled variables* were the textual and graphical contents displayed in the slides (arranged as shown in Figure 2).

The *dependent* variables, besides the questionnaire outcomes for Phase 1, in Phase 2 were the *percentage of fixation duration* (i.e., the percentage of fixation time on the AOI), the *percentage of fixations* (i.e., the percentage of fixations detected on the AOI), and the *average fixation duration*. Percentages were preferred to absolute values because the time spent on each slide by each participant was different.

For a temporal analysis of eye behavior, we subdivided the whole time spent by each participant on a slide into ten intervals. For each slide and each interval, we calculated the percentage of fixation duration (over the total time spent on the slide), the percentage of fixations (over the total number of fixations detected on the slide), and the average fixation duration up to that interval.

Unfortunately, six of the 90 participants did not fill in the questionnaire completely. Other four participants failed the eye tracking calibration procedure (consisting in fixating the center of a circle appearing in different positions of the screen). Moreover, 25 participants tried the test more than once, due to problems occurring in the data recording phase. Thus, in the end, we decided to consider only eye data from the surely reliable 55 participants.

5.1. Score Distributions

As shown by the histograms in Figure 4, the scores obtained from the Felder-Silverman questionnaire were not evenly distributed.

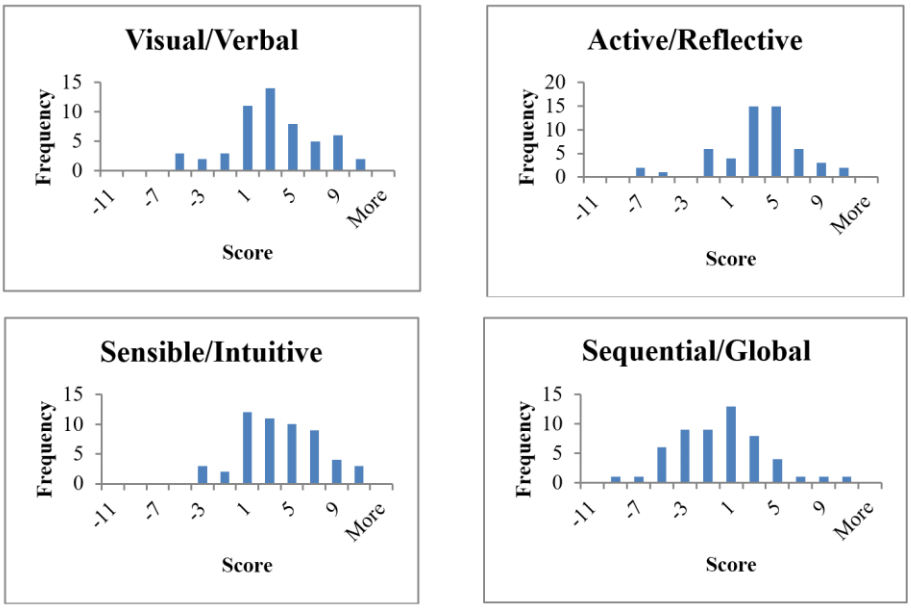


Fig. 4. Score histograms obtained from the answers to the Felder-Silverman questionnaire

For this reason, instead of classifying participants by the score threshold (as suggested by Felder and Soloman), we grouped them based on the median (MED) and median absolute deviation (MAD) values of the score. Specifically, we identified three groups:

- *Group 1*, with score < MED – MAD
- *Group 2*, with score > MED + MAD
- *Group 3*, with score in the range (MED – MAD) ÷ (MED + MAD)

Since learning styles are bipolar measurements, this classification can be interpreted as a learning style "tendency" of participants in the three groups. For example, for the Visual/Verbal case, Group 1 means "more verbal than visual", Group 2 "more visual than verbal", and Group 3 "between visual and verbal". Table 2 shows the number of participants in each group for the four kinds of learning styles.

Table 2. Studies using eye tracking for learning style detection

	Visual/Verbal **MED = 3, MAD = 2**	**Active/Reflective** **MED = 3, MAD = 2**	**Sensible/Intuitive** **MED = 3, MAD = 2**	**Sequential/Global** **MED = 1, MAD = 2**
Group 1	8	9	5	17
Group 2	13	11	16	7
Group 3	34	35	34	31

5.2. Statistical Procedure

For all the three selected metrics (percentage of fixation duration, percentage of fixations, and average fixation duration), we used the Shapiro-Wilk test to verify the normality of data distributions in the ten intervals. Since distributions were not normal in numerous cases, and various attempts to transform data using several functions were not successful, we carried out a non-parametric statistical analysis. We therefore considered medians instead of means.

The next step of the analysis had the purpose to verify whether the specific slide influenced the three metrics. This was done by means of the Friedman's test applied to each learning style group (Group 1, Group 2, and Group 3) separately. The obtained results indicated that differences among slides were significant in most cases, and therefore it was not possible to consider the four slides together. Thus, we investigated whether a common behavior could be found considering the couples of slides with the same structure, (i.e., the two slides with the picture on the left and the text on the right slides *a* and *c* in Figure 2), and the two slides with the opposite arrangement (i.e., slides *b* and *d*).

We considered the four kinds of learning styles – Visual/Verbal (VV), Active/Reflective (AR), Sensible/Intuitive (SI), and Sequential/Global (SG) – separately, and the three metrics for each of them. In both the text and picture areas, in each of the six (3 groups x 2 areas) cases of each learning style and metric, we counted the number of occurrences in which the influence of the slide was not significant, with a 5% significance level. This value is traditionally and universally used in statistics as the significance level for decisions.

Although it was not possible to find cases in which the values of the metrics were independent of the slide in all 10 intervals for all three learning style groups, we considered as acceptable, or *valid*, those cases where the effect of the *slide* factor was not significant in at least seven intervals out of ten. For the two slides with the picture on the left and the two slides with the picture on the right, respectively, Tables 3 and 4 show these valid occurrences for each learning style category and metric, indicating whether they are related to the picture area (P), the text area (T), or both (PT).

As can be seen from Table 3, when the picture is on the left, it is never possible to consider the average fixation duration (no pair of bipolar styles has at least seven non-significant differences between slides *a* and *c*). The percentage of fixations is potentially useful for the VV, AR, and SG categories only on the text area. Lastly, the percentage of fixation duration is exploitable on both the picture and text areas for all learning style categories except SG (for which only the text region can be analyzed).

Table 3. Cases with at least seven non-significant differences between slides a and c (picture on the left and text on the right), for each learning style category and metric (P = picture area, T = text area)

	VV	**AR**	**SI**	**SG**
%FixDur	PT	PT	PT	T
%Fix	T	T		T
AvgFixDur				

Table 4. Cases with at least seven non-significant differences between slides b and d (picture on the right and text on the left), for each learning style category and metric (P = picture area, T = text area)

	VV	**AR**	**SI**	**SG**
%FixDur	T			
%Fix	P	PT	PT	PT
AvgFixDur	P	P		PT

When the picture is on the right (Table 4), the percentage of fixation duration can be considered only for the VV category and on the text area. The percentage of fixations is potentially useful on both the picture and text regions for all learning styles, except for VV (for which only the picture area can be studied). In regard to the average fixation duration, it can be used on both the text and the picture areas for SG, and only on the picture area for VV and AR.

After identifying the valid cases for metrics, learning style categories, and slide regions, the last step was using the Kruskal-Wallis test to find possible connections (i.e., relationships) between the metrics' values and learning style groups (Group 1, Group 2, and Group 3). Slides *a* and *c* (picture on the left) and slides *b* and *d* (picture on the right) were considered distinctly.

5.3. Results

For each metric, learning style category, AOI, and interval, we searched for valid cases with significant relations (5% significance level) between metric value and learning style group. This happened in very few occurrences, as shown in Tables 5 and 6.

As can be seen, the only metric with significant relations in both slides was, for slides a and c (i.e., picture on the left and text on the right), the percentage of fixation duration in the text area, for the Visual/Verbal style and in intervals 9 and 10. Hence, according to our analysis, in slides having a picture on the left and a corresponding text description on the right, the percentage of fixation duration up to the last part of the interaction (intervals 9 and 10), can be exploited to distinguish the groups of Visual/Verbal learners.

Table 5. Picture on the left and text on the right

Slide a:	*Slide c:*
• Percentage of fixation duration, VV, text area: in interval 9 ($\chi^2(2) = 7.237$, $p = .027$) and in interval 10 ($\chi^2(2) = 8.306$, $p = .016$)	• Percentage of fixation duration, VV, text area: in interval 9 ($\chi^2(2) = 6.421$, $p = .04$) and in interval 10 ($\chi^2(2) = 8.092$, $p = .017$) • Percentage of fixation duration, SI, text area: in interval 9 ($\chi^2(2) = 6.709$, $p = .035$) and in interval 10 ($\chi^2(2) = 6.304$, $p = .043$)

Table 6. Picture on the right and text on the left

Slide b:	*Slide d*:
• Average fixation duration, AR, text area: in interval 1 ($\chi 2(2) = 7.449$, $p = .024$) • Average fixation duration, AR, picture area: in interval 2 ($\chi 2(2) = 7.217$, $p = .027$)	No significant instances

In particular, pairwise comparisons (carried out using the Dunn-Bonferroni test) allowed to determine that, for both slides and both intervals, the difference was significant for Group 1 and Group 2, with the percentage of fixation duration always higher for Group 1. This means that, considering at least the first 90% of the interaction time with the slide, the text area was observed more than the picture region by Verbal learners and less by Visual learners. This is also evident from Figure 5, which shows the evolution over time (medians calculated up to each interval) of the percentage of fixation duration for the VV learning style in the text area in slides *a* (left) and *c* (right). Figure 6 shows the box plots indicating the values of the medians of each group for each slide (*a* and *b*) and intervals 9 and 10.

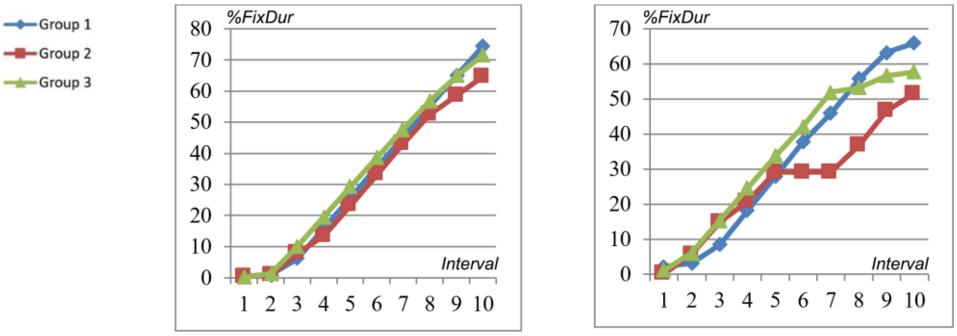


Fig. 5. Evolution of the percentage of fixation duration on the text area for the Visual/Verbal learning styles and the three learner groups in slide *a* (left) and in slide *c* (right)

In a boxplot, the bottom and top of the box indicate the 25th and 75th percentiles (i.e., the percentages of fixation duration corresponding to, respectively, the 25% and the 75% of the gathered data), while the inner band designates the 50th percentile (i.e., the medians); the ends of the whiskers represent the smallest and largest non-outlier values; circles denote outliers standing more than 1.5 box-lengths above or below the box; and stars indicate extreme values, standing more than three box-lengths above or below the box.

6. Discussion

Eye-tracking technology can be useful for implicitly classifying users based on their high-level cognitive processes (i.e., cognitive styles) in real-time while performing activities with varying characteristics (e.g., type complexity). In the study presented in this paper, we exploited eye tracking technology to assess the user's Visual/Verbal learning styles from the way some slides presenting basic computer science topics (on the notion of *variable*, the concept of *algorithm*, and the *sequence*, *selection*, and *iteration* basic imperative programming constructs).

Three metrics were selected to characterize the user's gaze behavior, namely, percentage of fixation duration, percentage of fixations, and average fixation duration. Our experiments were designed around this main purpose, through the presentation of basic computer science concepts in both textual and graphical form. However, as a secondary research objective, we also considered the possibility to *recognize the Active/Reflective, Sensible/Intuitive,* and *Sequential/Global* bipolar styles from learners' gaze behavior.

Gaze data were coupled with the outcomes of the Index of Learning Styles (ILS) questionnaire [6], one of the most widespread methods to evaluate users' learning styles. A connection between the Visual/Verbal learning styles was found for a specific information layout, which gives a constructive contribution to the field of e-learning in general, and to the area of automatic learning style assessment specifically. Exploiting eye tracking in this field is of paramount importance because it can enable "intelligent" e-learning systems in which learning styles are assessed in a seamless way.

According to our results, the answer to the primary research question of our study (i.e., "Is it possible to distinguish Visual and Verbal learners from the features of their gaze behavior – percentage of fixation duration, percentage of fixations, and average fixation duration – recorded by an eye tracker?") is partially positive:

- A relation between gaze behavior and learners' group (groups obtained from our modified interpretation of the Felder-Silverman questionnaire outcomes, as illustrated in sub-section 5.1, Score Distributions) could be found only for Group 1 (participants who were classified as more verbal than visual) and Group 2 (participants who were classified as more visual than verbal), but not for Group 3 (participants who were classified as being between visual and verbal), which was the largest.
- The relation between gaze behavior and Groups 1 and 2 could be found only for slides having the picture on the left and the text description on the right, not for the opposite case.

Specifically, the percentage of fixation duration on the text area, computed up to intervals 9 and 10 (i.e., up to the last part of the slide reading/observation process), gives clear information about the user's style group (Group 1 or Group 2). This indicates that, if most of the time (at least 90%) spent on the slide is evaluated, the Visual/Verbal learner can be successfully recognized.

As regards the secondary research question of our study, (i.e., "Is it possible to recognize AR, SI, and SG learners from the features of their gaze behavior – percentage of fixation duration, percentage of fixations, and average fixation duration – recorded by an eye tracker?"), the answer is negative: for no metric, significant relations with the three learners' groups (Group 1, Group 2, and Group 3) could be found. This, however, was partially expected, as our experiments were specifically designed with the first research question in mind. Indeed, also looking at the literature, eye tracking has been rarely used in studies aimed at recognizing styles other than Visual and Verbal.

Due to the peculiarity of the experiments, we have implemented in our study, a direct comparison with previous works is not possible. The novelty of our approach is due to three main factors:

- The subdivision of participants into three groups, based on the outcomes of the Felder-Silverman questionnaire, using MED and MAD to define score intervals;
- The absence of time limits for participants while reading or observing the content of the presented slides, to make the experiment more similar to real learning scenarios;
- A temporal analysis carried out by subdividing the time taken by each participant to read/observe the content of each slide into (ten) intervals.

A limitation of our study is the simple structure of slides, which may prevent our results to be generalized to more complex layouts. Moreover, the subjects of the slides (notion of variable, concept of algorithm, and the three basic imperative programming constructs) are very specific, and this may have influenced the results. Also, all the participants were about the same age (18) and engineering students.

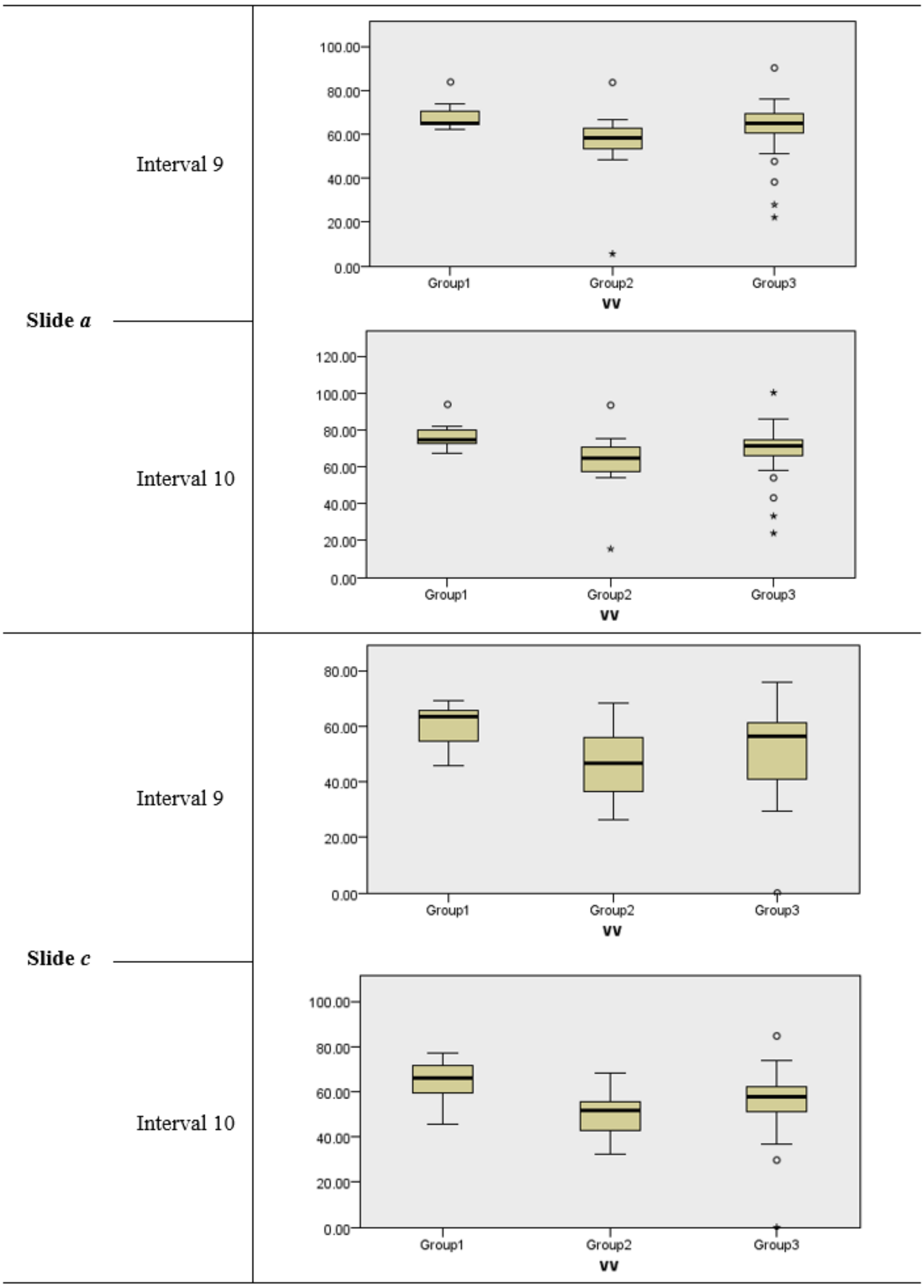


Fig. 6. Boxplots for each group, interval, and slide (Visual/Verbal learning styles)

7. Conclusions

In this paper, we have studied the possibility to recognize learning styles from the way users look at learning material, focusing in particular on the Visual/Verbal case. The content was basically structured into a two-column layout, with either an image on the left and text on the right or vice versa. The Index of Learning Styles questionnaire was exploited to preliminarily assess the styles of the participants in the experiments, to find possible connections between their gaze behavior and potential associated styles. The participants were grouped based on median and median absolute deviation of the scores obtained from the questionnaire. For a given bipolar learning style category, three groups were created which included participants who were "more towards one style" (Group 1), "more towards the other style" (Group 2), or "somewhere in the middle" (Group 3). This allowed us to deal with the unbalanced subdivisions of the participants in the two opposite learning style sets (such as Visual/Verbal).

Since gaze data distributions were not normal, the examination was carried out using non-parametric statistics. Three gaze metrics were considered, namely percentage of fixation duration, percentage of fixations, and average fixation duration. Percentage values allowed us to take into account the fact that the participants had no time limits and could read/observe a slide for how long they wanted. The time-dependent analysis was implemented through the subdivision of the whole interaction time with the slide into ten intervals.

Significant relations between the Visual/Verbal style and gaze behavior were found for the content layout in which the image is on the left and text is on the right. Specifically, clear distinctions between Groups 1 and 2 were identified using the percentage of fixation duration: considering at least the first 90% of the interaction time with the slide (i.e., measuring values of the metric up to intervals 9 or 10), the text region was looked at more than the picture area by verbal learners and less by visual learners.

Further research can continue to explore different design formats and deal with various types of illustrations, different difficulties of text and topics, and their impact on the learning styles of visualizers and verbalizers. It would have been useful to observe in detail how verbalizers learn only from text and how visualizers learn only from images.

We also tried to recognize other kinds of learning styles (Active/Reflective, Sensible/Intuitive, and Sequential/Global) using the same experimental material. However, as we could have expected, the results were not satisfying, because the investigation of these learning styles would have required different presentations of content, which we will consider in the future. Future work will also include further experiments with new topics, different content layouts, and more varied participants.

The automatic recognition of users' learning styles is a very important step towards *intelligent adaptive learning platforms*. To achieve an adaptive e-learning system, it is essential to monitor the learner behavior dynamically to diagnose their learning style. Eye tracking can serve that purpose by investigating the eye gaze movement while engaging in the e-learning environment. It would be also useful to consider an application of eye tracking technology in combination with other biosensor systems. Additional tools and analytical data might explore hidden patterns in user behavior and activities. In particular, this should be taken into account when working on the implementation of adaptive tutoring systems..We think that the research presented in this paper can provide a useful contribution to gaze-based learning style research, stimulating further studies on the subject.

References

1. Jena, R. K.: Predicting students' learning style using learning analytics: a case study of business management students from India. Behaviour & Information Technology, Vol. 37, N. 10-11, 978-992. (2018)
2. Heidrich, L., Barbosa, J.L.V., Cambruzzi, W., Rigo, S.J., Martins, M.G., Dos Santos, R.B.S.: Diagnosis of learner dropout based on learning styles for online distance learning. Telematics and Informatics, Vol. 35, Issue 6, 1593-1606. (2018)
3. Felder, R.M., Silverman, L.K.: Learning and Teaching Styles in Engineering Education. Engr. Education, 78(7), 674–681. (1988)
4. Franzoni, A.L., Assar, S.: Student Learning Styles Adaptation Method Based on Teaching Strategies and Electronic Media. Educational Technology and Society, 12(4), 15–29. (2009)
5. Duchowski, A.: Eye Tracking Methodology – Theory and Practice (2nd edition), Springer, London. (2007)
6. Felder, R.M., Soloman, B.A.: Index of Learning Styles (ILS). Available: http://www4.ncsu.edu/unity/lockers/users/f/felder/public/ILSpage.html [Accessed: 18 May 2020].
7. Hughes, A., Wilkens, T., Wildemuth, B.M., Marchionini, G.: Text or pictures? An eyetracking study of how people view digital video surrogates. In Proc. 2003 International Conference on Image and Video Retrieval. Springer, 271–280. (2003)
8. Kirby, J.R., Moore, P.J., Schofield, N.J.: Verbal and visual learning styles. Contemp. Educ. Psychol. 13, 169–184. (1988)
9. Tsianos, N., Germanakos, P., Lekkas, Z., Mourlas, C., Samaras, G.: Eye-tracking users' behavior in relation to cognitive style within an e-learning environment. In Proc. 9th IEEE International Conference on Advanced Learning Technologies (ICALT 2009), 329–333. (2009)
10. Riding, R., Cheema, I.: Cognitive styles—an overview and integration. Educ. Psychol. 11, 193–215. (1991)
11. Al-Wabil, A., ElGibreen, H., George, R.P., Al-Dosary, B: Exploring the validity of learning styles as personalization parameters in eLearning environments: An eyetracking study. In Proc. 2010 2nd International Conference on Computer Technology and Development (ICCTD), 174–178. (2010)
12. Mehigan, T.J., Barry, M., Kehoe, A., Pitt, I.: Using eye tracking technology to identify visual and verbal learners. In Proc. 2011 IEEE International Conference on Multimedia and Expo, 1–6. (2011)
13. Cao, J., Nishihara, A.: Understanding Learning Style by Eye Tracking in Slide Video Learning. J. Educ. Multimed. Hypermedia 21, 335–358. (2012)
14. Alyahya, S.M.: Minard's graph of Napoleon's March to Moscow: An eye tracking study of cognitive processing. Ph.D. Thesis (advisers: Linda L.L., Gall, J.E.), University of Northern Colorado. (2014)
15. Mayer, R.E., Massa, L.J.: Three facets of visual and verbal learners: Cognitive ability, cognitive style, and learning preference. J. Educ. Psychol. 95, 833. (2003)
16. Nisiforou, E., Laghos, A.: Field Dependence–Independence and Eye Movement Patterns: Investigating Users' Differences Through an Eye Tracking Study. Interact. Comput. 28, 407–420. (2016)
17. Ekstrom, R.B., French, J.W., Harman, H.H., Dermen, D.: Manual for kit of factor-referenced cognitive tests. Educational Testing Service. Available: https://www.ets.org/Media/Research/

pdf/Manual_for_Kit_of_Factor-Referenced_Cognitive_Tests.pdf [Accessed: 12 June 2020]. (1976)

18. Goswami, A., Walia, G., McCourt, M., Padmanabhan, G.: Using Eye Tracking to Investigate Reading Patterns and Learning Styles of Software Requirement Inspectors to Enhance Inspection Team Outcome. In Proc. 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM '16. ACM, New York, NY, USA, 34:1–34:10. (2016)
19. Koć-Januchta, M., Höffler, T., Thoma, G. B., Prechtl, H., Leutner, D.: Visualizers versus verbalizers: Effects of cognitive style on learning with texts and pictures–An eye-tracking study. Computers in Human Behavior, 68, 170-179. (2017)
20. Höffler, T. N., Koć-Januchta, M., and Leutner, D.: More Evidence for Three Types of Cognitive Style: Validating the Object-Spatial Imagery and Verbal Questionnaire Using Eye Tracking when Learning with Texts and Pictures. Applied cognitive psychology, 31(1), 109-115. (2017)
21. Raptis, G. E., Katsini, C., Belk, M., Fidas, C., Samaras, G., and Avouris, N.: Using eye gaze data and visual activities to infer human cognitive styles: method and feasibility studies. In Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, 164-173. (2017)
22. Alhasan, K., Chen, L., hen, F.: An Experimental Study of Learning Behavior in an ELearning Environment. In Proc. of the 16th IEEE International Conference on Smart City (SmartCity-2018). (2018)
23. Blazhenkova, O., Kozhevnikov, M.: The new object–spatial–verbal cognitive style model: Theory and measurement. Applied Cognitive Psychology, 23, 638–663. (2009)
24. Paivio, A., Harshman, R.: Factor analysis of a questionnaire on imagery and verbal habits and skills. Canadian Journal of Psychology/Revue canadienne de psychologie, 37(4), 461. (1983)
25. Marks, D.F.: Visual imagery differences in the recall of pictures. British Journal of Psychology, 64, 17-24. (1973)
26. Richardson, A.: Verbalizer-Visualizer: a cognitive style dimension. Journal of Mental Imagery, 1, 109-126. (1977)
27. The Eye Tribe. Available: https://en.wikipedia.org/wiki/The_Eye_Tribe [Accessed: 21 July 2020].
28. Luo, Z., O'Steen, B., & Brown, C. (2020). The use of eye-tracking technology to identify visualisers and verbalisers: accuracy and contributing factors. *Interactive Technology and Smart Education*.
29. Luo, Z., & Wang, Y. (2019). Eye-tracking technology in identifying visualizers and verbalizers: data on eye-movement differences and detection accuracy. *Data in brief*, 26, 104447.30. El Guabassi, I., Bousalem, Z., Al Achhab, M., & EL Mohajir, B. E. (2019). Identifying learning style through eye tracking technology in adaptive learning systems. *International Journal of Electrical & Computer Engineering (2088-8708)*, *9*(5).
31. Wibirama, S., Sidhawara, A. P., Pritalia, G. L., & Adji, T. B. (2020, September). A Survey of Learning Style Detection Method using Eye-Tracking and Machine Learning in Multimedia Learning. In *2020 International Symposium on Community-centric Systems (CcS)* (pp. 1-6). IEEE.
32. Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education*, *125*, 413-428.

Nahumi Nugrahaningsih is a researcher at the University of Palangkaraya (Indonesia). She obtained her master's degree in Informatics from the Bandung Institute of Technology (Indonesia) in 2007 and her PhD in Electronics, Computer Science and

Electrical Engineering from the University of Pavia (Italy) in 2017. Her research interests include eye tracking and vision-based user interfaces for e-learning and biometric applications.

Marco Porta is associate professor at the University of Pavia (Italy). He received the master’s degree in Electronic Engineering from the Politechnic of Milan (Italy) in 1996 and the PhD degree in Electronic and Computer Engineering from the University of Pavia in 1999. His research interests include eye tracking, vision-based perceptive interfaces, e-learning, user centered design, and Human-Computer Interaction in general.

Aleksandra Klasnja-Milicevic holds the position of Associate Professor at the Faculty of Sciences, University of Novi Sad (Serbia), where she received her PhD in Computer science, in 2013. Her research interests include e-learning, technology-enhanced learning (TEL), adaptivity and personalization, educational technology, learning analytics, and recommender systems.

Received: December 20, 2019; Accepted: May 02, 2020

Computer Science and Information Systems 18(4):1311–1332 https://doi.org/10.2298/CSIS210227048T

Compensation of Degradation, Security, and Capacity of LSB Substitution Methods by a new Proposed Hybrid n-LSB Approach

Kemal Tütüncü and Özcan Çataltaş*

Faculty of Technology,Selcuk University,
42130 Konya, Turkey
{ktutuncu, ozcancataltas}@selcuk.edu.tr

Abstract. This study proposes a new hybrid n-LSB (Least Significant Bit) substitution-based image steganography method in the spatial plane. The previously proposed n-LSB substitution method by authors of this paper is combined with the Rivest-Shamir-Adleman (RSA), RC5, and Data Encryption Standard (DES) encryption algorithms to improve the security of the steganography, which is one of the requirements of steganography, and the Lempel-Ziv-Welch (LZW), Arithmetic and Deflate lossless compression algorithms to increase the secret message capacity. Also, embedding was done randomly using a logistic map-based chaos generator to increase the security more. The classical n-LSB substitution method and the proposed hybrid approaches based on the previously proposed n-LSB were implemented using different secret messages and cover images. When the results were examined, it has been seen that the proposed hybrid n-LSB approach showed improvement in all three criteria of steganography. The proposed hybrid approach that consists of previously proposed n-LSB, RSA, Deflate, and the logistic map had the best results regarding capacity, security, and imperceptibility.

Keywords: image steganography, lossless compression, logistic map, data encryption.

1. Introduction

Today, the use of the internet and other technological tools in communication between people is widespread. According to a survey conducted among OECD countries in 2019, internet access on an individual basis has increased from 45.7% to 85.6% between 2005-2018[1]. As the use of the internet increases in communication between people, privacy concerns also increase. For this reason, efforts to ensure privacy in communication have increased.

Steganography is one of the data hiding sciences that aims to increase confidentiality in communication [2-3]. The primary purpose of steganography is to conceal the existence of a secret message. This purpose is the most crucial feature that distinguishes it from other data hiding sciences. Since a media file containing a secret message will

* Corresponding author

not be attracted by the third party viewing the message, the secret message will not arise. Therefore, researchers' interest in this subject has increased gradually.

Unlike other data hiding methods, the message is hidden in another media file called a cover or carrier in steganography. This file type can be text, image, audio, video and, etc. The embedded message received by the recipient is converted to the original message by reverse conversion. Since the image is used more in communication between people, the studies have focused on the image file as cover media [4].

Cryptography and watermarking are two other concepts used to provide digital information security. In cryptography, the encrypted output of the encryption algorithm attracts the third person's attention to extract the original message[5]. On the other hand, although both techniques have a data hiding scheme, the intent of watermarking is different from steganography. Steganography aims to conceal the existence of any secret message, while watermarking makes it challenging to remove or manipulate the message.

Steganography algorithms can generally be divided into two categories: spatial domain and transform domain [6]. In the spatial domain, the secret message's bits are embedded into the cover image by directly manipulating the pixel values. On the other hand, in the transform domain, a secret message is placed in the frequency coefficients calculated from the cover image's pixel values using some mathematical functions. The methods applied in the spatial domain have less computation and time complexity but are relatively less resistant to some attacks. The algorithms in the spatial domain have a very high embedding capacity with very poor perceptibility.

In practice, while designing a steganography algorithm, three main features must be considered carefully: imperceptibility, embedding capacity, and security [7]. The embedding capacity and imperceptibility of the stego image are inversely proportional. As the embedding capacity increases, the quality of the stego image decreases. Therefore, using compression methods before embedding the secret message will increase the capacity of the cover media and reduce the detectability of the secret message.

The third feature, security, provides resistance against attacks that is subject to steganalysis. Although steganography's main feature is that it is not suspicious, the message can be obtained in case of possible detection of embedding algorithm. Therefore, encrypting the secret message with known cryptology algorithms before embedding it will increase communication security.

Another way to improve security in steganography is to embed the secret message randomly instead of sequentially. For this purpose, the embedding process can be done with the help of numbers generated by random number generators [8]. In literature, pseudo-random generators and chaos-based generators are generally used as random number generators.

In this study, we have hybridized the different compression methods to increase the embedding capacity of the n-LSB substitution method we introduced in another study [9] and with different encryption methods to increase security. Additionally, we increased security by using a chaos-based (logistic map) embedding algorithm regarding compressed and encrypted messages. These hybrid approachesare tested in different size messages and different images, and the results were compared. It has been seen that the proposed hybrid system compensated degradation, security, and capacity of classical n-LSB based image steganography.

The paper is organized as follows: In the second part, the existing studies in the literature are examined. In the third part, classical n-LSB substitution method, data compression methods, data encryption methods, random number generators, and image quality evaluation methods are mentioned. In the fourth section, the n-LSB substitution method [9] and the proposed hybrid methods are explained. The obtained results are shown in the fifth chapter. In the sixth section, the results are interpreted, and suggestions are made about future works.

2. Related Works

In this study, the proposed hybrid methods have been compared with the classical LSB substitution method as can be seen in the following section. Thus, we will include studies in the literature that modified the LSB substitution method or combined it with compression and encryption methods.

In our previous study [9], the classical n-LSB substitution method was improved and a new version of n-LSB was proposed and tested on different images. Obtained stego images were compared with stego images obtained by the classical n-LSB substitution method. The proposed n-LSB method caused an increase of 6.6% in the Peak Signal to Noise Ratio (PSNR) value regarding the classical n-LSB substitution method.

In their study, Rajput et al. used RSA cryptography and Spatial Orientation Tree Wavelet (STW) compression methods for hiding a secret message in color and gray-scale images. The secret message was encrypted using the RSA encryption algorithm, then embedded in the cover image compressed by the STW compression algorithm. They tested their method using 8 different cover images and obtained PSNR values ranging from 77.3 dB to 83.9dB [10].

Chen has proposed a new module-based LSB substitution method. In this method, the repeated bits in the secret message are detected and the repeated bits are coded with a code. He tested his method by hiding 7 different gray-scale images at 256x512 pixel resolution in 2 different gray-scale images at 512x512 pixel resolution and obtained the PSNR values ranging from 34dB to 36dB in the test result [11].

Akhtar and colleagues [12] proposed a new module-based LSB steganography method by developing the algorithm proposed by Chen [11]. They tested their method by hiding 10 different gray-scale images with 256x256 pixel resolution in 2 different gray-scale cover images with 512x512 pixel resolution. They obtained PSNR values ranging from 34dB to 40dB in the test result. According to the classical LSB method, they obtained increases of between 3% and 25% regarding PSNR. At the same time, they also applied the method suggested by Chen and emphasized that they achieved a higher PSNR value in their method.

Chikouche combined the classical LSB substitution method with the Advanced Encryption Standard (AES) cryptography method and the Deflate compression method in his work. The LSB substitutionmethod was implemented randomly with a pseudo-random generator rather than sequentially. They embedded 3264-bit data in a color cover image with 512x512 pixel resolution and emphasized that their method is better than according to the security criterion [13].

In their study, Manjula and Shivakumar compressed the message they encrypted with AES and Elliptic Curve Cryptography (ECC) with the LZW algorithm and embedded it with the classical LSB substitution method. 32-bit messages were hidden in different images and the PSNR values ranging from 79dB to 81dB values were obtained. Then the messages with a length ranging from 32 bits to 288 bits were hidden in different images and the PSNR values ranging from 77dB to 81dB were obtained. Also, they stated that they have 2 times security because the message is encrypted twice [14].

Kasapbaşi proposed a new image steganography scheme including chaos-based Huffman encoding algorithm and fractal encryption. Firstly, he calculated the frequency of the alphabets and other characters in a section of Turkish newspaper and encoded them with Huffman encoding. He encoded the compressed text with random numbers generated by the logistic map. The message was embedded in the selected LSBs of the cover image. The proposed method was found to be successful in terms of encryption [15].

Rachmawanto et al. proposed a hybrid method consists of the AES cryptology method and classical LSB substitution method. They tested their method and obtained PSNR values ranging from 58 dB to 80dB [16].

SupriadiRustad et al. proposed a new image steganography method based on finding an adaptive pattern in inverted LSB steganography. They obtained thePSNR value ranges from 52.49 to 57.45, and the SSIM ranges from 0.9991 to 0.9999 [7].

3. Materials and Methods

3.1. LSB Substitution Method

The basic principle of the LSB substitution method is to replace the LSB of each pixel with the message bit in the order of the cover image[17]. It can be applied to RGB or gray-scale images. The value of each pixel, which consists of 8 bits, 0 to 255, is either increased by 1, decreased by 1, or unchanged. A change of ±1 in the image pixel will not make a big difference on the image.

3.2. Data Compression Methods

According to the compression formats, data compression methods are divided into two categories: lossy compression and lossless compression [18]. If the original data can be recovered without any changes after compressing the data, this type of compression is called lossless compression. Lossless compression methods ensure that the original data is preserved precisely and that no detail is desired to be lost. In the other category of compression algorithms, lossy compression, original data cannot be obtained precisely after the recovery. In this study, LZW, Arithmetic, and Deflate algorithms had been chosen to compress the message before it was hidden. Detailed information about these algorithms is shown below.

Lempel-Ziv-Welch

The LZW algorithm is a compression method derived from the LZ78 algorithm [18]. It was discovered in 1984 by Terry A. Welch and introduced in his paper titled “A Technique for High-Performance Data Compression” (1984).

There is no preset dictionary in the LZW compression method. Dictionary is created dynamically according to the context to be compressed. For this reason, when the LZW method is used to compress the secret message in steganography, the sender does not need to transmit a dictionary to the recipient. Once the recipient has extracted the compressed message from the stego image, the dictionary will be dynamically created, and the secret message will be obtained.

Arithmetic

The primary purpose of arithmetic coding is to assign an interval to each character. Then, this range is assigned a decimal number. The algorithm starts with 0 and 1 intervals. After reading each character in the input data, the interval is divided into subparts as a smaller range than the input character's probability. This sub-range becomes the new range and is partitioned according to the probability of that character. This process is repeated for each input character. When this is done, every floating point in the last interval uniquely represents the input data [18-20].

Deflate

Deflate is a popular compression method used in well-known algorithms such as Zip and Gzip. Deflate method is used by many important programs such as PNG image, HTTP protocol, and PDF. The Deflate method is a dictionary-based compression technique based on LZ77 and Huffman coding. There are three different modes in Deflate method. In the first mode, input symbols are subdivided without compression. This mode is used for non-compressible files or when someone wants to partition a file without compression. The second mode is a single-pass compression solution. In this mode, a predetermined coding table is used during coding. This mode is used in real-time applications [21]. The third mode of Deflate is a two-pass compression solution based on the dictionaries produced according to the statistical properties of the input file.

3.3. Data encryption methods

RSA

The RSA encryption algorithm was proposed by Ron Rivest, Adi Shamir, and Leonard Adleman in 1977 [22]. The expansion of the RSA consists of the initials of the names of the developers. The RSA algorithm is one of the asymmetric encryption methods. Two different keys are used for encryption and decryption, but these two keys are related to each other. The key used for encryption is a public key and is known by everyone. The key used for decryption is a private key and is only found on the receiving side [23].

The encryption steps of the RSA algorithm are as follows:

1. Two prime numbers, such as p and q, are input parameters.
2. The value of $n = p * q$ is the base value, and $\varphi(n) = (p - 1) * (q - 1)$Euler valueis calculated.
3. The number of e (public keys) is selected as $1 < e < \varphi(n)(\varphi(n)$isa prime number).
4. d value is selected so that $d * e = 1\ mod(\varphi(n))$. This value is a private key.
5. The $c = m^e\ mod(n)$ formula encrypts each message character.

To extract an encrypted message using the RSA algorithm, the first four steps are applied in the same way, andthen the secret message is obtained with the formula $m = c^d\ mod(n)$.

RC5

The RC5 algorithm is one of the symmetric encryption algorithms. It was introduced by Ron Rivest in 1994. The RC5 algorithm is simple to implement because it uses basic mathematical and logical operators. Furthermore, the variable key length distinguishes RC5 from traditional encryption methods such as Data Encryption Standard (DES). The implementation steps of the RC5 encryption algorithm are presented below [24, 25]:

1. Firstly, define *w*, *r*, and Key parameters.
2. Obtain *P* and *Q* constants.

```
  P=odd(e-2)2w
  Q=odd(Φ-2)2^w
```

3. Convert Key *K* byte to words.

```
for i=b-1 to 0
    L[i/u] = (L[u/i] << 8) + K[i]
```

4. Initialize key-independent array, *S*.

```
S[0] = P
for i = 1 to 2(r+1)-1
    S[i] = S[i-1] + Q)
```

5. Mix secret key in the L and S array.

```
i = j = 0
A = B = 0
do 3 * max(t, c) times:
    A = S[i] = (S[i] + A + B) << 3
    B = L[j] = (L[j] + A + B) << (A + B)
i = (i + 1) % t
    j = (j + 1) % c
```

6. Divide the input text into w-bit blocks (A and B are two of these blocks) and encrypt each block.

```
A = A + S[0]
B = B + S[1]
for i = 1 to r do:
    A = ((A ^ B) << B) + S[2 * i]
    B = ((B ^ A) << A) + S[2 * i + 1]
```

7. Decrypt using A and B.

```
for i = r down to 1 do:
    B = ((B - S[2 * i + 1]) >> A) ^ A
    A = ((A - S[2 * i]) >> B) ^ B
B = B - S[1]
A = A - S[0]
```

Data encryption standard (DES)

DES was one of the symmetric encryption methods introduced by the National Institute of Standards & Technology (NIST) in 1976 for all government communications. It has also been used for a long time in bank transactions[26].

In the DES algorithm, the input text is divided into blocks. Each has a 64-bit message. A 64-bit key is required in the encryption process.Eight bits of this key are used as parity bits. Encryption is done in 16 rounds.Each round uses a new key that is the 48-bit length. These keys are obtained using the input key. The block diagram of the DES algorithm is shown in Fig.1.

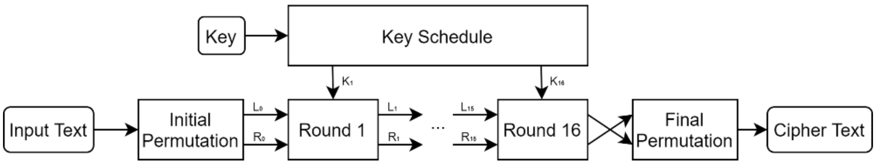


Fig. 1. Block diagram of DES encryption algorithm

3.4. Chaos generator

A random number is a series of numbers or symbols that are not predictable with random luck and do not repeat in a particular pattern. Random number generators have many field-critical presets, such as secure communications, data transmission, and storage. Random number generators are examined in two categories: pseudo-random generators and real random generators.

Chaos-based generators are used more extensively than pseudo-random generators because they are real random generators. Since the chaos generators are very sensitive to input parameters, the numbers they will produce constantly are not predictable. For this reason, it is frequently used in information security applications [27].

One of the simple chaos systems widely used and applied in the literature is the logistic map. The formula of the logistic map is as follow:

$$x_{k+1} = \mu * x_k * (1 - x_k) \tag{1}$$

Here, x_0 and μ are input parameters. When $3.57 < \mu \leq 4$, the system goes into a chaotic state and generates random numbers.

3.5. Image Quality Evaluation Criteria

Image evaluation criteria are methods used to learn the amount of change in the cover image. The image evaluation criteria used in this study are explained below.

Peak Signal-to-Noise Ratio (PSNR)

PSNR is one of the essential criteria used to evaluate image quality. The PSNR is the ratio of the power of the highest possible power of the cover image to the power of the difference between the cover image and the stego image. A high PSNR value means little distortion in the stego image, while a low PSNR value means more distortion in the stego image [28].

The PSNR value can be calculated using the following formula:

$$PSNR = 20 * log \frac{255}{\sqrt{MSE}} \tag{2}$$

$$MSE = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\left(S(i,j) - I(i,j)\right)^2}{m * n} \tag{3}$$

Average Difference

The average difference (AD) metric equals the mean of the sum of the differences between the cover image pixels and the stego image pixels. The low average difference value means that there is less distortion in the stego image.

The average difference formula is:

$$AD = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{S(i,j) - I(i,j)}{m * n} \tag{4}$$

Universal Image Quality Index (UIQI)

UIQI (Universal Image Quality Index) is an index that attempts to model any distortion on the image[29]. These distortions can be in the form of a combination of the following three factors: correlation, luminance distortion, and contrast distortion. The UIQI value is between [-1, 1]. 1 means the images are identical.UIQI formula is shown in (5).

$$UIQI = \frac{\sigma_{xy}}{\sigma_x \sigma_y} * \frac{2\bar{x}\bar{y}}{\bar{x}^2 + \bar{y}^2} * \frac{2\sigma_x \sigma_y}{{\sigma_x}^2 + {\sigma_y}^2} \tag{5}$$

Here, σ denotes standard deviation and $\bar{x}$and $\bar{y}$ denote the average of the cover and stego images, respectively.

4. Proposed Methods

The classical 1-LSB, 2-LSB, and 3-LSB substitution methods are the most common methods used in steganography because of their ease of implementation and high capacity. Many different methods developed are built on these methods. However, they need to be improved because it is easy to detect, and third-person can directly access the message when it is detected.

4.1. Proposed n-LSB Method

The pseudo-code of the application steps of the n-LSB substitution method [9] by the author of this paper is presented in Fig. 2. The proposed n-LSB method reduces the excess change between the cover and stego image pixels while applying the classical n-LSB substitution methods. The proposed n-LSB method has only been applied to 2-LSB and 3-LSB methods because there is no effect on the 1-LSB method and also 4 and more bit substitution methods that are not used in the literature. The 2 and 3-bit implementation of the proposed n-LSB method is described and illustrated below.

Proposed 2-LSB substitution method

When $n = 2$ in the pseudo-code given in Fig. 2, the secret message is embedded in the cover image using the classical 2-LSB substitution method. At the end of the embedding process, the difference between the cover image and the stego image pixels is examined. Since at most two LSBs of the cover image pixels can be changed, the decimal difference will be one of 3, 2, 1, 0, -1, -2, -3. If this difference is -2, -1, 0, 1, or 2, then the pixel of the stego image is left unchanged. In another case, if this difference is 3, all bits from the 3rd LSB of the stego image are examined, the first 0 is converted to 1, and the previous 1s are converted to 0. So the related pixel of the stego image is added to the decimal number 4, the difference between the pixels falls from 3 to -1. If there are no 0's between the 3rd and 8th bits (most significant bit), the related pixels are left unchanged in the stego image. If the difference is -3, all bits from the 3rd bit of the related pixel are examined, the first 1 encountered is converted to 0, and the previous 0s are converted to 1. Thus, the decimal number 4 is subtracted from the corresponding pixel, and the difference between the pixels falls from -3 to 1. If there is no 1 between the 3rd and 8th bits, then the corresponding pixel is left unchanged in the stego image. With this method, the deterioration of the pixels where the degradation is excessive is reduced [9]. Examples of the implementation of the proposed 2-LSB are shown in Table 1.

```
Input:C,S,n
Output: S
C: Cover image pixels
S: Stego image pixels
n: Embedding method (n=2, 3)
$C(p)_i$,$S(p)_i$: *i*th bit of related pixel; i=1,..,8
**For each** p **in***every pixel of* C,S
**If**$C(p) - S(p) > 2^{(n-1)}$
        **If**$S(p)_{n+1} = 0$
        $S(p)_{n+1} = 1$
**Elseif**$S(p)_{n+2} = 0$
        $S(p)_{n+2} = 1$
                    $S(p)_{n+1} = 0$
…
**Elseif**$S(p)_8 = 0$
                    $S(p)_8 = 1$
          $S(p)_7 = \cdots = S(p)_{n+1} = 0$
End If
**Elseif**$C(p) - S(p) < -2^{(n-1)}$
        **If**$S(p)_{n+1} = 1$
        $S(p)_{n+1} = 0$
        **Elseif**$S(p)_{n+2} = 1$
                    $S(p)_{n+2} = 0$
                    $S(p)_{n+1} = 1$
…
**Elseif**$S(p)_8 = 1$
                    $S(p)_8 = 0$
          $S(p)_7 = \cdots = S(p)_{n+1} = 1$
End If
End If
End For each
```

Fig. 2. The pseudo-code of the proposed n-LSB substitution method

In Table 1, randomly generated 2-bit messages were hidden in the randomly generated 20 pixels cover image pixels consisting of 8 bits by the 2-LSB method. After the embedding process, the differences between the pixels of the cover image and the stego image were examined. If the difference is -3 or 3, the proposed method was applied. As shown in the table, the difference in 5 of 20 pixels is -3 or 3, so the pixels outside these 5 pixels were not changed. The AD before the enhancement was 1.2, but after the enhancement, this difference was reduced to 0.9. Thereby the amount of distortion in the image was reduced.

Table 1. Proposed 2-LSB substitution example. The changed bits of cover image pixel obtained after embedding with the 2-LSB method and the bits of stego image pixel obtained after applying the proposed compensation method were shown in red and green color, respectively

Pixel No	Cover image	Message to be embedded (2-bit)	Stego image	Difference	New stego image	New difference
1	01111001	01	011110***01***	0	01111001	0
2	00100011	10	001000***10***	1	00100010	1
3	10010000	11	100100***11***	-3	100***011***11	1
4	00000001	10	000000***10***	-1	00000010	-1
5	11000011	00	110000***00***	3	11000***1***00	-1
6	11000010	11	110000***11***	-1	11000011	-1
7	00101110	01	001011***01***	1	00101101	1
8	11111000	01	111110***01***	-1	11111001	-1
9	00110000	10	001100***10***	-2	00110010	-2
10	11010100	11	110101***11***	-3	11010***0***11	1
11	11100000	11	111000***11***	-3	11***0111***11	1
12	00100000	10	001000***10***	-2	00100010	-2
13	01010001	01	010100***01***	0	01010001	0
14	01101101	10	011011***10***	-1	01101110	-1
15	11101001	10	111010***10***	-1	11101010	-1
16	11001101	01	110011***01***	0	11001101	0
17	01110111	00	011101***00***	3	0111***10***00	-1
18	01111110	10	011111***10***	0	01111110	0
19	01001101	10	010011***10***	-1	01001110	-1
20	10000011	10	100000***10***	1	10000010	1

Proposed 3-LSB substitution method

When $n = 3$ in the pseudo-code given in Fig. 2, the secret message is embedded in the cover image using the classical 3-LSB substitution method. At the end of the embedding process, the difference between the pixels of the cover image and the stego image is examined. Since at most three LSBs of the cover image can be changed, the difference will get one of the values from -7 to 7. If the decimal difference between the pixels is -4, -3, -2, -1, 0, 1, 2, 3, or 4, then the pixel of the stego image is left unchanged. If the difference is 5, 6, or 7, all bits from the 4th bit of that pixel are examined, the first 0 value encountered is converted to 1, and the previous 1s are converted to 0. So the related pixel is added to the decimal number 8, and the difference between the pixels falls from 5, 6, 7 to -3, -2, and -1, respectively. If there is no 0 value between the 4th and 8th bits, the related pixel of the stego image is left unchanged. Similarly, if the difference is -5, -6, or -7, all bits of the pixel are examined from the 4th bit, the first 1 value encountered is converted to 0, and the previous 0s are converted to 1. Thus, the decimal number 8 is subtracted from the related pixel, the difference between the pixels falls from -5, -6, -7 to 3, 2, and 1, respectively. If there is no 1 value between the 4th and 8th bits, then the related pixel is left unchanged. With this method, the deterioration of the pixels where the degradation is excessive is reduced [9]. Examples of the implementation of the proposed compensation method for 3-LSB are presented in Table 2.

Table 2. Proposed 3-LSB substitution example. The changed bits of cover image pixel obtained after embedding with the 3-LSB method and the bits of stego image pixel obtained after applying the proposed compensation method were shown in red and green color, respectively

Pixel No	Cover image	Message to be embedded (3-bit)	Stego image	Difference	New stego image	New difference
1	01111001	111	01111***111***	-6	0111***0***111	2
2	00100011	001	00100***001***	2	00100001	2
3	10010000	111	10010***111***	-7	100***01***111	1
4	00011111	010	00011***010***	5	00***100***010	-3
5	11000011	100	11000***100***	-1	11000100	-1
6	11000010	111	11000***111***	-5	1***0111***111	3
7	00101110	111	00101***111***	-1	00101111	-1
8	11111000	001	11111***001***	-1	11111001	-1
9	00110000	010	00110***010***	-2	00110010	-2
10	11010100	110	11010***110***	-2	11010110	-2
11	11100111	000	11100***000***	7	1110***1***000	-1
12	00100000	101	00100***101***	-5	00***011***101	3
13	01010001	101	01010***101***	-4	01010101	-4
14	01101101	101	01101***101***	0	01101101	0
15	11101001	110	11101***110***	-5	1110***0***110	3
16	11001101	100	11001***100***	1	11001100	1
17	01110111	001	01110***001***	6	0111***1***001	-2
18	01111110	111	01111***111***	-1	01111111	-1
19	01001111	001	01001***001***	6	010***10***001	-2
20	10000011	001	10000***001***	2	10000001	2

In Table 2, randomly generated 3-bit messages were hidden on the randomly generated 20 cover image pixels consisting of 8 bits by the 3-LSB method. After the embedding process, the differences between the pixels of the cover image and the stego image were examined. If the difference is -7, -6, -5, 5, 6, or 7, the proposed method was applied. As shown in the table, the difference in 9 of 20 pixels is -7, -6, -5, 5, 6, or 7, so the pixels outside these 9 pixels were not changed in the stego image. The AD before the enhancement was 3.45, but after the enhancement, this difference was reduced to 1.85. Thereby the amount of distortion in the stego image was reduced.

4.2. Proposed hybrid-1

With the proposed n-LSB method [9], the degradation of pixels of the stego image has been reduced, which is one of the three basic principles of steganography. However, there has been no change in the other principles of steganography, which are security and capacity. To improve these two criteria, existing compression and encryption methods are combined with the proposed n-LSB method.

Compressing the secret message before embedding it in the cover image will reduce the degradation of the stego image as it will reduce the number of secret message bits to be embedded and increase the message length that can be embedded on the cover image. Also, since the 3rd person will not know the used compression algorithm, the secret message will not be directly available, even if the hidden data in the stego image is recovered. For this purpose, the message is compressed using text compression algorithms before being embedded. Three different algorithms, LZW, Arithmetic, and Deflate, have been applied as text compression algorithms and compared among themselves.

Embedding the message sequentially in the cover image makes it easier to extract it by the 3rd person. Encrypting the message before embedding it is a big solution, but since the embedding process is sequential, the 3rd person can quickly get the encrypted message. Although cryptography methods are challenging to break, it is not impossible to break. For this reason, the secret message may be passed on to other people. To overcome this problem, the message in steganography is often randomly embedded in the image with various random number generators rather than sequentially. In our proposed hybrid n-LSB approach, random numbers are generated with the logistic map-based chaos generator to overcome this problem, and the message is randomly embedded in the cover image. The x_0 and μ values of the chaotic generator are used as input parameters. For this reason, these parameters must be transmitted to the recipient to extract the hidden message.

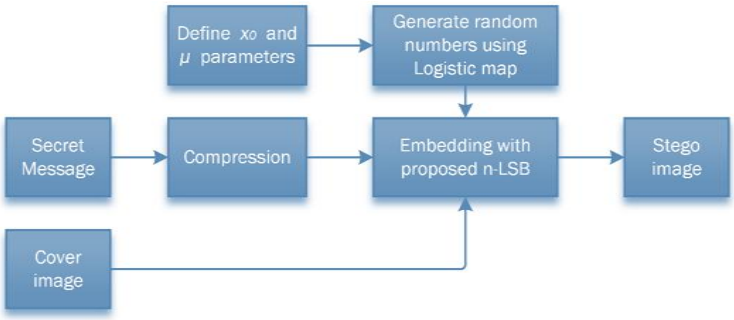

Define x_0 and μ parameters
Generate random numbers using Logistic map
Secret Message
Compression
Embedding with proposed n-LSB
Stego image
Cover image

Fig. 3. The flowchart of the proposed hybrid-1 method

To eliminate the shortcomings of the n-LSB method and obtain a safer embedding algorithm, twodifferent hybrid methods were created.In the first hybrid method, named as proposed hybrid-1, the secret message was compressed using Deflate compression algorithm, and then this compressed secret message was embedded to the cover image randomly using the proposed 2-LSB and 3-LSB substitution method and logistic map-based chaos generator. The reason for choosing Deflate compression algorithm is being superior to the other two methods in compression ratios according to applications presented in Results Section. The flowchart of the proposed hybrid-1 method was shown in Fig. 3.

4.3. Proposed hybrid-2

Encrypting the message before embedding it in the cover image will make it difficult to reach the secret message, even if the third party can completely get the information embedded in the image. For this purpose, the message is encrypted with different encryption algorithms. RSA, DES, and RC5 algorithms were tested and compared for this purpose.

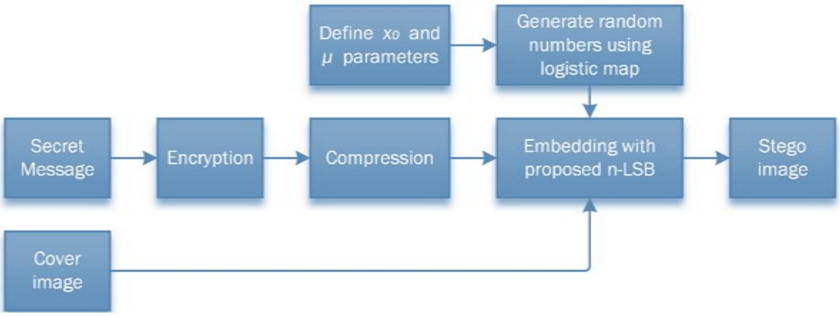

Define x_0 and μ parameters
Generate random numbers using logistic map
Secret Message
Encryption
Compression
Embedding with proposed n-LSB
Stego image
Cover image

Fig. 4. The flowchart of the proposed hybrid-2 method

In this hybrid method, named as proposed hybrid-2, in addition to the proposed hybrid-1 method, the secret message was encrypted before compressing process using the RSA encryption algorithm. The reasons why the RSA algorithm is chosen are its widespread use and being asymmetric encryption algorithm. Additionally, the RSA encryption algorithm is superior to the other two methods both in the average encryption time and in the file size to be encrypted per second according to applications presented in Results Section. The (p, q) values required to generate the public and private keys in RSA encryption are used as input parameters. For this reason, for the receiver to reach the secret message, the values p and q must be transmitted to the receiver. The flowchart of the proposed hybrid-2 method was shown in Fig. 4.

5. Results

In this section, the results obtained by applying the classical LSB substitution and proposed hybrid n-LSB approaches are evaluated. For this aim, three different text files and four different cover images were used. The cover images used are shown in Fig. 5. These images are RGB 24-bit images with a resolution of “Lena” 225x225, “Mandrill” 512x512, “Cat” 960x603, and “Peppers” 600x600 pixels. These images are in the ".bmp" file format. The methods we propose in this study are independent of these images and can be applied to any desired image without any constraints.The main reasons forchoosing these images as cover are being in different resolutions and well-known in the literature.

Three text files with sizes 6.95 kB, 13.59 kB,and 17.13 kB were selected as secret messages. These text files contain the standard English alphabet as well as some special characters (., ?, -, !, “”, ,).

When the results are evaluated, a comparison is made only with the classical LSB substitution method, as seen in the literature. The reason is that each of the methods in the literature is tested on different cover images using different messages. There is no common ground between methods proposed by other authors in the literature.

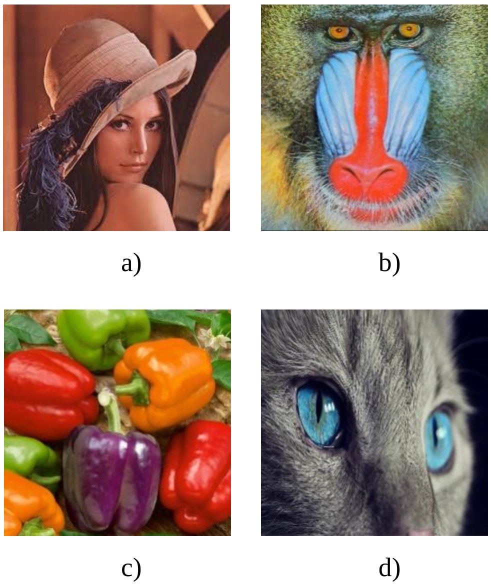


Fig. 5. Cover images used to test proposed methods

5.1. Comparison of data compression methods

In the proposed hybrid methods, the secret message file is compressed using different compression methods before being hidden in the cover image. The new message length and the compression rates resulting from the compression are shown in Table 3. The compression ratio values were calculated according to (6).

$$\%CompressionRatio = \frac{InputSize}{OutputSize} * 100 \tag{6}$$

The original message files with 56980, 111405, and 140364-bit lengths were compressed with the LZW algorithm; their size decreased to 38448, 72969, and 84643-bit, respectively. Compression ratios were obtained as 148.2%, 152.67% and 165.83% respectively. It is seen that as the message length increases, the compression ratio increases, and the highest compression ratio is obtained when the Text-3 file is compressed with LZW. The reason is that as the message length increases, the possibility of finding new words added to the dictionary increases. In other words, when there are 2, 3, or more character words added to the dictionary, the probability of encountering these words increases as the size of the message increases so that the coded words in the dictionary can be used instead of these words.

With the arithmetic algorithm, the secret message files, which are 56980, 111405, and 140364-bit length, have been reduced to the size of 35448, 69361, and 88109 bits respectively. Compression ratios were calculated as 160.74%, 160.62% and 159.31%, respectively. It is seen that as the message size increases, the compression ratio decreases, and the highest compression ratio is obtained when the Text-1 file is compressed with the arithmetic algorithm.

Table 3. Comparison of compression algorithms

Filename	Uncompressed size (bit)	LZW	Arithmetic	Deflate
		Compression ratio		
Text 1	56980	38448 148.20%	35448 160.74%	29496 193.18%
Text 2	111405	72969 152.67%	69361 160.62%	56832 196.03%
Text 3	140364	84643 165.83%	88109 159.31%	65744 213.50%

With the Deflate algorithm, the original message files, which are 56980, 111405, and 140364-bits in length, have been reduced to the size of 29496, 56832, and 65744 bits, respectively. Compression ratios were calculated as 193.18%, 196.03% and 213.50%, respectively. Since the Deflate algorithm is a hybrid algorithm consisting of Huffman and LZ77 codes, the highest compression ratios according to other algorithms were obtained with this algorithm. Also, as the message length increased, the compression ratio increased, and the highest compression ratio was obtained when the Text-3 file was compressed.

Since the highest compression ratio between LZW, Arithmetic, and Deflate algorithms is obtained by Deflate method, it is used as a compression method in the proposed hybrid-1 and hybrid-2 algorithms.

5.2. Comparison of data encryption methods

Determining the security and success of encryption algorithms is a complex process. To compare such algorithms on a common basis, encryption times are generally used. Therefore, the encryption times of the RSA, RC5, and DES encryption algorithms used in this study were calculated using texts of different lengths and are shown in Table 4.

Table 4. Comparison of encryption times of RSA, RC5, and DES

Text Size	RSA	RC5	DES
Text-1 (7122 byte)	10.23	24.91	21
Text-2 (13925 byte)	20.05	48.29	40.05
Text-3 (17545 byte)	24.95	60.97	50.42
Average	18.41	44.72	37.15
Bytes/sec	698.75	287.65	346.27

When Table 4 is examined, it can be seen that the RSA encryption algorithm is superior to other methods both in the average encryption time and in the file size to be encrypted per second.Therefore, RSA was chosen as the encryption method in the proposed hybrid-2 algorithm.

5.3. Test results

Three different secret message files were embedded in 4 different cover images using the classical LSB substitution method, the proposed n-LSB method [9], and proposed hybrid methods. The stego images were compared with the cover images by using image comparison criteria, and the results are shown in the sub-sections.

The following algorithms are used for embedding:

- Classical 1-LSB, 2-LSB, and 3-LSB methods
- Proposed n-LSB method (consist of proposed 2-LSB or 3-LSB methods) [9]
- Proposed hybrid-1 (consist of proposed 2-LSB or 3-LSB method combined with Deflate compression algorithm and logistic map-based random embedding method)
- Proposed hybrid-2 (consist of proposed 2-LSB or 3-LSB method combined with Deflate compression algorithm, RSA encryption algorithm, and logistic map-based random embedding method)

The input parameters used during embedding are:

- RSA encryption: $p = 3, q = 41$.
- Logistic map: $x_0 = 0.675$ and $\mu = 3.9763$.

PSNR

The cover and stego images were submitted to the PSNR test and the obtained results are shown in Appendix.Since the PSNR value is the ratio of the peak signal to the noise in the image, the higher PSNR value means that the image degradation is less. When the obtained results are examined, it is seen that the PSNR values of the proposed 2-LSB and 3-LSB methods are higher than the classical 2-LSB and 3-LSB methods. It can be said that the proposed hybrid-1 is better than the proposed n-LSB method and proposed hybrid-2. The second highest PSNR value was obtained by embedding Text-1 in Image-4 by the proposed hybrid-1 2-LSB method with 64.86235 dB and comes after the classical 1-LSB method. Also, the highest increase was 10.8426% (from 42.96577 dB to 47.62437 dB) when the Text-3 file was embedded to Image-1 by proposed hybrid-1 3-LSB compared to the classical 3-LSB method. It can be said that the highest PSNR increase can be achieved when the high-size message file is embedded into the low-resolution image.

Average Difference

The stego images and cover images obtained after applying the proposed and classical LSB substitution methods are compared according to the AD criterion, and the obtained results are presented in Appendix.The average difference is equal to the average of differences between the cover and stego image pixels. Since the stego image is desired to be similar to the cover image, it is expected that the average difference value is small. When the obtained results are examined, embedding Text-1 message on image-4 is obtained with the smallest mean difference value of 0.042 by the proposed hybrid-1 2-LSB algorithm. It is also seen that the proposed hybrid-1 is superior to other classical and proposed methods. However, as the resolution of the image increases or the length of the secret message decreases, the average difference value decreases.

UIQI

The stego images and cover images obtained after applying the proposed and classical LSB substitution method were compared according to the UIQI criterion, and the obtained results are shown in Appendix.It is preferable to have a high UIQI value because the difference between the stego image and the cover image is desired to be small. The classical 1-LSB method is the least corrupted method because it only changes the last pixels of the cover image. When we compare the proposed n-LSB method with the classical 2-LSB and 3-LSB algorithms, the proposed n-LSB method yields higher UIQI values. Additionally, the results obtained by the proposed hybrid-1 are compatible with the results obtained by the proposed n-LSB.However, since the UIQI value consists of a combination of correlation, luminance distortion, and contrast distortion, the results show differences in different cover images. Therefore, it is not possible to make a clear conclusion about which method is superior according to the UIQI value.

5.4. Capacity test

Capacity can be defined as the maximum amount of secret messages hidden in the cover image. Thus, it is essential to check the capacity of images when steganography methods are compared.

The effect of data compression on stego image capacity is shown in Table 5. The capacity values shown here are estimated values calculated from the compression ratios calculated in Table 3. Besides, since the application of encryption algorithms and other embedding methods such as proposed n-LSB and classical n-LSB do not affect the embedding capacity,only the classical 3-LSB method was used as an embedding method. According to Table 5, the highest capacity increase was achieved by the Deflate algorithm. With this algorithm, the capacity of image-4 was increased from 635.98 kB to 1357.81 kB, a 113.5% increase was obtained. Furthermore, according to the results shown in Table 3, it was obtained that the compression ratio increased as the message size increased. Accordingly, it is expected that with the LZW and the Deflate algorithm, the cover images will have a larger message capacity than the values shown in Table 5.

Table 5. Capacity test results

Image No	Uncompressed Capacity (kB)	Compressed Capacity (Expected) (kB)		
		LZW	Arithmetic	Deflate
Image 1	55.62	92.23	89.40	118.75
Image 2	288.00	477.59	462.93	614.88
Image 3	395.51	655.87	635.74	844.41
Image 4	635.98	1054.64	1022.27	1357.81

6. Conclusion and Discussion

In this paper, steganography methods which are one of the information security methods are examined, and a new hybrid method in image steganography is proposed. This method, which we propose, is based on the proposed n-LSB substitution method of the authors of this paper and tries to reduce the pixel differences between cover and stego image. In this regard, improvement in the perceptibility criterion, one of the three main criteria of steganography, has been achieved and confirmed with an implemented test.

Since the proposed n-LSB method is based on the LSB substitution method, the third party can extract the secret message easily. To solve this problem, instead of embedding the message bits in sequence, they are randomly embedded using a chaos-based random number generator. To increase the security a step ahead, RSA, RC5, and DES encryption algorithms are used to encrypt the secret message before being embedded. Then, data compression methods were combined with the proposed n-LSB method to provide improvement in both the capacity criterion and compensating for the increase in data size that would result in the use of encryption algorithms. Three compression methods, LZW, Arithmetic, and Deflate, were applied. The best compression ratio was obtained by the Deflate algorithm. For this reason, the secret message was compressed with the Deflate algorithm before being hidden in the cover image.

These proposed hybrid methods based on the proposed n-LSB method have been tested by hiding three message files in different sizes in 4 cover images with different resolutions and sizes. The highest PSNR value was obtained as 64.86 dB with proposed hybrid-1 (2-LSB), and the highest PSNR increase rate was 10.84% with proposed hybrid-1 (3-LSB) when the stego images and the cover images were compared according to image quality evaluation criteria. PSNR values were higher in all the different combinations of the proposed n-LSB method than in the classical LSB method. Moreover, the use of the Deflate compression algorithm in the proposed hybrid-1 method resulted in an increase of 113.5% in the embedding capacities of the cover images.

Thanks to the proposed hybrid methods, the shortcomings in using the n-LSB method have been eliminated, and more reliable methods have been obtained for data hiding. The proposed n-LSB and hybrid methods can be used regardless of the message and the cover image, as long as the secret message size does not exceed the capacity of the cover image.

The authors of this paper study the effects of the application of the combination of different compression and encryption algorithms with the proposed n-LSB method to different color spaces. Furthermore, the authors think that investigating the applicability of the proposed methods in the frequency domain will be a good research step.

Appendix

Table 6. Test results obtained by applying proposed methods

Metric	Image No	Text No	Classical			Proposed n-LSB		Proposed Hybrid-1		Proposed Hybrid-2	
			1-LSB	2-LSB	3-LSB	2-LSB	3-LSB	2-LSB	3-LSB	2-LSB	3-LSB
PSNR	Image 1	Text-1	55.37	51.30	46.85	53.62	49.69	54.74	50.42	52.79	48.57
		Text-2	52.46	48.41	43.98	50.69	46.82	52.21	48.06	50.52	46.49
		Text-3	51.46	47.39	42.97	49.68	45.83	51.72	47.62	50.04	46.10
	Image 2	Text-1	62.48	58.51	53.99	60.73	56.86	61.51	57.12	59.37	55.00
		Text-2	59.61	55.58	51.10	57.86	53.97	58.74	54.35	56.60	52.26
		Text-3	58.60	54.59	50.14	56.85	52.98	58.14	53.78	55.93	51.57
	Image 3	Text-1	63.83	59.82	55.36	62.06	58.10	62.87	58.44	60.74	56.24
		Text-2	60.93	56.91	52.44	59.09	55.19	60.04	55.60	57.83	53.45
		Text-3	59.96	55.89	51.48	58.03	54.13	59.46	55.01	57.14	52.73
	Image 4	Text-1	65.94	61.94	57.46	64.09	60.08	64.86	60.42	62.77	58.28
		Text-2	63.06	59.03	54.49	61.17	57.12	62.12	57.63	59.90	55.46
		Text-3	62.07	58.04	53.47	60.17	56.09	61.49	57.05	59.21	54.74
AD	Image 1	Text-1	0.0625	0.0793	0.1108	0.0627	0.0840	0.0443	0.0666	0.0703	0.1030
		Text-2	0.1226	0.1548	0.2155	0.1228	0.1634	0.0807	0.1167	0.1224	0.1708
		Text-3	0.1544	0.1953	0.2716	0.1548	0.2055	0.0907	0.1298	0.1379	0.1891
	Image 2	Text-1	0.0121	0.0152	0.0214	0.0121	0.0162	0.0092	0.0141	0.0151	0.0230
		Text-2	0.0236	0.0297	0.0417	0.0236	0.0316	0.0175	0.0267	0.0288	0.0435
		Text-3	0.0298	0.0375	0.0522	0.0298	0.0396	0.0200	0.0305	0.0336	0.0509
	Image 3	Text-1	0.0089	0.0112	0.0156	0.0089	0.0119	0.0067	0.0103	0.0110	0.0171
		Text-2	0.0174	0.0218	0.0305	0.0174	0.0232	0.0129	0.0199	0.0215	0.0327
		Text-3	0.0218	0.0275	0.0382	0.0220	0.0294	0.0148	0.0228	0.0252	0.0385
	Image 4	Text-1	0.0055	0.0069	0.0096	0.0055	0.0074	0.0042	0.0066	0.0069	0.0107
		Text-2	0.0107	0.0134	0.0190	0.0108	0.0146	0.0080	0.0125	0.0134	0.0207
		Text-3	0.0134	0.0169	0.0240	0.0136	0.0185	0.0092	0.0143	0.0157	0.0243
UIQI	Image 1	Text-1	0.9914	0.9789	0.9645	0.9859	0.9750	0.9908	0.9786	0.9863	0.9708
		Text-2	0.9867	0.9690	0.9441	0.9791	0.9612	0.9844	0.9683	0.9787	0.9595
		Text-3	0.9847	0.9629	0.9357	0.9750	0.9560	0.9829	0.9662	0.9771	0.9571
	Image 2	Text-1	1	0.9999	0.9999	1	1	0.9998	0.9997	0.9997	0.9995
		Text-2	0.9999	0.9995	0.9998	0.9999	0.9999	0.9996	0.9994	0.9994	0.9991
		Text-3	0.9999	0.9994	0.9997	0.9999	0.9999	0.9996	0.9993	0.9993	0.9989
	Image 3	Text-1	0.9994	0.9940	0.9985	0.9994	0.9991	0.9949	0.9938	0.9927	0.9904
		Text-2	0.9999	0.9991	0.9998	0.9999	0.9999	0.9996	0.9994	0.9993	0.9990
		Text-3	0.9971	0.9962	0.9945	0.9971	0.9962	0.9992	0.9885	0.9870	0.9825
	Image 4	Text-1	0.9989	0.9945	0.9989	0.9991	0.9991	0.9961	0.9933	0.9940	0.9869
		Text-2	0.9972	0.9922	0.9969	0.9976	0.9976	0.9937	0.9894	0.9909	0.9852
		Text-3	0.9961	0.9962	0.9958	0.9968	0.9968	0.9928	0.9885	0.9898	0.9840

References

1. OECD:ICT Access and Usage by Households and Individuals. [Online]. Available: https://stats.oecd.org/Index.aspx?DataSetCode=ICT_HH2 (current September 2021)
2. Luo, X., Wang, D., Wang, P., Liu, F.: A Review On Blind Detection For Image Steganography. Signal Processing, Vol. 88, No. 9, 2138-2157. (2008)
3. Chen, K., Lin, C., Zhong, S., Guo, L.:A Parallel SRM Feature Extraction Algorithm For Steganalysis Based On GPU Architecture. Computer Science and Information Systems, Vol. 12, No. 4, 1345-1359. (2015)

4. Karakus, S., Avci, E.:A New Image Steganography Method With Optimum Pixel Similarity For Data Hiding In Medical Images. Medical Hypotheses, Vol.139. (2020)
5. Tian, Q., Han, D., Jiang, Y.: Hierarchical Authority Based Weighted Attribute Encryption Scheme. Computer Science and Information Systems,Vol. 16, No. 3, 797-813. (2019)
6. Sharafi, J., Khedmati, Y., Shabani, M.M.:Image Steganography Based On A New Hybrid Chaos Map And Discrete Transforms. Optik, Vol. 226, No. 2. (2021)
7. Rustad, S., Setiadi, D., Syukur, A., Andono, P.:Inverted LSB Image Steganography Using Adaptive Pattern To Improve Imperceptibility. Journal of King Saud University - Computer and Information Sciences, Early Access. (2021)
8. Roy, R.,Sarkar, A., Changder, S.:Chaos based Edge Adaptive Image Steganography. Procedia Technology,Vol. 10, 138-146. (2013)
9. Cataltas, O., Tutuncu, K.: Improvement Of LSB Based Image Steganography. International Journal Of Electrical, Electronics And Data Communication, Vol. 5, 1-5. (2017)
10. Rajput, V., Tiwari, S.K., Gupta, R.: An Enhanced Image Security Using Improved RSA Cryptography And Spatial Orientation Tree Compression Method. International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES).Paralakhemundi, India, 327-331. (2016)
11. Chen, S.K.: A Module-Based LSB Substitution Method With Lossless Secret Data Compression. Computer Standards & Interfaces,Vol. 33, 367-371. (2011)
12. Akhtar, N., Ahamad, V., Javed, H.: A Compressed LSB Steganography Method. 3rd International Conference on Computational Intelligence & Communication Technology (CICT).Ghaziabad, India, 1-7. (2017)
13. Chikouche,S.L., Chikouche,N.: An Improved Approach For Lsb-Based Image Steganography Using AES Algorithm. 5th International Conference on Electrical Engineering - Boumerdes (ICEE-B). Boumerdes, Algeria, 1-6. (2017)
14. Manjula,Y., Shivakumar,K.B.: Enhanced Secure Image Steganography Using Double Encryption Algorithms. 3rd International Conference on Computing for Sustainable Global Development (INDIACom).New Delhi, India, 705-708. (2016)
15. Kasapbaşi,M.C.: A New Chaotic Image Steganography Technique Based on Huffman Compression of Turkish Texts and Fractal Encryption With Post-Quantum Security. IEEE Access,Vol. 7, 148495-148510. (2019)
16. Rachmawanto,E.H., Amin,R.S., Setiadi,D.I.M., Sari,C.A.: A Performance Analysis StegoCrypt Algorithm based on LSB-AES 128 bit in Various Image Size. International Seminar on Application for Technology of Information and Communication (Isemantic). Semarang, Indonesia, 16-21. (2017)
17. Jain,N., Meshram,S., Dubey,S.: Image Steganography Using LSB and Edge – Detection Technique. International Journal of Soft Computing and Engineering, Vol. 2, 217-222. (2012)
18. Shanmugasundaram,S., Lourdusamy,R.: A Comparative Study Of Text Compression Algorithms. International Journal of Wisdom Based Computing, Vol. 1, 68-76. (2011)
19. Langdon,G.G., Rissanen,J.: Compression of Black-White Images with Arithmetic Coding. IEEE Transactions on Communications, Vol. 29, No. 6, 858-867.(1981)
20. Langdon,G.G.: An Introduction to Arithmetic Coding. IBM Journal of Research and Development, Vol. 28, No. 2, 135-149. (1984)
21. Tahghighi,M., Mousavi,M., Khadivi,P.: Hardware Implementation Of A Novel Adaptive Version Of Deflate Compression Algorithm.18th Iranian Conference on Electrical Engineering.Isfahan, Iran, 566-569. (2010)
22. Rivest,R.L., Shamir,A., Adleman,L.: A Method For Obtaining Digital Signatures Ad Public-Key Cryptosystems. Communications of the ACM, Vol. 21, No. 2,120-126.(1977)
23. Xin,Z., Xiaofei,T.: Research And Implementation Of RSA Algorithm For Encryption And Decryption. Proceedings of 2011 6th International Forum on Strategic Technology.Harbin, China, 1118-1121. (2011)

25. Rivest,R.L., Preneel, B. (ed.):The RC5 Encryption Algorithm.Fast Software Encryption. Springer Berlin Heidelberg, Berlin, Heidelberg, 86-96. (1995)
25. Shahzadi,R., Anwar,S.M., Qamar,F., Ali,M., Rodrigues,J.J.P.C.: Chaos Based Enhanced RC5 Algorithm ForSecurity And Integrity Of Clinical Images In Remote Health Monitoring. IEEE Access, Vol. 7, 52858-52870.(2019)
26. McLoone,M., McCanny,J.V.: High-Performance FPGA Implementation OfDES Using A Novel Method For Implementing The Key Schedule. IET Proceedings - Circuits, Devices and Systems,Vol. 150, No. 5, 373-378. (2003)
27. Patidar,V., Sud,K.K.: A Pseudo Random Bit Generator Based on Chaotic Logistic Map and its Statistical Testing. Informatica,Vol. 33, No. 4, 441-452.(2009)
28. Kadhim,I.J., Premaratne, P., Vial,P.J., Halloran,B.: Comprehensive Survey Of Image Steganography: Techniques, Evaluations, And Trends In Future Research. Neurocomputing,Vol. 335, 299-326. (2019)
29. Zhou,W., Bovik,A.C.: A Universal Image Quality Index. IEEE Signal Processing Letters,Vol. 9, No. 3, 81-84. (2002)

Kemal Tütüncü was born in Konya, Turkey, in 1975. He received the master’s degree from Free University of Brussel, Belgium. He received Ph.D. degrees from Selcuk University, Turkey. His research interest includescryptology, information security, natural language processing, and artificial intelligence.

Özcan Çataltaş was born in Konya, Turkey, in 1992. He received the master’s degree from Selcuk University, Turkey. His research interest includes information security and artificial intelligence.

Received: February 27, 2021; Accepted: July 05, 2021.

Computer Science and Information Systems 18(4):1333–1358 https://doi.org/10.2298/CSIS200820025D

The Dynamic Two-echelon MSW Disposal System Study under Uncertainty in Smart City

Feng Dai[1], Gui-hua Nie[2], and Yi Chen[3]

[1] Research Center for Mining and Metallurgy Culture and Social-economic Development in the
Middle Reaches of Yangtze River, Hubei Polytechnic University,
Huangshi, China
13964571@qq.com
[2] School of Economics, Wuhan University of Technology,
Wuhan, China
niegh@whut.edu.cn
[3] School of Economics and Management, Hubei Polytechnic University,
Huangshi, China
208052@ hbpu.edu.cn

Abstract. The municipal solid waste (MSW) disposal system is the key for building the smart city. In the MSW disposal system, the MSW is allocated among the disposal plants in the first echelon, and then the derivatives (incineration residues and RDF) are allocated between residues disposal plants and markets in the second echelon. In the two-echelon optimal allocation of MSW disposal system, two objectives, cost and environmental impact, should be considered. Considering the uncertainty in the MSW disposal system, this paper constructs a grey fuzzy multi-objective two-echelon MSW allocation model. The model is divided into two sub models and the expected value sorting method is applied to solve the model. The proposed model successfully was applied to a real case in Huangshi, China. The numerical experiments showed RDF technology has advantages on both cost and environmental impact comparing to other disposal technology on disposing MSW.

Keywords: smart city, Two-echelon allocation, MSW, uncertainty.

1. Introduction

With population increasing in the city, the municipal solid waste (MSW) generation grows fast, which causes many issues such as public health, resource utilization and environment. Therefore, MSW management has become an urgent problem in the smart city management. And the MSW disposal is critical to the sustainable development of MSW management.

In the MSW disposal system, the two-echelon optimal allocation model consists of MSW allocation and residues allocation. Sustainable MSW management system requires the incorporation of economic, environmental, and social aspects. Since the social aspect is difficult to measure with data, the economy and environment will be considered into the allocation model.

This paper aims at establishing a dynamic two-echelon MSW optimal allocation model for enhancing MSW management. Generally, the MSW generation is stochastic and unplanned [1]. And it is influenced by many factors, such as people's living habits, consumption pattern and resident income and so on. As a result, the optimal allocation mode is under uncertainty.

The present study area is Huangshi, a city of Hubei Province, China, which consists three administrative districts with a total population of 537,733. Household waste is the major source of MSW. In Huangshi, the MSW generation rate per capita in the study area is about 1.31 kg/day. Currently, MSW incineration is the main technology to dispose the MSW. And there is only one waste incineration power plant located in Huangjinshan. All the MSW of the 3 administrative districts are transported to the waste incineration power plant to be disposed. And in the near future, Refuse Derived Fuel (RDF) disposal technology will be introduced into the MSW disposal system in Huangshi.

The rest of this paper is organized as follows: Section 2 presents a review of the relevant literature and clarifies how we bridge a research gap. A description of the two-echelon allocation of MSW disposal system and the formulation of the mathematical model are presented in Section 3. The computational experiments' results from the case are examined in Section 4. The sensitivity analysis is discussed in Section 5. Finally, Section 6 presents a conclusion, along with suggestions for future research directions.

2. Literature review

The MSW disposal is a complex system. Scholars usually apply mathematical programming models to analyze many MSW management problems, especially the optimal MSW allocation solution. Huang et al. [2] proposed the mathematical programming model, line programming, to obtain the optimal MSW allocation solutions by minimizing the MSW disposal cost. Then, Chang and Wang [3] developed a multi-objective integer programming model based on the model proposed by Huang et al. They took economy and environment into account, and constructed a fuzzy multi-objective integer programming model to seek the optimal allocation solutions of MSW and the capacity expansion solutions of MSW disposal plants. Fiorucci et al. [4] developed a non-linear optimization model to determine the optimal amount and types of MSW transported to landfill, incineration and recycling. Rathi [5] proposed a linear programming model, and took into account the economic and environmental factors in the MSW system to optimize the allocation of MSW in Mumbai.

Many scholars also apply Mixed-Integer Linear Programming (MILP) to find the optimal allocation solution. For example, a MILP model is applied to find the optimal MSW allocation solution in Port Said, Egypt with the goal of minimize transportation cost [6]. Considering the minimize MSW system cost, Dai, Li, and Huang [7] applied the MILP model to obtain the optimal MSW allocation solution in Beijing, China. Chatzouridis and Komilis [8] proposed an MILP model to find the optimal location of MSW transfer stations. Lee et al. [9] proposed a MILP model to find the optimal decision of MSW management system. Tan et al. [10] utilized a MILP model to obtain the optimal MSW disposal facilities capacity and MSW allocation solution by minimizing MSW system cost in Iskandar, Malaysia. The MILP model is developed by Harijani et al. [11]to decide optimal solution of the MSW facilities location and MSW allocation by maximizing the system profit.

There are many uncertain variables in the system, such as MSW generation, MSW recycle rate, transportation and disposal cost, residue conversion rate, etc. Usually, three approaches are applied to represent the uncertain variables: Interval value, Fuzzy and Stochastic programming [12, 13]. Xu et al. [14] optimized the MSW allocation solution by establishing a fuzzy-stochastic programming. Later, an interval-stochastic programming was developed to minimize the MSW system cost by a combination of interval, fuzzy and stochastic programming model [15].

Recently, MSW optimal allocation models have been developed to incorporate multiple disposal technology, especially waste-to-energy (WTE) technology [16,17]. Considering several WTE technologies, Santibañez Aguilar et al. [18] proposed an optimization model to achieve the optimal MSW allocation. Xiong et al. [19] took into account a hybrid WTE system and found that an optimal incorporation of WTE technologies is more economically advantageous. Some scholars took MSW logistics planning and transportation costs into account, applying Location-Routing Problem (LRP) models to find the optimal MSW allocation. Asefi, Lim, Maghrebi, and Shahparvari [20] took minimize MSW transportation and disposal cost as the objectives to optimize the MSW transportation route and allocation solutions. Khattak [21] et al designed a Cross-layer and optimization techniques in wireless multimedia sensor networks for smart cities.

In the previous studies, waste incineration and landfill are considered as the main disposal technology, and refuse derived fuel (RDF) is less involved in the MSW disposal system. Besides, only the MSW allocation is considered in the MSW management system, the residues allocation after MSW disposal is neglected.

In this paper, considering the dynamic and uncertainty of the MSW generation and multiple MSW disposal technologies in the MSW disposal system, a dynamic two-echelon MSW optimal allocation model under uncertainty is established to minimize the economic cost and environmental impact.

3. Methodology

As is mentioned above, many scholars only concern MSW optimal allocation in MSW disposal system. In fact, MSW residues produced during MSW disposal process need to be allocated too. So, in this section, MSW and residues allocation are all considered into the MSW disposal system. Since the MSW generation is uncertain and dynamic, in this paper, the uncertain data or the missing data will be described by grey number and fuzzy number [22, 23], and three periods will be considered. So, a dynamic two-echelon MSW optimal allocation model under uncertainty will be established.

3.1. Problem description

There is two-echelon allocation in the MSW disposal system. The first echelon allocation is the MSW allocation from the MSW transfer stations to the MSW disposal plants. After simple compression and compaction in the MSW transfer stations, the MSW is transported from the transfer station to the each MSW disposal plant (incineration plant, composting plant, RDF plant and sanitary landfill).

The second echelon allocation is the residues allocation. The MSW residues include MSW incineration residues and RDF. During the MSW incineration process, residues, fly ash and bottom ash, will be produced. In the RDF plant, MSW can be converted into RDF. So, residues and RDF need to be redistributed. There are two main disposal methods of residues: landfill and co-disposal in cement kiln. RDF, an alternative fuel for cement plant, can be transported to cement plant to dispose. In addition, RDF can be sold on the market. In summary, the two-echelon allocation of MSW disposal system is shown in Figure.1.

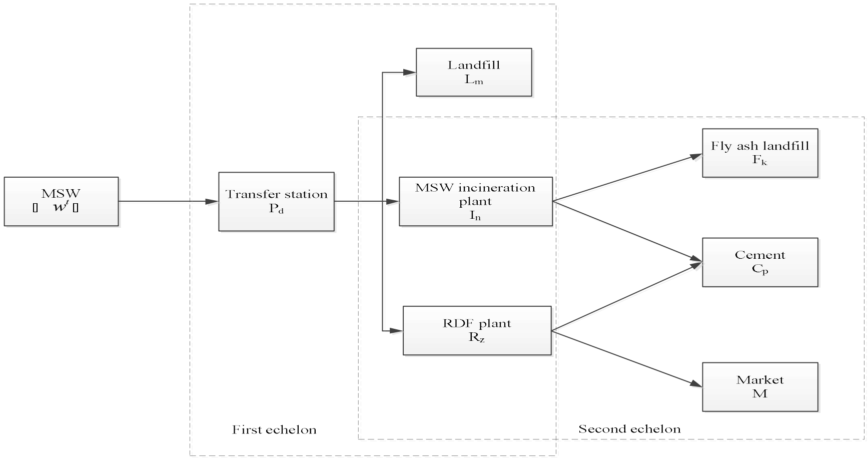


Fig. 1. The two-echelon allocation diagram of MSW disposal system

3.2. Model description

The model built in this paper is based on the following assumptions:

(1) All MSW disposal plants and residues plants applying the same technology have the same disposal efficiency;

(2) In the MSW disposal plant, if the MSW received exceeds its maximum capacity, the capacity expansion will be considered;

(3) The operation cost of each enterprise is only considered;

(4) The transportation cost is only related to the transportation distance.

The parameters involved in the model are as follows:

w^t : The MSW quantity in t period

λ : Converted ratio after pretreatment in transfer station

P_d^t : There are d transfer stations to collect and pre-treat the MSW in period t, $d = 1,...,D$

R_z^t : There are z RDF plants to dispose the MSW in period t, $z = 1,...,Z$

I_n^t : There are n incineration plants to dispose the MSW in period t, $n = 1,...,N$

L_m^t : There are m landfills to dispose the MSW in period t, $m = 1,...,M$

C_p^t : There are p cement plants to dispose the RDF and fly ash in period t, $p = 1,...,P$

F_g^t : There are g fly ash landfills to dispose fly ash in period t, $g = 1,\cdots,G$

$Q_{p_d,I}^t$: The amount of MSW from transfer station P_d^t to incineration plant I_n^t in period t (thousand ton)

Q_{P_d,R_z}^t : The amount of MSW from transfer station P_d^t to RDF plant R_z^t in period t (thousand ton)

Q_{P_d,L_m}^t : The amount of MSW from transfer station P_d^t to landfill L_m^t in period t (thousand ton)

σ : The RDF conversion rate

Q_{R_z,C_q}^t : The amount of MSW from RDF plant R_z^t to cement plant C_p^t in period t (thousand ton)

$Q_{R_z,M}^t$: The amount of MSW from RDF plant R_z^t to market in period t (thousand ton)

ε : Electric generated by incinerating 1 ton MSW (kw*h)

μ : The fly ash conversion rate

Q_{I_n,F_g}^t : The amount of fly ash from incineration plant I_n^t to fly ash landfill F_g^t in period t (thousand ton)

Q_{I_n,C_p}^t : The amount of fly ash from incineration plant I_n^t to cement plant C_p^t in period t (thousand ton)

$Q_{a,b}^t$: The amount of MSW from a to b in period t (thousand ton)

$D_{a,b}$: The distance from a to b

C_T^t : Transport cost in period t

C_m^t : The operation cost of each enterprises in period t

$C_{R_z}^t$: The operation cost of RDF plant in period t

$C_{I_n}^t$: The operation cost of incineration plant in period t

$C_{L_m}^t$: The operation cost of landfill in period t

$C_{C_p}^t$: The operation cost of disposing MSW in cement plant in period t

$C_{CF_p}^t$: The operation cost of disposing fly ash in cement plant in period t

$C_{F_g}^t$: The operation cost of disposing fly ash in fly ash landfill in period t

P_{RDF}^{t} : The price of RDF in period t

S_{RDF}^{t} : The subsidy of disposing MSW in RDF plant in period t

B_{RDF}^{t} : The revenue of RDF plant in period t

S_{I}^{t} : The subsidy of disposing MSW in incineration plant in period t

E_{n}^{t} : The electricity on grid power from MSW incineration plant n in period t

P_{I}^{t} : The price of the electricity on grid power from MSW incineration plant in period t

ν : The substitution rate of RDF for fossil fuel

P_{C}^{t} : The price of fossil fuel

B_{C}^{t} : The revenue of cement plant in period t

$M_{I_n,a}$: The minimum MSW disposal requirement of incineration plant

$M_{I_n,b}$: The maximum MSW disposal capacity of incineration plant

α^{t} : Binary variable, if the MSW incineration plant expands in t period, $\alpha^{t} = 1$, otherwise 0

M_{I_nE} : The expanded capacity of incineration plant

$M_{R_z,a}$: The minimum MSW disposal requirement of RDF plant

$M_{R_z,b}$: The maximum MSW disposal capacity of RDF plant

β^{t} : Binary variable, if the MSW incineration plant expands in t period, $\beta^{t} = 1$, otherwise 0

M_{R_zE} : The expanded capacity of RDF plant

C_{ER_z} : The unit expansion cost of RDF plant

C_{EI_n} : The unit expansion cost of incineration plant

C_{E}^{t} : The total expansion cost of all the plants

M_{L_m} : The maximum MSW disposal capacity of landfill

M_{C_p} : The maximum RDF disposal capacity of cement

M_{F_g} : The maximum fly ash disposal capacity of fly ash landfill

Z_{1}^{t} : The total cost of MSW disposal system

GWP_{L} : The greenhouse gas emission of disposing MSW in landfill

GWP_{I} : The greenhouse gas emission of disposing MSW in incineration plant

GWP_{R} : The greenhouse gas emission of disposing MSW in RDF plant

GWP_{FL} : The greenhouse gas emission of disposing fly ash in fly ash landfill

GWP_{FC} : The greenhouse gas emission of disposing fly ash in cement

GWP_{C} : The greenhouse gas emission of disposing RDF in cement

GWP_{TC} : The greenhouse gas emission of transportation vehicle fleet

Z_{2}^{t} : The total greenhouse gas emission of MSW disposal system

W_1: The cost weight of MSW disposal system

W_2: The greenhouse gas emission weight of MSW disposal system

Z^t: The comprehensive evaluation value (GEV) of MSW disposal system

3.3. Mathematical model

The MSW disposal system mainly considers two objectives (cost and environment). The objective function is as follows:

$$\min Z^t = W_1 Z_1^t + W_2 Z_2^t \tag{1}$$

Where Z^t represents total comprehensive evaluation index, Z_1^t represents total system cost, and Z_2^t represents total environmental impact.

The total system cost Z_1^t mainly consists of four parts: transportation cost, operation cost, expansion cost and economic revenue. The details are as follows:

(1) Transportation cost

Let X be the set of all transportation routes, and $(a,b) \in X$ denotes the route from a to b. $C_{a,b}^t$ represents the transportation cost per kilometer from a to b in period t, and $D_{a,b}$ represents the distance from a to place b. Then the total transportation cost in period t is:

$$C_T^t = \sum_{(a,b)\in X} C_{a,b}^t \square D_{a,b} \tag{2}$$

(2) Operation cost

Suppose $C_{R_z}^t$, $C_{I_n}^t$, $C_{L_m}^t$, $C_{C_p}^t$, $C_{CF_p}^t$, $C_{F_g}^t$ represent the MSW operating cost of RDF plant, MSW incineration plant, landfill, and the residues operating cost of cement plant and fly ash, fly ash landfill respectively in period t. Then the total operating cost in period t is

$$\begin{aligned} C_m^t = {} & \sum_{d=1}^{D}\sum_{z=1}^{Z} Q_{P_d,R_z}^t C_{R_z}^t + \sum_{d=1}^{D}\sum_{n=1}^{N} Q_{P_d,I_n}^t C_{I_n}^t + \sum_{d=1}^{D}\sum_{m=1}^{M} Q_{P_d,L_m}^t C_{L_m}^t \\ & + \sum_{p=1}^{P}\Big(\sum_{z=1}^{Z} Q_{R_z,C_p}^t C_{C_p}^t + \sum_{n=1}^{N} Q_{I_n,C_p}^t C_{CF_p}^t\Big) + \sum_{n=1}^{N}\sum_{g=1}^{G} Q_{I_n,F_g}^t C_{F_g}^t \end{aligned} \tag{3}$$

(3) The expansion cost

When the MSW transported to the disposal plants exceeds their disposal capacity, the expansion decision will be considered. This paper only considers the expansion of MSW incineration plant and RDF plant. The total expansion cost in period t is

$$C_E^t = \alpha^t \sum_{n=1}^{N} C_{EI_n}^t + \beta^t \sum_{z=1}^{Z} C_{ER_z}^t \tag{4}$$

(4) The RDF plant revenue

The revenue of RDF plant in period t mainly consists of two parts: the sales revenue in the market and the subsidy revenue. The revenue of RDF plant can be expressed as follows:

$$B_{RDF}^{t}=\sum_{z=1}^{Z}P_{RDF}^{t}Q_{R_z,M}^{t}+S_{RDF}^{t}\sum_{d=1}^{D}\sum_{z=1}^{Z}Q_{P_d,R_z}^{t} \qquad \textbf{(5)}$$

(5) The incineration plant revenue

It is assumed that the electricity price on grid converted from MSW is 0.65 yuan per kilowatt hour. In addition, the MSW incineration plant can also obtain the subsidy S_n^t for disposing MSW. The ε kW electricity can be obtained by disposing 1 ton MSW. Then the incineration plant revenue in period t is

$$B_E^t=\sum_{n=1}^{N}P_I^tE_n^t+\sum_{d=1}^{D}\sum_{n=1}^{N}Q_{P_d,I_n}^{t}S_n^t \qquad \textbf{(6)}$$

where $E_n^t=\sum_{d=1}^{D}\sum_{n=1}^{N}Q_{P_d,I_n}^{t}\varepsilon$.

(6) The cement plant revenue

RDF can be disposed in cement plant, which can replace fossil fuel. So, disposing RDF can be regarded as the revenue of cement plant. Then the revenue of cement plant in period t is

$$B_C^t=P_C^t\sum_{p=1}^{P}\sum_{z=1}^{Z}\nu Q_{R_z,C_p}^{t} \qquad \textbf{(7)}$$

Above all, the objective function of total cost of MSW disposal system in period t is

$$\min Z_1^t=C_T^t+C_m^t+C_E^t-B_{RDF}^t-B_E^t-B_C^t \qquad \textbf{(8)}$$

(7) Constraints

Capacity constraint of incineration plant: The amount of MSW transported to the incineration plant in period t should be between the minimum and maximum disposal capacity and expansion capacity. Then the constrain is:

$$M_{I_n,a}\le\sum_{n=1}^{N}\sum_{d=1}^{D}Q_{P_d,I_n}^{t}\le M_{I_n,b}+\alpha^tM_{I_nE} \qquad \textbf{(9)}$$

Capacity constraint of RDF plant: The amount of MSW transported to the RDF plant in period t should be between its minimum and maximum disposal capacity and expansion capacity. Then the constrain is:

$$M_{R_z,a}\le\sum_{d=1}^{D}\sum_{z=1}^{Z}Q_{P_d,R_z}^{t}\le M_{R_z,b}+\beta^tM_{R_zE} \qquad \textbf{(10)}$$

Capacity constraint of landfill: The amount of MSW transported to the landfill should be no more than the landfill capacity. Then the constrain is:

$$\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{m=1}^{M}Q_{P_d,L_m}^{t}\le M_{L_m} \qquad \textbf{(11)}$$

Capacity constraint of cement plant: The RDF and fly ash transported to the cement plant in period t should be no more than the cement plant capacity respectively. Then the constrain is:

$$\sum_{z=1}^{Z}\sum_{p=1}^{P}Q_{R_z,C_p}^{t}\le M_{C_p} \qquad \textbf{(12)}$$

$$\sum_{n=1}^{N}\sum_{p=1}^{P}Q_{I_n,C_p}^{t}\le M_{C_{p,F}} \qquad \textbf{(13)}$$

Capacity constraint of fly ash landfill: The amount of fly ash transported to the landfill should be no more than the landfill capacity. Then the constrain is:

$$\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{g=1}^{G} Q^{t}_{I_n,F_g} \le M_{F_g} \qquad \textbf{(14)}$$

Material balance constraint: The material balance constraints of MSW transfer station, incineration plant, RDF plant and landfill are as follows.

$$\sum_{z=1}^{Z} Q^{t}_{P_d,R_z} + \sum_{n=1}^{N} Q^{t}_{P_d,I_n} + \sum_{m=1}^{M} Q^{t}_{P_d,L_m} = \lambda w^{t} \quad d = 1,\ldots,D \qquad \textbf{(15)}$$

$$\sum_{p=1}^{P} Q^{t}_{R_z,C_p} + Q^{t}_{R_z,M} = \sigma \sum_{d=1}^{D} Q^{t}_{P_d,R_z} \quad z = 1,\ldots,Z \qquad \textbf{(16)}$$

$$\sum_{g=1}^{G} Q^{t}_{I_n,F_g} + \sum_{p=1}^{P} Q^{t}_{I_n,C_p} = \mu \sum_{d=1}^{D} Q^{t}_{P_d,I_n} \quad n = 1,\ldots,N \qquad \textbf{(17)}$$

In this paper, all decision variables are nonnegative.

The environmental impact is measured by greenhouse gas emissions (GHG). The environmental impact Z_2^t in period t is:

$$\begin{aligned}\min Z_2^t = (&\sum_{d=1}^{D}\sum_{m=1}^{M} Q^{t}_{P_d,L_m} \square d_{P_d,L_m} + \sum_{d=1}^{D}\sum_{n=1}^{N} Q^{t}_{P_d,I_n} \square d_{P_d,I_n} + \sum_{d=1}^{D}\sum_{z=1}^{Z} Q^{t}_{P_d,R_z} \square d_{P_d,R_z} \\ &+ \sum_{z=1}^{Z}\sum_{p=1}^{P} Q^{t}_{R_z,C_p} \square d_{R_z,C_p} + \sum_{n=1}^{N}\sum_{g=1}^{G} Q^{t}_{I_n,F_g} \square d_{I_n,F_g} + \sum_{n=1}^{N}\sum_{p=1}^{P} Q^{t}_{I_n,C_p} \square d_{I_n,C_p}) \square GWP_{TC} \\ &+ \sum_{d=1}^{D}\sum_{m=1}^{M} Q^{t}_{P_d,L_m} \square GWP_L + \sum_{d=1}^{D}\sum_{n=1}^{N} Q^{t}_{P_d,I_n} \square GWP_I + \sum_{d=1}^{D}\sum_{z=1}^{Z} Q^{t}_{P_d,R_z} \square GWP_R \\ &+ \sum_{z=1}^{Z}\sum_{p=1}^{P} Q^{t}_{R_z,C_p} \square GWP_C + \sum_{g=1}^{G} Q^{t}_{I_n,F_g} \square GWP_{FL} + \sum_{n=1}^{N}\sum_{p=1}^{P} Q^{t}_{I_n,C_p} \square GWP_{FC}\end{aligned} \qquad \textbf{(18)}$$

Where the first three terms represent the greenhouse gas emissions of MSW transportation from the waste transfer station to the landfill, incineration plant and RDF Plant respectively; the fourth term represents the greenhouse gas emissions of the RDF transportation from RDF plant to the cement plant, the fifth and sixth terms represent the greenhouse gas emissions of the fly ash transportation from incineration plant to the fly ash landfill and the cement plant respectively; the seventh, eighth and ninth terms represent the greenhouse gas emissions of disposing MSW in landfill, incineration plant and RDF plant respectively; Since RDF can replace the fossil fuel, disposing the RDF can reduce the greenhouse gas emission the tenth term represents the greenhouse gas emissions of disposing RDF in cement plant; the eleventh and twelfth terms represent the greenhouse gas emissions of disposing the fly ash in fly ash landfill and cement plant respectively.

3.4. The uncertain multi-objective two-echelon optimal MSW allocation model

Since there are many uncertain factors in the MSW disposal system, considering the uncertain factors, the above model is transformed into gray fuzzy multi-objective programming model. The grey fuzzy minimum system cost function is:

$$\min \otimes \tilde{Z}_1^t = \otimes \tilde{C}_T^t + \otimes \tilde{C}_m^t + \otimes \tilde{C}_E^t - \otimes \tilde{B}_{RDF}^t - \otimes \tilde{B}_E^t - \otimes \tilde{B}_C^t \qquad \textbf{(19)}$$

Where $\otimes \tilde{C}_T^t$ represents the fuzzy grey value of transportation cost in period t, $\otimes C_m^t$ represents the operation cost in period t, $\otimes C_E^t$ represents the expansion cost in period t, $\otimes \tilde{B}_{RDF}^t$ represents the RDF plant revenue in period t, $\otimes \tilde{B}_E^t$ represents the electric revenue in period t and $\otimes \tilde{B}_C^t$ represents the cement plant revenue in period t.

$\tilde{GWP}_{TC}$, $\tilde{GWP}_L$, $\tilde{GWP}_I$, $\tilde{GWP}_R$, $\tilde{GWP}_C$, $\tilde{GWP}_{FL}$, $\tilde{GWP}_{FC}$ represent the fuzzy greenhouse gas emissions value of transportation, landfill, incineration plant, RDF plant, disposing RDF in cement plant , fly ash landfill and disposing fly ash in cement plant respectively. The fuzzy function of environmental impact in period t is

$$\begin{aligned}\min \tilde{Z}_2^t = (&\sum_{d=1}^{D}\sum_{m=1}^{M} Q_{P_d,L_m}^t \square d_{P_d,L_m} + \sum_{d=1}^{D}\sum_{n=1}^{N} Q_{P_d,I_n}^t \square d_{P_d,I_n} + \sum_{d=1}^{D}\sum_{z=1}^{Z} Q_{P_d,R_z}^t \square d_{P_d,R_z} \\ &+ \sum_{z=1}^{Z}\sum_{p=1}^{P} Q_{R_z,C_p}^t \square d_{R_z,C_p} + \sum_{n=1}^{N}\sum_{g=1}^{G} Q_{I_n,F_g}^t \square d_{I_n,F_g} + \sum_{n=1}^{N}\sum_{p=1}^{P} Q_{I_n,C_p}^t \square d_{I_n,C_p}) \square \tilde{GWP}_{TC} \\ &+ \sum_{d=1}^{D}\sum_{m=1}^{M} Q_{P_d,L_m}^t \square \tilde{GWP}_L + \sum_{d=1}^{D}\sum_{m=1}^{M} Q_{P_d,I_n}^t \square \tilde{GWP}_I + \sum_{d=1}^{D}\sum_{z=1}^{Z} Q_{P_d,R_z}^t \square \tilde{GWP}_R \\ &+ \sum_{z=1}^{Z}\sum_{p=1}^{P} Q_{R_z,C_p}^t \square \tilde{GWP}_C + \sum_{g=1}^{G} Q_{I_n,F_g}^t \square \tilde{GWP}_{FL} + \sum_{n=1}^{N}\sum_{p=1}^{P} Q_{I_n,C_p}^t \square \tilde{GWP}_{FC}\end{aligned} \qquad \textbf{(20)}$$

The grey fuzzy comprehensive evaluation function of two-echelon allocation model is

$$\min \otimes \tilde{Z}^t = W_1 \otimes \tilde{Z}_1^t + W_2 \otimes \tilde{Z}_2^t \qquad \textbf{(21)}$$

The constraints of grey fuzzy multi-objective two-echelon allocation model are

$$M_{I_n,a} \le \sum_{n=1}^{N}\sum_{d=1}^{D} \otimes Q_{P_d,I_n}^t \le M_{I_n,b} + \alpha^t M_{I_nE} \qquad \textbf{(22)}$$

$$M_{R_z,a} \le \sum_{d=1}^{D}\sum_{z=1}^{Z} \otimes Q_{P_d,R_z}^t \le M_{R_z,b} + \beta^t M_{R_zE} \qquad \textbf{(23)}$$

$$\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{m=1}^{M} \otimes Q_{P_d,L_m}^t \le M_{L_m} \qquad \textbf{(24)}$$

$$\sum_{z=1}^{Z}\sum_{p=1}^{P} \otimes Q_{R_z,C_p}^t \le M_{C_p} \qquad \textbf{(25)}$$

$$\sum_{n=1}^{N}\sum_{p=1}^{P} \otimes Q_{I_n,C_p}^t \le M_{C_p,F} \qquad \textbf{(26)}$$

$$\sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{g=1}^{G} \otimes Q_{I_n,F_g}^t \le M_{F_g} \qquad \textbf{(27)}$$

$$\sum_{z=1}^{Z} \otimes Q_{P_d,R_z}^t + \sum_{n=1}^{N} \otimes Q_{P_d,I_n}^t + \sum_{m=1}^{M} \otimes Q_{P_d,L_m}^t = \lambda \otimes w^t \quad d = 1,\ldots,D \qquad \textbf{(28)}$$

$$\sum_{p=1}^{P} \otimes Q_{R_z,C_p}^t + \otimes Q_{R_z,M}^t = \sigma \sum_{d=1}^{D} \otimes Q_{P_d,R_z}^t \quad z = 1,\ldots,Z \qquad \textbf{(29)}$$

$$\sum_{g=1}^{G}\otimes Q^{t}_{I_n,F_g} + \sum_{p=1}^{P}\otimes Q^{t}_{I_n,C_p} = \mu\sum_{d=1}^{D}\otimes Q^{t}_{P_d,I_n} \quad n=1,...,N \qquad \textbf{(30)}$$

4. Case study

4.1. Experimental Design and Environment

Huangshi is located in the southeast of Hubei Province, China. Huangshi consists of three administrative districts (Huangshigang, Xisaishan, Xialu). This section first forecasts the MSW generation per capita in the three administrative districts, and then combines the population data in each administrative region with the MSW generation per capita prediction data to obtain the MSW generation allocation in Huangshi. There are 76 communities in the three administrative districts, where 30 communities are located in Huangshigang, 19 communities are located in Xisaishan and 27 communities are located in Xialu. In this paper, the community is regarded as collection point.

This paper will study the two-echelon optimal allocation in MSW disposal system in three periods. The system includes two parts: (1) the allocation of MSW among waste disposal plants; (2) the allocation of residues between residues disposal plants and the market. There are 18 waste transfer stations (TS) in Huangshi, the longitude and latitude coordinates of 18 waste transfer stations and the amount of MSW are shown in Table 1.

Table 1. The received waste amount of waste transfer station (thousand ton)

No	T1		T2		T3	
	Low	Up	Low	Up	Low	Up
1	15.9	16.2	16.4	16.6	16.4	16.6
2	10.1	10.2	10.3	10.4	10.3	10.4
3	16.9	17.1	17.3	17.5	17.3	17.5
4	46.1	46.6	47	47.5	47	47.5
5	6.6	6.8	7	7.2	7	7.2
6	14.5	14.9	15.3	15.8	15.4	15.8
7	10.7	11.1	11.4	11.7	11.4	11.7
8	17.4	17.9	18.4	19.0	18.4	19.0
9	15.9	16.4	16.9	17.4	16.9	17.4
10	15.6	16.1	16.6	17.1	16.6	17.1
11	11.5	11.8	12	12.3	12.0	12.3
12	11.2	11.4	11.7	11.9	11.7	11.9
13	33.3	34	34.7	35.5	34.7	35.5
14	27.3	27.9	28.5	29.1	28.5	29.1
15	11.7	11.8	11.9	12	11.9	12
16	12.2	12.3	12.4	12.5	12.4	12.5
17	3.1	3.1	3.2	3.3	3.2	3.3
18	28.4	29	29.6	30.3	29.6	30.3

Based on the MSW management system in Huangshi, this paper will consider three MSW disposal technologies, namely landfill, MSW incineration and RDF. The disposal capacity of each disposal plant is shown in Table 2. The transportation distance between each disposal plant is shown in Table 3. The operating cost of each plant in three periods is shown in Table 4. The subsidy for MSW disposal is shown in Table 5.

Table 2. The disposal capacity of each disposal plant

Disposal plant	Disposal object	Disposal capacity	Unit
Incineration plant （$M_{I_n,a}$，$M_{I_n,b}$）	MSW	[63.4，190.4]	Thousand ton/year
RDF plant （$M_{R_q,a}$，$M_{R_q,b}$）	MSW	[47.4，158.7]	Thousand ton/year
Cement plant	RDF（M_{C_p}）	≤328.5	Thousand ton/year
	Fly ash （$M_{C_p,F}$）	[15.8，19]	Thousand ton/year
Fly ash landfill （M_{F_g}）	Fly ash	≤120.45	Thousand ton/year

Table 3. The transportation distances between each disposal plant Unit:km

To From	RDF plant	Incineration plant	Fly ash landfill	Cement plant
Transfer station 1	17.2	13.1	9.5	-
Transfer station 2	18.7	14.2	7.5	-
Transfer station 3	20.0	15.3	5.8	-
Transfer station 4	14.3	10.6	12.2	-
Transfer station 5	5.1	7	22.5	-
Transfer station 6	4.5	8.3	24.8	-
Transfer station 7	6.5	11.3	27.9	-
Transfer station 8	4.4	8.7	25.5	-
Transfer station 9	7.5	6.8	19.5	-
Transfer station 10	10.0	7.4	16.5	-
Transfer station 11	16.3	13.2	12.5	-
Transfer station 12	14.0	11.6	15.0	-
Transfer station 13	15.7	13.5	15.2	-
Transfer station 14	14.5	12.9	16.9	
Transfer station 15	15.2	11.3	11.3	
Transfer station 16	17.1	12.7	9.0	
Transfer station 17	7.4	8.8	22.3	
Transfer station 18	16.0	12.3	11.3	
RDF plant	-	-	-	3.94
Incineration plant	-	-	19.6	9.4

Table 4. The operating costs of each plant in three periods

Disposal plant	Operation cost					
	T1		T2		T3	
	Low	Up	Low	Up	Low	Up
RDF plant (yuan/ton)	135.0	137.7	140.4	143.2	146.0	148.9
Incineration plant (yuan/ton)	90	91.9	93.9	96.1	98.4	100.8
Landfill (yuan/ton)	89.0	90.9	92.9	95.0	97.3	99.7
Fly ash landfill (yuan/ton)	446.1	455	463.9	473.2	482.5	492.1
Cement plant (yuan/ton)	1500	1531.5	1565.2	1601.2	1639.6	1680.6

Table 5. The subsidy for MSW disposal

Disposal technology	Subsidy form	Subsidy standard
Incineration	Disposal （S_I^t）	150 yuan/ton
Incineration	Electricity price on grid （P_I^t）	0.65 yuan/kwh
RDF	Disposal （S_{RDF}^t）	80 yuan/ton

There is some fuzzy data in MSW disposal system, such as expansion capacity, expansion cost, conversion rate, greenhouse gas emissions and so on. For the fuzzy data, the lower limit of the fuzzy data is based on 90% of the original data, the upper limit of the fuzzy data is based on 120% of the original data, and the de-fuzzy data is obtained according to the expected value sorting method. The de-fuzzy data of related parameters is shown in Table 6. The de-fuzzy greenhouse gas emissions of each disposal plant are shown in Table 7.

Table 6. The de-fuzzy data of related parameters

Disposal plant	Parameter	Value
Transfer station	The rest rate after pretreatment（λ）	96.58%
Incineration plant	The electricity on grid after disposing MSW （$E^t_n - E^t_{c,n}$）	325.33 kwh/ton
	Production rate of fly ash（μ）	3.05%
	Expansion capacity（M_{I_nE}）	100 thousand ton
	Expansion cost（C_{EI_n}）	50.83 yuan/ton
RDF plant	RDF conversion rate（σ）	50.83%
	Expansion capacity（M_{I_nE}）	100 thousand ton
	Expansion cost（C_{ER_z}）	40.67 yuan/ton
	Price（P^t_{RDF}）	203.33yuan/ton
Cement plant	Substitution rate of RDF for coal（ν）	50.83%
	Coal rice（P^t_C）	610 yuan/ton

Table 7. The de-fuzzy data of greenhouse gas emissions of each disposal plant

Emission source	GWP（t CO_2 eqv. t^{-1} MSW）
Landfill（GWP_L）	2.755
Incineration plant（GWP_I）	0.464
RDF plant（GWP_R）	0.203
Fly ash landfill（GWP_{FL}）	0.015
Disposal fly ash（GWP_{FC}）	0
Disposal RDF（GWP_C）	-1.169

4.2. Results and discussion

The uncertain multi-objective two-echelon optimal MSW allocation model is divided into two sub models, the weight of cost objective function (W_1) is 0.6, the weight of environment objective function (W_2) is 0.4, and then it is solved by matlab2016a.

It can be seen from Table 8-10 that in the next three periods, the annual MSW is mainly allocated to RDF plant and incineration plant. During the MSW allocation process, the transportation distance is fully considered. When the transfer station is close to the RDF plant, the MSW is prior to be allocated to the RDF plant (such as transfer station 8), otherwise, it is prior to be allocated to the incineration plant (such as transfer station 4). Since RDF plant has environmental advantages in MSW disposal, MSW is prior to be allocated to the RDF plant. So the MSW allocated to RDF plant reaches its maximum disposal capacity. The MSW allocated to landfill is zero in the three periods.

In addition, all RDF produced in RDF plant is disposed in cement plants, and there is no market sale for RDF. Due to the high cost of disposing fly ash in cement plant, all fly ashes are transported to fly ash landfill to be disposed. The system revenue increases from T1 to T2, mainly due to the MSW growth. The system revenue decrease from T2 to T3, mainly because the MSW growth slows down and the operating cost of the MSW disposal plants increases. The revenue of MSW disposal plants mainly comes from the subsidy and the MSW recycle income.

It also can be seen that the environmental impact is proportional to the MSW amount. When the MSW increases, the environmental pressure increases. According to the comprehensive evaluation index, the operation effect of the above-mentioned system is decreasing, mainly because the MSW amount can’t meet the demand of all the MSW disposal plant.

5. Sensitivity analysis

In this part, two sensitivity analysis cases will be discussed:

(1) Adjust the weight of cost and environment in the comprehensive evaluation, and compare the difference of the optimal solutions.

(2) Don’t consider the landfill disposal technology, and assume that the MSW generation increases by 30% in T4 period, then the capacity expansion of the MSW disposal plant will be discussed.

5.1. Weight adjustment of cost and environmental

The weight of cost and environment is adjusted to W_1=0.4， W_2=0.6, which means policy makers pay more attention to environmental impact. The result is shown in Table11-13. Comparing the data in Table 8-10 with the data in Table 11-13, the MSW allocation in each transfer station has little change and MSW is still allocated to RDF plant preferentially. The amount of fly ash allocated to cement kiln collaborative disposal technology is still 0. The reason is that although the cement kiln collaborative disposal technology has environmental advantages, due to the high cost of fly ash collaborative disposal, the MSW disposal system still cannot apply this technology. Considering comprehensive evaluation, the overall effect of this case is worse compared with the previous case. The main reason is that the MSW allocated to RDF plant is not sufficient, so the environmental advantage of RDF technology is difficult to present.

5.2. The capacity expansion of the MSW disposal plant

In order to discuss the capacity expansion of the MSW disposal plant, T4 period is added. In this period, it is assumed that the MSW in each transfer station increases by 30%, and the landfill technology is not considered, the operation cost of each MSW disposal plant is the same as that of the T3 period, the weight of cost and environment is 0.6 and 0.4, the result is shown in Table 14. According to the Table 14, it can be found that when the MSW exceeds the maximum disposal capacity of the plant, the capacity expansion of the RDF plant is considered preferentially. Only the minimum MSW disposal requirement of the incineration plant is satisfied, and the rest MSW is allocated to the RDF plant. Comparing the data in T4 period with the data in T3 period of the previous two cases, it is found that the system revenue decreases, but the environmental impact significantly increase. It also can be seen that when the MSW is sufficient, the environmental advantage of RDF technology can be reflected, and the whole system runs better.

Table 8. The result of uncertain multi-objective two-echelon optimal MSW allocation model in period T1 (W_1=0.6，W_2=0.4)

To From	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement
	Lower amount of collected MSW （297.9 thousand ton）					Upper amount of collected MSW（303.8 thousand ton）				
TS 1	0.79	0.75	0	-	-	0.86	0.66	0	-	-
TS 2	0.48	0.50	0	-	-	0.58	0.39	0	-	-
TS 3	0.83	0.81	0	-	-	0.84	0.80	0	-	-
TS 4	1.95	2.50	0	-	-	1.97	2.54	0	-	-
TS 5	0.36	0.28	0	-	-	0.42	0.23	0	-	-
TS 6	0.89	0.51	0	-	-	0.85	0.60	0	-	-
TS 7	0.65	0.39	0	-	-	0.58	0.49	0	-	-
TS 8	1.12	0.56	0	-	-	1.06	0.68	0	-	-
TS 9	0.85	0.68	0	-	-	0.97	0.63	0	-	-
TS 10	0.82	0.69	0	-	-	0.90	0.67	0	-	-
TS 11	0.57	0.54	0	-	-	0.62	0.51	0	-	-
TS 12	0.57	0.51	0	-	-	0.48	0.62	0	-	-
TS 13	1.78	1.44	0	-	-	1.48	1.84	0	-	-
TS 14	1.49	1.15	0	-	-	1.31	1.41	0	-	-
TS 15	0.57	0.56	0	-	-	0.57	0.54	0	-	-
TS 16	0.58	0.59	0	-	-	0.61	0.55	0	-	-
TS 17	0.17	0.13	0	-	-	0.13	0.15	0	-	-
TS 18	1.41	1.33	0	-	-	1.64	1.19	0	-	-
Incineration plant	-	-	-	0.42	0	-	-	-	0.44	0
RDF plant	-	-	-	-	8.07	-	-	-	-	8.07
Revenue			51.77					52.53		
GHG			2.75					5.53		
GEV			-2.00					-0.94		

Table 9. The result of uncertain multi-objective two-echelon optimal MSW allocation model in period T2 (W_1=0.6， W_2=0.4)

To From	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement
	Lower amount of collected MSW （297.9 thousand ton）					Upper amount of collected MSW（303.8 thousand ton）				
TS 1	0.79	0.80	0	-	-	0.26	1.34	0	-	-
TS 2	0.29	0.71	0	-	-	0.31	0.70	0	-	-
TS 3	0.94	0.70	0	-	-	0.36	1.33	0	-	-
TS 4	1.58	2.96	0	-	-	1.19	3.40	0	-	-
TS 5	0.41	0.27	0	-	-	0.59	0.11	0	-	-
TS 6	1.11	0.38	0	-	-	1.49	0.04	0	-	-
TS 7	0.81	0.30	0	-	-	0.97	0.16	0	-	-
TS 8	1.05	0.73	0	-	-	1.71	0.12	0	-	-
TS 9	0.85	0.79	0	-	-	1.25	0.43	0	-	-
TS 10	0.98	0.62	0	-	-	1.02	0.63	0	-	-
TS 11	0.72	0.44	0	-	-	0.52	0.67	0	-	-
TS 12	0.52	0.61	0	-	-	0.56	0.59	0	-	-
TS 13	1.52	1.83	0	-	-	2.02	1.40	0	-	-
TS 14	1.58	1.18	0	-	-	1.84	0.97	0	-	-
TS 15	0.51	0.65	0	-	-	0.41	0.75	0	-	-
TS 16	0.51	0.70	0	-	-	0.24	0.97	0	-	-
TS 17	0.20	0.11	0	-	-	0.24	0.08	0	-	-
TS 18	1.52	1.34	0	-	-	0.88	2.04	0	-	-
Incineration plant	-	-	-	0.46	0	-	-	-	0.48	0
RDF plant	-	-	-	-	8.07	-	-	-	-	8.07
Revenue			53.29					54.04		
GHG			8.23					11.12		
GEV			0.095					1.21		

Table10. The result of uncertain multi-objective two-echelon optimal MSW allocation model in period T3 (W_1=0.6， W_2=0.4)

To From	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement
	Lower amount of collected MSW （297.9 thousand ton）					Upper amount of collected MSW（303.8 thousand ton）				
TS 1	0.80	0.78	0	-	-	0.29	1.31	0	-	-
TS 2	0.47	0.53	0	-	-	0.26	0.75	0	-	-
TS 3	0.86	0.81	0	-	-	0.14	1.55	0	-	-
TS 4	2.42	2.12	0	-	-	1.67	2.92	0	-	-
TS 5	0.35	0.33	0	-	-	0.61	0.09	0	-	-
TS 6	0.78	0.70	0	-	-	1.46	0.07	0	-	-
TS 7	0.56	0.55	0	-	-	1.06	0.07	0	-	-
TS 8	0.94	0.84	0	-	-	1.76	0.07	0	-	-
TS 9	0.85	0.79	0	-	-	1.32	0.36	0	-	-
TS 10	0.83	0.78	0	-	-	0.76	0.89	0	-	-
TS 11	0.59	0.57	0	-	-	0.60	0.59	0	-	-
TS 12	0.56	0.57	0	-	-	0.60	0.55	0	-	-
TS 13	1.72	1.63	0	-	-	1.72	1.70	0	-	-
TS 14	1.37	1.38	0	-	-	1.67	1.14	0	-	-
TS 15	0.57	0.58	0	-	-	0.40	0.76	0	-	-
TS 16	0.55	0.66	0	-	-	0.33	0.88	0	-	-
TS 17	0.16	0.15	0	-	-	0.24	0.08	0	-	-
TS 18	1.50	1.36	0	-	-	0.98	1.94	0	-	-
Incineration plant	-	-	-	0.46	0	-	-	-	0.48	0
RDF plant	-	-	-	-	8.07	-	-	-	-	8.07
Revenue			51.66					52.28		
GHG			8.29					11.08		
GEV			0.22					1.29		

Table11． The result of uncertain multi-objective two-echelon optimal MSW allocation model in period T1 (W_1=0.4， W_2=0.6)

To From	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement
	Lower amount of collected MSW （297.9 thousand ton）					Upper amount of collected MSW（303.8 thousand ton）				
TS 1	0.78	0.76	0	-	-	0.29	1.27	0	-	-
TS 2	0.42	0.56	0	-	-	0.07	0.92	0	-	-
TS 3	0.78	0.86	0	-	-	0.02	1.63	0	-	-
TS 4	2.50	1.95	0	-	-	1.76	2.74	0	-	-
TS 5	0.45	0.19	0	-	-	0.63	0.03	0	-	-
TS 6	0.88	0.52	0	-	-	1.39	0.05	0	-	-
TS 7	0.68	0.36	0	-	-	1.06	0.01	0	-	-
TS 8	0.96	0.72	0	-	-	1.72	0.01	0	-	-
TS 9	0.84	0.70	0	-	-	1.08	0.50	0	-	-
TS 10	0.72	0.79	0	-	-	0.70	0.85	0	-	-
TS 11	0.51	0.60	0	-	-	0.19	0.95	0	-	-
TS 12	0.59	0.49	0	-	-	0.44	0.66	0	-	-
TS 13	1.90	1.32	0	-	-	2.12	1.17	0	-	-
TS 14	1.30	1.34	0	-	-	2.05	0.64	0	-	-
TS 15	0.48	0.65	0	-	-	0.15	0.99	0	-	-
TS 16	0.57	0.60	0	-	-	0.14	1.05	0	-	-
TS 17	0.19	0.10	0	-	-	0.27	0.03	0	-	-
TS 18	1.33	1.41	0	-	-	1.79	1.01	0	-	-
Incineration plant	-	-	-	0.42	0	-	-	-	0.44	0
RDF plant	-	-	-	-	8.07	-	-	-	-	8.07
Revenue			51.75					52.51		
GHG			2.73					5.49		
GEV			-0.43					1.19		

Table12． The result of uncertain multi-objective two-echelon optimal MSW allocation model in period T2 (W_1=0.4， W_2=0.6)

To From	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement
	Lower amount of collected MSW（297.9 thousand ton）					Upper amount of collected MSW（303.8 thousand ton）				
TS 1	0.85	0.73	0	-	-	0.44	1.16	0	-	-
TS 2	0.50	0.51	0	-	-	0.53	0.48	0	-	-
TS 3	0.86	0.81	0	-	-	0.82	0.87	0	-	-
TS 4	2.44	2.10	0	-	-	2.34	2.25	0	-	-
TS 5	0.34	0.34	0	-	-	0.15	0.55	0	-	-
TS 6	0.76	0.72	0	-	-	1.03	0.50	0	-	-
TS 7	0.58	0.52	0	-	-	0.61	0.52	0	-	-
TS 8	0.90	0.88	0	-	-	1.21	0.63	0	-	-
TS 9	0.84	0.79	0	-	-	0.97	0.71	0	-	-
TS 10	0.84	0.76	0	-	-	0.97	0.68	0	-	-
TS 11	0.60	0.56	0	-	-	0.44	0.75	0	-	-
TS 12	0.57	0.56	0	-	-	0.36	0.79	0	-	-
TS 13	1.72	1.63	0	-	-	1.85	1.57	0	-	-
TS 14	1.28	1.47	0	-	-	1.64	1.17	0	-	-
TS 15	0.60	0.56	0	-	-	0.46	0.70	0	-	-
TS 16	0.59	0.61	0	-	-	0.17	1.04	0	-	-
TS 17	0.15	0.16	0	-	-	0.16	0.17	0	-	-
TS 18	1.44	1.42	0	-	-	1.72	1.20	0	-	-
Incineration plant	-	-	-	0.46	0	-	-	-	0.48	0
RDF plant	-	-	-	-	8.07	-	-	-	-	8.07
Revenue			53.25					54.02		
GHG			8.12					11.05		
GEV			2.74					4.47		

Table13. The result of uncertain multi-objective two-echelon optimal MSW allocation model in period T3 (W_1=0.4，W_2=0.6)

To From	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement
	Lower amount of collected MSW （297.9 thousand ton）					Upper amount of collected MSW（303.8 thousand ton）				
TS 1	0.87	0.71	0	-	-	0.32	1.28	0	-	-
TS 2	0.50	0.51	0	-	-	0.02	0.99	0	-	-
TS 3	0.94	0.73	0	-	-	0.10	1.59	0	-	-
TS 4	1.24	3.30	0	-	-	2.18	2.41	0	-	-
TS 5	0.38	0.30	0	-	-	0.67	0.03	0	-	-
TS 6	0.95	0.53	0	-	-	1.48	0.05	0	-	-
TS 7	0.69	0.41	0	-	-	1.11	0.02	0	-	-
TS 8	1.23	0.55	0	-	-	1.82	0.01	0	-	-
TS 9	0.98	0.65	0	-	-	1.13	0.55	0	-	-
TS 10	0.93	0.67	0	-	-	0.55	1.10	0	-	-
TS 11	0.60	0.56	0	-	-	0.31	0.88	0	-	-
TS 12	0.61	0.52	0	-	-	0.41	0.74	0	-	-
TS 13	1.42	1.93	0	-	-	2.29	1.13	0	-	-
TS 14	1.28	1.47	0	-	-	1.65	1.16	0	-	-
TS 15	0.62	0.53	0	-	-	0.24	0.92	0	-	-
TS 16	0.63	0.57	0	-	-	0.13	1.08	0	-	-
TS 17	0.17	0.14	0	-	-	0.29	0.03	0	-	-
TS 18	1.82	1.04	0	-	-	1.16	1.76	0	-	-
Incineration plant	-	-	-	0.46	0	-	-	-	0.48	0
RDF plant	-	-	-	-	8.07	-	-	-	-	8.07
Revenue			51.56					52.22		
GHG			8.28					11.04		
GEV			2.91					4.54		

Table14． The result of uncertain multi-objective two-echelon optimal MSW allocation model in period T4 (W_1 =0.6， W_2 =0.4)

To / From	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement	RDF plant	Incineration plant	Landfill	Fly ash landfill	Cement
	Lower amount of collected MSW （297.9 thousand ton）					Upper amount of collected MSW（303.8 thousand ton）				
TS 1	1.74	0.32	0	-	-	1.76	0.33	0	-	-
TS 2	0.86	0.43	0	-	-	1.01	0.30	0	-	-
TS 3	1.86	0.31	0	-	-	1.86	0.33	0	-	-
TS 4	5.42	0.49	0	-	-	5.32	0.65	0	-	-
TS 5	0.70	0.18	0	-	-	0.65	0.26	0	-	-
TS 6	1.56	0.36	0	-	-	1.65	0.33	0	-	-
TS 7	1.07	0.36	0	-	-	1.15	0.32	0	-	-
TS 8	2.06	0.26	0	-	-	2.08	0.30	0	-	-
TS 9	1.80	0.32	0	-	-	1.84	0.34	0	-	-
TS 10	1.76	0.32	0	-	-	1.63	0.51	0	-	-
TS 11	1.15	0.35	0	-	-	1.19	0.35	0	-	-
TS 12	1.20	0.27	0	-	-	1.17	0.33	0	-	-
TS 13	3.72	0.64	0	-	-	4.00	0.45	0	-	-
TS 14	3.05	0.53	0	-	-	3.30	0.35	0	-	-
TS 15	1.20	0.30	0	-	-	1.18	0.33	0	-	-
TS 16	1.29	0.27	0	-	-	1.23	0.34	0	-	-
TS 17	0.25	0.16	0	-	-	0.25	0.17	0	-	-
TS 18	3.26	0.46	0	-	-	3.43	0.37	0	-	-
Incineration plant	-	-	-	0.19	0	-	-	-	0.19	0
RDF plant	-	-	-	-	17.25	-	-	-	-	17.25
Revenue			45.1					44.6		
GHG			-103					-101		
GEV			-63.671					-62.654		

6. Conclusion and future work

In the MSW disposal system, the MSW is allocated among the disposal plants firstly, and then the residues (incineration residues and RDF) are allocated between the residue disposal plants and market. So there is a two-echelon allocation in the MSW disposal system. In the two-echelon optimal allocation of MSW system, two objectives, cost and environmental impact, should be considered. Considering the uncertainty and dynamic in the MSW disposal system, this paper constructs a grey fuzzy multi-objective two-echelon MSW allocation model. The model is divided into two sub models firstly, and then the expected value sorting method is applied to solve the models. According to the result, the MSW is prior to be allocated to RDF plant and incineration plant. The MSW allocated to landfill is zero in the three periods, because the landfill will cause more environment pollution. Two sensitivity analysis cases are studied, and it is found that RDF technology has greater environmental advantage in all disposal technology. When the MSW is sufficient, the environmental advantage of RDF technology can be reflected, and the whole system runs better.

In the future work, stochastic MSW generation rates can be considered. Besides that, waste classification can be considered into the MSW disposal system. How to allocate the different waste type among the disposal plants can be an interesting research direction in the future.

Acknowledgment. The study was supported by the Science and Technology Research Project of Hubei Polytechnic University (Grant no. 14xjr08Q). Research Center for Mining and Metallurgy Culture and Social-economic Development in the Middle Reaches of Yangtze River (Grant no. 2018kyy07, 2015kyb06).

References

1. Feng, D., Guihua, Nie., Yi, C.: The municipal solid waste generation distribution prediction system based on FIG–GA-SVR model. Journal of Material Cycles and Waste Management, Vol. 22, No. 5, 1352-1369. (2020).
2. Huang, G., Baetz, B. W., Patry, G. G.: A grey linear programming approach for municipal solid waste management planning under uncertainty. Civil Engineering Systems, Vol. 9, No. 4, 319-335. (1992)
3. Nibin, C., Wang S.F.: A fuzzy goal programming approach for the optimal planning of metropolitan solid waste management systems. European Journal of Operational Research, Vol. 99, No. 2, 303-321. (1997)
4. Fiorucci, P., Minciardi, R., Robba, M., Sacile, R.: Solid waste management in urban areas. Resources Conservation & Recycling, Vol. 37, No. 4, 301-328 (2003)
5. Rathi, S.: Alternative approaches for better municipal solid waste management in Mumbai, India. Waste Management, Vol. 26, No. 10, 1192-1200. (2006)
6. Badran, M. F., El-Haggar, S. M.: Optimization of municipal solid waste management in Port Said–Egypt. Waste Management, Vol. 26, No. 5, 534-545. (2006)

7. Dai, D, Li Y. P., Huang, G. H.: A two-stage support-vector-regression optimization model for municipal solid waste management – A case study of Beijing, China. Journal of Environmental Management, Vol. 92, No. 12, 3023-3037. (2011)
8. Chatzouridis, C., Komilis, D.: A methodology to optimally site and design municipal solid waste transfer stations using binary programming. Resources, Conservation and Recycling, Vol. 60, 89-98. (2012)
9. Lee, C. K. M., Yeung, C.L., Xiong, Z. R., Chung, S. H.: A mathematical model for municipal solid waste management –a case study in Hong Kong. Waste Management. Vol. 58, 430–441. (2016)
10. Tan, S. T., Lee, C. T., Hashim, H., Ho, W. S., Lim, J. S.: Optimal process network for municipal solid waste management in iskandar malaysia. Journal of Cleaner Production, Vol. 71, No. 4, 48-58. (2014)
11. Harijani, A. M., Mansour, S., Karimi. B., Lee, C. G.: Multi-period sustainable and integrated recycling network for municipal solid waste – a case study in Tehran. Journal of Cleaner Production, Vol. 151, 96-108. (2017).
12. Li, Y. P., Huang, G. H., Nie, S. L., Qin, X. S.: ITCLP: An inexact two-stage chance-constrained program for planning waste management systems. Resources, conservation and recycling, Vol. 49, No. 3, 284-307. (2007)
13. Xu, Y.: SRFILP: A Stochastic Robust Fuzzy Interval Linear Programming Model for Municipal Solid Waste Management under Uncertainty. Journal of Environmental Informatics, Vol. 14, No. 2, 72-82. (2009)
14. Xu, Y., Huang, G. H., Qin, X. S.: An interval-parameter stochastic robust optimization model for supporting municipal solid waste management under uncertainty. Waste management, Vol. 30, No. 2, 316-327. (2010)
15. Li, P., Chen, B.: FSILP: Fuzzy-stochastic-interval linear programming for supporting municipal solid waste management. Journal of environmental management, Vol. 92, No.4, 1198-1209. (2011)
16. Erkut, E., Karagiannidis, A., Perkoulidis, G., Tjandra, S. A.: A multicriteria facility location model for municipal solid waste management in north greece. European Journal of Operational Research, Vol. 187, No. 3, 1402-1421. (2008)
17. Minoglou, M., Komilis, D.: Optimizing the treatment and disposal of municipal solid wastes using mathematical programming—A case study in a Greek region. Resources, Conservation and Recycling, Vol. 80, 46-57. (2013)
18. Santibañez-Aguilar, J. E., Martinez-Gomez, J., Ponce-Ortega, J. M.: Optimal planning for the reuse of municipal solid waste considering economic, environmental, and safety objectives. AIChE Journal, Vol. 61, No. 6, 1881-1899. (2015)
19. Xiong. J., Ng, T. S. A., Wang. S.: An optimization model for economic feasibility analysis and design of decentralized waste-to-energy systems. Energy, Vol. 101, 239-251. (2016)
20. Asefi, H., Shahparvari, S., Chhetri, P.: Integrated Municipal Solid Waste Management under uncertainty: A tri-echelon city logistics and transportation context. Sustainable Cities and Society, Vol. 50, 101606. (2019)
21. Khattak, H. A., Ameer, Z., Din, I. U., Khan, M. K.: Cross-layer Design and Optimization Techniques in Wireless Multimedia Sensor Networks for Smart Cities. Computer Science and Information Systems, Vol. 16, No. 1. (2019)
22. SENTURK, A., KARA, R., OZCELIK, I.: Fuzzy Logic and Image Compression Based Energy Efficient Application Layer Algorithm for Wireless Multimedia Sensor Networks. Computer Science and Information Systems, Vol. 17, No. 2, 509–536. (2020)
23. Vrbaški, D., Kupusinac, A., Doroslovački, R., Stokić, E., Ivetić, D.: Missing Data Imputation in Cardiometabolic Risk Assessment: A Solution Based on Artificial Neural Networks. Computer Science and Information Systems, Vol. 17, No. 2, 379–401. (2020)

Feng Dai, PhD, is Lecture at Hubei Polytechnic University, School of Economics and Management. His research interests concern: distributed systems, applications of mathematical logic in municipal solid waste management.

Gui-hua Nie, PhD, is Professor at Wuhan University of Technology, School of Economy. His research interests concern: applications of mathematical logic in computer science, distributed systems.

Yi Chen, M.B.A., is Lecture at Hubei Polytechnic University, School of Economics and Management. Her research interests concern: distributed systems, the use of temporal epistemic logic in describing and verifying distributed protocols.

Received: August 20, 2020; Accepted: March 15, 2021.

Computer Science and Information Systems 18(4):1359–1378 https://doi.org/10.2298/CSIS200917026Z

The Application of E-commerce Recommendation System in Smart Cities based on Big Data and Cloud Computing

Yiman Zhang

College of Oujiang, Wenzhou University, Wenzhou, 325035, China
00194012@wzu.edu.cn

Abstract. In the era of big data, the amount of Internet data is growing explosively. How to quickly obtain valuable information from massive data has become a challenging task. To effectively solve the problems faced by recommendation technology, such as data sparsity, scalability, and real-time recommendation, a personalized recommendation algorithm for e-commerce based on Hadoop is designed aiming at the problems in collaborative filtering recommendation algorithm. Hadoop cloud computing platform has powerful computing and storage capabilities, which are used to improve the collaborative filtering recommendation algorithm based on project, and establish a comprehensive evaluation system. The effectiveness of the proposed personalized recommendation algorithm is further verified through the analysis and comparison with some traditional collaborative filtering algorithms. The experimental results show that the e-commerce system based on cloud computing technology effectively improves the support of various recommendation algorithms in the system environment; the algorithm has good scalability and recommendation efficiency in the distributed cluster, and the recommendation accuracy is also improved, which can improve the sparsity, scalability and real-time problems in e-commerce personalized recommendation. This study greatly improves the recommendation performance of e-commerce, effectively solves the shortcomings of the current recommendation algorithm, and further promotes the personalized development of e-commerce.

Keywords: e-commerce, personalized recommendation, cloud computing, big data

1. Introduction

Under the background of big data era, the development of e-commerce is relatively rapid, and the trading volume in this field shows geometric growth [1]. Online shopping has become an indispensable part of people's life. Due to the increasing variety and quantity of goods on e-commerce websites, when a wide range of goods are provided, it provides users with more choices, and causes the problem of information overload [2]. In the face of massive information, personalized recommendation has become one of the most effective means to solve the problem of information overload, and it is a hot spot in the academic and e-commerce circles [3]. In this context, e-commerce recommendation system comes into being. It can capture key data from rich data information, mine potential customers for businesses, expand sales scope, and provide commodity recommendation for old customers to expand user groups [4]. As the continuous improvement of user demand, the recommendation quality of e-commerce recommendation cannot meet the requirements of users and businesses. The continuous development and expansion of e-commerce lead to the diversified e-commerce mode. On the one hand, the increasing amount of product data leads to the untimely data processing of recommendation system, and users cannot quickly and accurately search for the products they want [5]; on the other hand, the user's demand becomes more and more diversified, so that the recommendation system cannot recommend the products that users are potentially interested in, and the recommendation content is not diversified enough [6]. With the wide use of e-commerce recommendation system in various websites, a large number of user' browsing records and purchase records have been accumulated in the database. The huge amount of data and the complexity of data structure are beyond the load of ordinary single-machine programs. However, high-performance computers are expensive, which forces the original calculation and storage model to be upgraded and improved [7]. How to make the recommended content generated by the system closer to the needs of users has been the core issue in this field.

The emergence of cloud computing is just a good solution to this problem, and the cloud computing framework based on ordinary computers is more suitable for the needs of data processing in the era of big data. Cloud computing is a kind of distributed computing. In the case of increasing data volume and unstructured data and semi-structured data, cloud computing only needs to dynamically expand data storage resources and data computing resources to maintain the timely response of the recommendation system [8]. Wang et al. (2018) used cloud computing and high performance computing (HPC) technology to implement large-scale RS data management and data for dynamic environmental monitoring, effectively solving the problem of data processing in remote sensing [9]; in order to solve the problems of data sparsity and cold start in the recommendation system, Mezni and Abdeljaoued (2018) proposed a cloud service recommendation system based on collaborative filtering. The experimental results confirmed the effectiveness of the method [10]; Mahmood et al. (2018) developed a service selector system based on the advantages of cloud providers' computer trust, and implemented a multi-agent system approach. This method can fully use the data processing advantages of cloud computing to provide better agent-based intelligent cloud solutions for end users [11]; Jiang et al. (2019) proposed a cloud computing slope algorithm based on the fusion of trusted data and user similarity. The experimental results on Amazon dataset suggest that the recommended algorithm is more accurate than the traditional algorithm [12]. The above studies show that cloud computing has strong performance in various fields, especially in data processing. However, there are few studies on the application of cloud computing in e-commerce recommendation system.

Therefore, based on the analysis and research of collaborative filtering, content filtering recommendation and association rule algorithm and other key technologies, the specific implementation method of recommendation algorithm improvement and optimization based on cloud computing technology is proposed. Finally, the improved algorithm is empirically analyzed through the establishment of experimental environment. This study improves the processing capacity of personalized recommendation algorithm for big data, provides users with real-time, intelligent and accurate recommendation information of goods or logistics services, enhances the shopping experience of users, improves the marketing, sales and customer relationship management capabilities of e-commerce websites, promotes the development of regional small and medium-sized logistics enterprises, relieves the pressure of inventory and traffic, and reduces the cost of logistics services.

This exploration is divided into five parts. The first part is the introduction, which mainly puts forward the scientific problem that users are difficult to obtain key knowledge due to massive Internet data, makes a comparative analysis of previous research algorithms, and further puts forward the research content; the second part is the methods, which mainly introduces the problems faced by the current personalized recommendation algorithm, proposes data mining technology, designs e-commerce recommendation system based on cloud computing, and proposes data and computer configuration to verify the model; the third part is the results and discussion, which mainly introduces the comparison of recommendation efficiency of e-commerce recommendation algorithm based on cloud computing, the comparison of recommendation performance and scalability of improved e-commerce recommendation system; the fourth part is the discussion, which compares the personalized recommendation algorithm in previous studies with the method proposed in this exploration, and further puts forward the possibility of future application; the fifth part is the conclusion, which gives a detailed description of the main contributions and limitations of this study.

2. Methods

2.1. Personalized Recommendation System

As the most important part of e-commerce personalized recommendation system, personalized recommendation technology largely determines the type and performance of e-commerce personalized recommendation system. At present, according to the different recommendation methods, personalized recommendation can be divided into collaborative filtering recommendation, recommendation based on association rules, recommendation based on user statistics, and combined recommendation technology. The traditional personalized recommendation system consists of three parts: behavior record, analysis module and recommendation algorithm [13]. Figure 3 shows the specific workflow. Among them, the behavior record module is used to collect the user's behavior information (browsing and rating); the model analysis module mainly uses the information data collected by the behavior record module to analyze, obtain the potential user preferences and the degree of liking, and establish the corresponding user preference information model; according to the recommendation model, the recommendation algorithm module finds the products that the user may like from the product set based on the user's preference information, and then recommends it to the user.

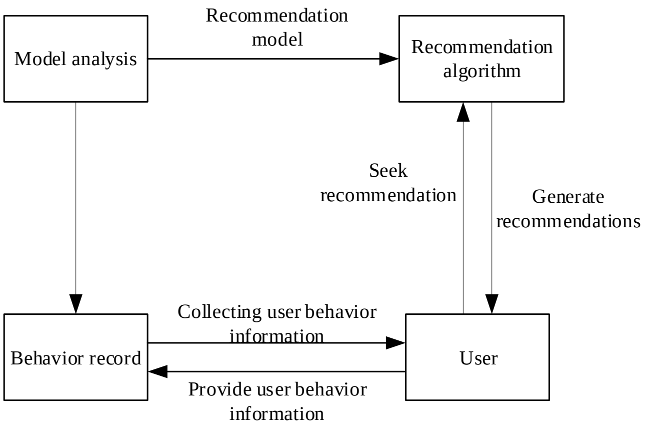


Fig. 1. Internet of things architecture

From the perspective of e-commerce, the recommendation system can be regarded as application software that can help e-commerce websites recommend commodities. User's behavior information data are collected, and they will be analyzed through statistical analysis, machine learning and other analysis methods. The commodities that users may be interested in are found from the product database and recommended to users, which can improve the sales level of e-commerce, increase user stickiness, promote consumption, and promote the development of e-commerce at last.

2.2. Big Data

Big data can be simply regarded as data with an extremely large scale. Since big data itself has an abstract concept, its definition has not been completely unified. First, big data is considered to be a dataset that takes more time to acquire, manage and process data with software tools than can be tolerated, but this definition is too unilateral and does not reflect the characteristics of big data. Then, the famous big data 3V model is proposed, which believes that big data has three characteristics: massive, diverse, and high-speed. Later, with the continuous development of the eras, it is proposed that big data also has the characteristics of value and authenticity [14].

2.3. Design of Recommendation System for E-commerce Based on Cloud Computing

(1) The framework of the recommendation system for e-commerce based on cloud computing: the e-commerce recommendation system based on cloud computing is constructed in a hierarchical structure, which can be divided into four layers from top to bottom: application layer, recommendation engine layer, cloud computing platform layer, and data source layer. The four layers are independent and mutual-restricted with

each other. They use the interface to interact information with each other, and design the level interior through the modular idea, so as to ensure the performance advantages of high cohesion, low coupling and easy to expand in the system architecture [15]. The data source layer mainly stores the data information on the e-commerce website, including the original data from different machines in various forms. After the integration of these data, the recommendation system obtains the information characteristics of users and commodities. The data information is preprocessed by the cloud computing platform layer. The information stored in the data source layer has the characteristics of multi-source, heterogeneity, and multi-type, and these characteristics lead to the increase of data noise; therefore, it is necessary to preprocess and filter the data to remove the noise before they are used. Data preprocessing includes five processes: data extraction, data cleaning, data conversion, data mapping and data integration. Different extraction methods are used to extract the corresponding characteristic data. Finally, the unified structure is used to store the data. The cloud computing platform layer mainly uses distributed computing and distributed storage systems to process and calculate data, which is mainly completed by the Hadoop platform. Figure 2 shows the overall architecture of the distributed recommendation system.

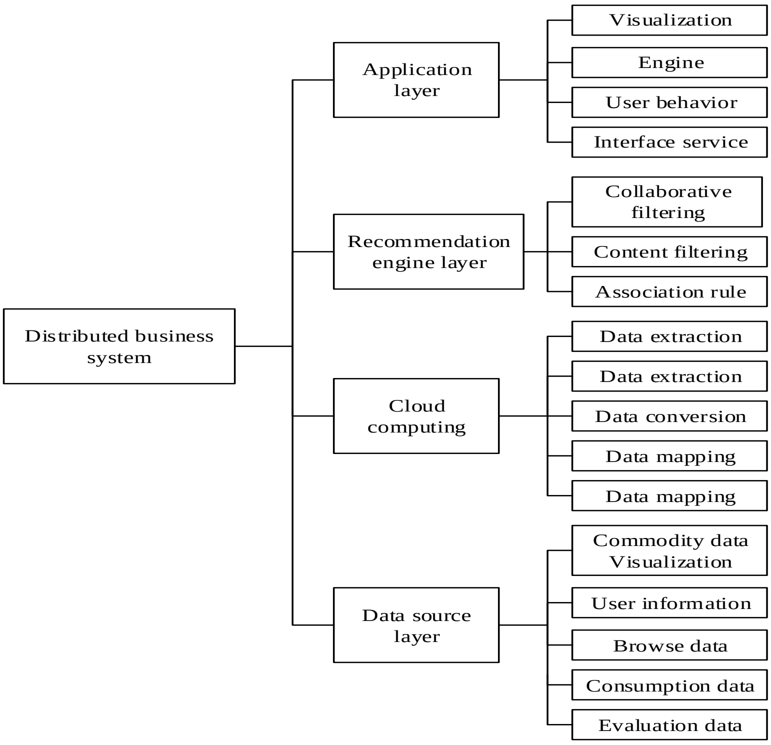


Fig. 2. Overall architecture of distributed recommendation system

The recommendation engine layer is the core layer of the recommendation system. It uses some general algorithm interfaces in the cloud computing platform to construct a recommendation algorithm and recommendation strategy into a recommendation engine that can operate independently. According to the diversified recommendation requirements, a corresponding recommendation engine is designed for each different recommendation requirement, which solves the scalability of the recommendation system. The recommendation engine based on collaborative filtering technology, content filtering calculation and association rule is designed.

(2) The cloud computing platform – Hadoop: it is a kind of distributed system infrastructure, which is supported by cheap computer cluster hardware to deal with massive data. Hadoop platform makes it easier for users to develop distributed programs; it can use cluster hardware to achieve massive data storage and high-speed computing, which has excellent scalability and high reliability [16]. Hadoop distributed file system (HDFS) and basic execution unit (MapReduce) of distributed computing tasks are the core components of the Hadoop platform. The HDFS is at the bottom of the Hadoop platform, which is mainly used to store files in all data nodes in the cluster. MapReduce is used to process massive data [17]. Figure 3 presents the architecture of the Hadoop platform.

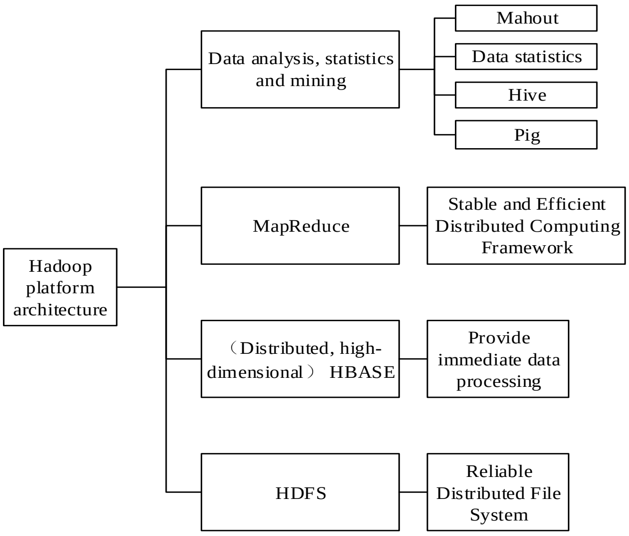

Hadoop platform architecture
Data analysis, statistics and mining
Mahout
Data statistics
Hive
Pig
MapReduce
Stable and Efficient Distributed Computing Framework
（Distributed, high-dimensional） HBASE
Provide immediate data processing
HDFS
Reliable Distributed File System

Fig. 3. The architecture of the Hadoop platform

HDFS is mainly used for the storage of data files and adopts the master/slave architecture, providing storage services for high-throughput, reliable and scalable large data files of upper layer distributed computing tasks. MapReduce is a programming model for parallel computing of large datasets, and it is easy to use and understand; the use of this programming model does not require users to understand its distributed and parallel programming, and the development of the program can be realized by using map function and reduce function [18]. MapReduce is used to process big data, which is realized mainly through the idea of dividing and ruling. When cloud computing technology is used to design a recommendation system of e-commerce, the concurrency elements involved in the traditional algorithm need to be found out. These parallel tasks are opposite to each other, so that the distributed computing method can be used directly; however, for the serial tasks, they should be decomposed as much as possible; then, the parallel tasks are found to calculate them. Figure 4 is the processing flow of tasks with MapReduce.

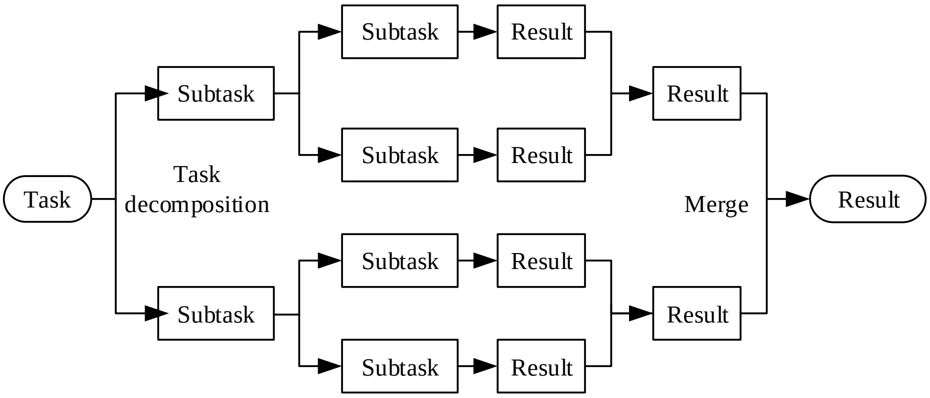


Fig. 4. Flow chart of task processing based on MapReduce

2.4. Personalized Recommendation Algorithm Based on Cloud Computing

(1) Collaborative filtering recommendation algorithm based on cloud computing: the collaborative filtering recommendation algorithm can be divided into user-based collaborative filtering (UserCF) and an item-based collaborative filtering algorithm (ItemCF) [19]. The widely used recommendation algorithm in e-commerce system is collaborative filtering algorithm, which simulates the scene of mutual recommendation between people in real life, uses the user's behavior characteristics in their historical information data to calculate the user's similarity, and uses the similarity data to recommend the product information to the user. $C=\{c_1,c_2,\ldots c_n\}$ is regarded as the set of all users in the system, and $S=\{s_1,s_2,\ldots s_n\}$ is a set of all products. The score of user c for the unevaluated product s is rc,s,. Then, the score is calculated as follows.

$$r_{c,s} = \frac{1}{N}\sum_{\hat{c}\in\hat{C}} r_{\hat{c},s} \tag{1}$$

$$r_{c,s} = k\sum_{\hat{c}\in\hat{C}} sim(c,\hat{c}) \bullet r_{\hat{c},s} \tag{2}$$

$$r_{c,s} = \bar{r}_c + k\sum_{\hat{c}\in\hat{C}} sim(c,\hat{c}) \bullet (r_{\hat{c},s} - \bar{r}_{\hat{c}}) \tag{3}$$

$\hat{C}$ represents the similarity set of c, k is a standardization factor, and $sim(c,\hat{c})$ represents the similarity between targeting user c and similar user $\hat{c}$. $\bar{r}_c$ represents the average score of user c, and $\bar{r}_{\hat{c}}$ represents the average score of the user $\hat{c}$.

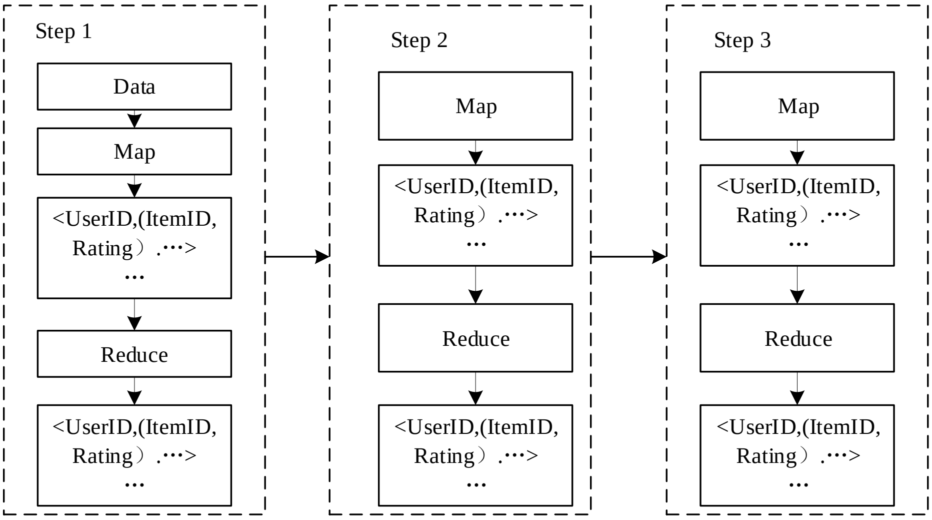

Step 1
Data
Map
<UserID,(ItemID, Rating）.···>
···
Reduce
<UserID,(ItemID, Rating）.···>
···
Step 2
Map
<UserID,(ItemID, Rating）.···>
···
Reduce
<UserID,(ItemID, Rating）.···>
···
Step 3
Map
<UserID,(ItemID, Rating）.···>
···
Reduce
<UserID,(ItemID, Rating）.···>
···

Fig. 5. Flow chart of system filtering recommendation algorithm improved by cloud computing technology

The principle of the collaborative filtering recommendation algorithm improved by cloud computing is the same as that of the traditional collaborative filtering recommendation algorithm. However, compared with the traditional collaborative filtering recommendation algorithm, the performance efficiency of the algorithm is improved because of the enhanced ability of distributed parallel computing.

(2) Content filtering recommendation algorithm based on cloud computing: the content-based recommendation (CBR) method recommends objects with similar attributes to users according to their selection objects, and Figure 6 shows the recommendation process based on content filtering.

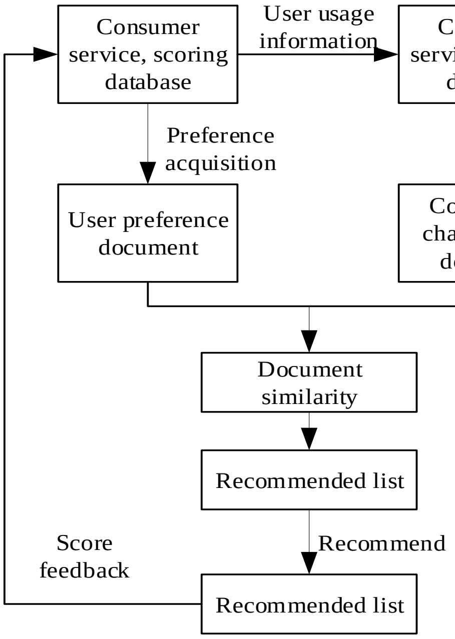


Fig. 6. Recommendation flow chart based on content filtering

Algorithm recommendation flow based on content filtering is as follows. First, according to the user's scoring database, the attribute features of the item are extracted, and the feature documents of the item are constructed. Combined with the user's historical information, the user's preference document is constructed, and the similarity between the two documents is calculated to find items similar to the user's preference and recommend them to the user. Both feature documents and user preference documents are built based on the content recommendation algorithm, and the content information of the item is represented by the vector space model [21]. If item i has k attributes, wij is used to represent the weight of the j-th attribute of item i; therefore, the contentProfile (i) of item i can be expressed in the following ways.

$$content\mathrm{Pr}\,ofile(i) = \{w_{i1}, w_{i2}, \cdots, w_{ik}\} \tag{4}$$

The user's preference information can be obtained by decision-tree, Bayesian classification algorithm, neural networks, and other machine learning algorithms. The importance of the j-th attribute to user u is represented by wuj, and the userProfile (u) can be calculated by the following equation.

$$user\mathrm{Pr}\,ofile(i) = \{w_{u1}, w_{u2}, \cdots, w_{uk}\} \tag{5}$$

Finally, cosine similarity is used to calculate the similarity between item i and user u in the item document and user preference document. The calculation method is as follows.

$$sim(u,i) = \cos(u,i) = \frac{\sum_{j=1}^{k} w_{u,j} w_{ij}}{\sqrt{\sum_{j=1}^{k} w_{u,j}^{2}} \sqrt{\sum_{j=1}^{k} w_{i,j}^{2}}} \tag{6}$$

MR-CBR, an improved content filtering recommendation algorithm based on cloud computing technology, can be regarded as a mutually independent and parallelizable process when user's preference documents are calculated, and can be regarded as a MapReduce. When the similarity between two document information is calculated, it is also an independent and feasible calculation process and regarded as another MapReduce, and there is a serial relationship between the two MapReduces [22]. Figure 7 presents the flow of the improved content filtering recommendation algorithm based on cloud computing.

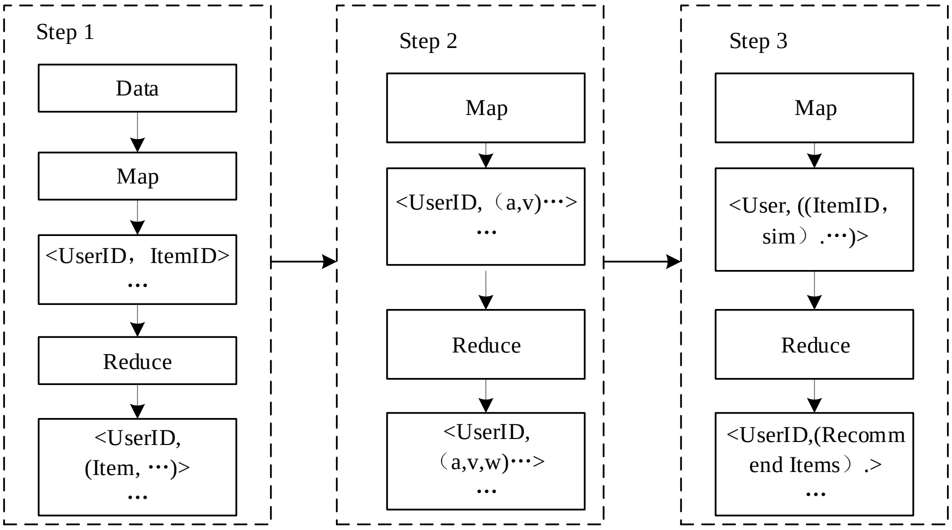


Figure. 7. The flow chart of the improved content filtering recommendation algorithm based on cloud computing

(3) An improved association rule recommendation algorithm based on cloud computing: association rules are based on mining the correlation between items from many data. It can analyze the transaction data of commodities, find out the commodities that are purchased frequently at the same time from the data, generate corresponding rules, and recommend for users based on the current behavior data of users [23]. The item set is set to I={i1,i2,…in}, the commodity transaction database is represented by D, and each transaction is represented by T. T is a subset of commodity items, and A is regarded as an item set when A is less than or equal to T. It can be considered that trade T contains A. The calculating method of ($\sup port(A \Rightarrow B)$) is as follows.

$$\sup port(A \Rightarrow B) = P(A \cup B) = \frac{\left|\{T : A \cup B \subseteq T, T \in D\}\right|}{|D|} \tag{7}$$

The $confidence(A \Rightarrow B)$ is calculated as follows.

$$confidence(A \Rightarrow B) = P(B \mid A) = \frac{\left|\{T : A \cup B \subseteq T, T \in D\}\right|}{\left|\{T : A \subseteq T, T \in D\}\right|} \tag{8}$$

It is difficult to extract rules due to the sparsity and high dimension of data in the association rule recommendation algorithm, resulting in quality instability. In addition, the offline building time of rule in this algorithm is relatively long. Meanwhile, the algorithm will increase the management difficulty with the increase in the number of rules.

For the improved association rule recommendation algorithm (MR-FP) based on cloud computing, it is not necessary to build frequent tree for the whole transaction set in the construction of a frequent tree. The frequent tree of the frequent term conditions is constructed with each calculated node, and the corresponding conditional frequent tree is established. Then, the final solution of the algorithm is obtained by combining conditional frequent trees, and the association rules of the frequent term set are obtained. Finally, according to the association rules, products are recommended to users [24].

2.5. Construction of a Ccomprehensive Evaluation System for Distributed Recommendation System

To verify the advantages and disadvantages of the distributed recommendation systems in three kinds of e-commerces, a comprehensive evaluation system of the corresponding distributed recommendation system will be established [25]. The following principles should be followed in indicator selection, as shown in Table 1.

Table 1. Selection principle of comprehensive evaluation system index

number	principle	selection method
one	objectiveness	comprehensive and easy to quantify
two	systematicness	can reflect essential features
three	independence	an indicator reflects a single element
four	science	scientific and reasonable to meet the theory of statistics, economics, e-commerce and other related disciplines

Precision, efficiency, coverage, diversity, and novelty are regarded as evaluation indicators of the e-commerce recommendation system. When the precision of the recommended system is calculated, the mean absolute error (MAE) can be used as the judgment method for the precision of the system score prediction. The calculation method is as follows.

$$MAE = \frac{\sum_{u,i\in T} \left| r_{ui} - \hat{r}_{ui} \right|}{|T|} \tag{9}$$

In the above equation, r_{ui} refers to the actual score of user u for product *i*, and $\hat{r}_{ui}$ refers to the predicting score of user u for the product i.

Users' preferences for recommended items can be calculated by the prediction precision recommended by TopN, expressed by two indicators of precision and recall, and fitted by F1-Score. The calculation method is as follows.

$$\Pr ecision = \frac{\sum_{u\in U} \left| R(u) \cap T(u) \right|}{\sum_{u\in U} \left| R(u) \right|} \tag{10}$$

$$\operatorname{Re} call = \frac{\sum_{u\in U} \left| R(u) \cap T(u) \right|}{\sum_{u\in U} \left| T(u) \right|} \tag{11}$$

$$F1 - Score = \frac{2 * \Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \tag{12}$$

In the above equations, *R(u)* refers to the user's behavior and the number of items in their recommendation list, and *T(u)* refers to the number of items the user likes.

Efficiency refers to the time consumed by the algorithm in calculating and processing, and the calculation of this indicator can be measured by the time consumed by the algorithm operating on the computer.

Coverage refers to the widespread degree of items recommended by e-commerce recommendation system to users. The calculation method is as follows.

$$Coverage = \frac{\left| U_{u\in U} R(u) \right|}{|I|} \tag{13}$$

Where *U* refers to the user set of the system, and *R(u)* refers to the list of products recommended to users.

Diversity is represented by the dissimilarity of products in the recommendation list and the dissimilarity of the list recommended to different users. Hamming distance (HM) is used to express the dissimilarity of recommendation lists of different users *u* and *v* [26], and the calculation method is as follows.

$$HM_{u,v} = 1 - \frac{R(u) \cap R(v)}{R(u) \cup R(v)} \tag{14}$$

In the above equations, *R(u)* and *R(v)* refer to the list of products recommended to users u and v. When *R(u)* and *R(v)* are identical, the value of HM is 0, and when there is no overlap, it is 1. When the value of HM is larger, the system diversity is higher.

Novelty means that the recommendation system recommends some non-popular new products to users, and the novelty of the recommended items can be evaluated according to the average popularity of the recommendation list. When the average popularity of products in the recommendation list is smaller, the novelty of the recommendation system is stronger.

The e-commerce recommendation system is built through the Hadoop platform. Because the number of nodes of the platform can be increased and decreased flexibly, the recommendation effect of the distributed recommendation system and in the single machine environment is compared by the acceleration ratio (R). The calculation method of *R* is as follows.

$$R = \frac{Ts}{Tc} \tag{15}$$

Ts in the above equation represents the operating time of the recommendation system in a single machine environment, and *Tc* represents the operating time of the distributed recommendation system.

3. Results and Analysis

3.1. Comparison of Recommendation Efficiency of E-commerce Recommendation Algorithm Based on Cloud Computing

The recommendation efficiency of the algorithm is compared mainly through the comparison of the operating speed of the algorithm. The operating time of the algorithm is inversely proportional to the computing ability of the algorithm. The shorter the operating time of the algorithm is, the stronger the computing ability of the algorithm is, and the higher the recommendation efficiency of the algorithm is. The recommended efficiency of several algorithms is compared, and the comparison results are shown in Figure 8. It suggests that when the amount of input data information increases, the running time of the algorithm under different nodes increases slowly, which shows that the greater the amount of information to be calculated is, the slower the calculation and processing speed of the algorithm are. However, the running time of the algorithm under 7 nodes is better than that of 5 nodes, the running time of algorithm with 5 nodes is better than that of 3 nodes, and the running time of algorithm with 3 nodes is better than that of 2 nodes. It shows that the in the distributed platform, the more the nodes are, the stronger the computing power of the algorithm is, and the faster the execution speed of the algorithm is. Therefore, the improved algorithm runs fast. The recommendation system has high recommendation efficiency.

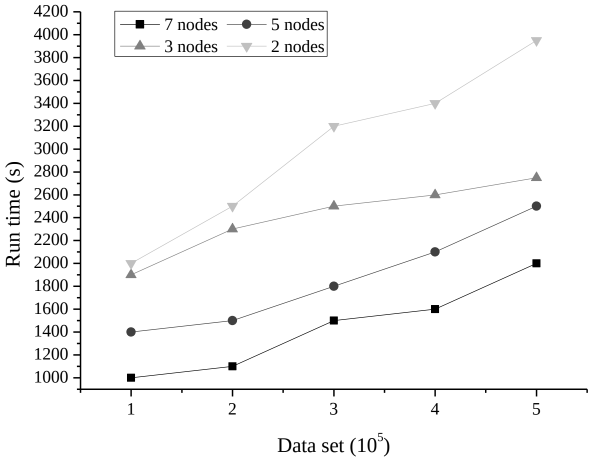


Fig. 8. Analysis chart of algorithm recommendation efficiency based on cloud computing improvement

3.2. Recommendation Performance Comparison of E-commerce Recommendation System Based on Cloud Computing iImprovement

The comprehensive evaluation system is used to analyze the performance of the e-commerce recommendation system based on cloud computing, and the analysis results are shown in Figure 9. It reveals that as far as the accuracy indicator is concerned, the final evaluation results are accuracy rate, recall rate and the average value of F1 in five experiments, because the three recommendation engines of collaborative filtering, content filtering and association rules are all based on TopN recommendation. The F1 values are 7.2%, 8.4% and 5.6%, respectively. It suggests that the content filtering recommendation engine is superior in recommendation accuracy; in terms of efficiency indicator, 5000 users are randomly selected for offline calculation, and the running time of the three is 12s, 23s and 42s, respectively, which shows that the collaborative filtering recommendation engine is the best in this respect; in terms of coverage indicator, the proportion of items recommended for all users by the three recommendation engines is counted. Coverage rate is 45.6%, 68.4% and 72.1%, respectively; in terms of diversity indicator, the recommended list of the three engines for the users at a certain time is selected, and the Hamming distance of the user pairs used. The diversity degree of the three recommendation engines is 74.6% 86.2%, and 64.5%, respectively. In terms of novelty indicator, the average popularity of items in the recommended list of the three recommendation engines is 15.2%, 28.6% and 15.6%, respectively. The above research results prove that the performance of content filtering recommendation system based on cloud computing is the best, and the system filtering recommendation system based on cloud computing has the least running time and the highest efficiency.

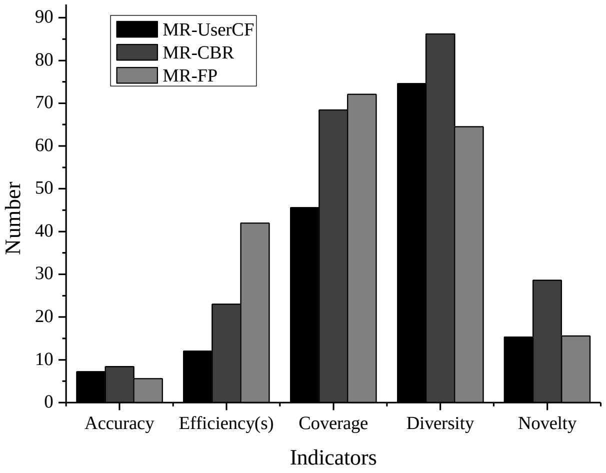


Fig. 9. Comprehensive evaluation indicator results chart of different recommendation algorithms

3.3. Scalable Performance Comparison of E-commerce Recommendation System Based on Cloud Computing

The R of several recommendation systems in a single machine environment (node number is 1) and distributed recommendation system (node number is at least 2) is compared, and the comparison results are shown in Figure 10 and Figure 11. It shows that the acceleration ratio of the algorithm increases with the increasing number of cluster nodes. The change is very fast in the early stage and slow in the later stage. The acceleration ratio of the algorithm with 7 nodes is larger than that of the algorithm with 5 nodes. The acceleration ratio of the algorithm with 5 nodes is larger than that of the algorithm with 3 nodes. The acceleration ratio of the algorithm under 3 nodes is larger than that of the algorithm under 1 node. It shows that in the distributed platform, the more the nodes are, the stronger the computing power of the algorithm is, the better the recommendation effect of the recommendation system is. However, the experiment suggests that when the number of nodes is less than 5, the acceleration ratio changes linearly. When the number of nodes is greater than 5, the acceleration ratio changes slowly with the increasing number of nodes. It reveals that simply increasing the number of cluster nodes cannot infinitely improve the performance of the algorithm.

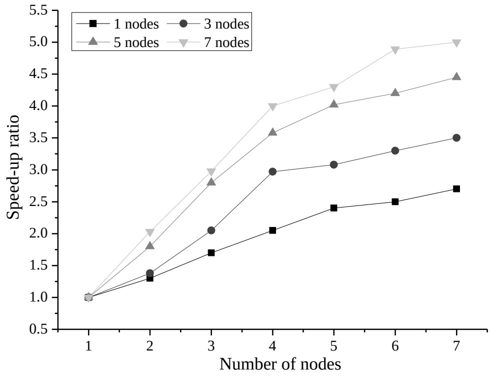


Fig. 10. Acceleration ratio analysis chart of the recommendation algorithm under different cluster nodes

Figure 10 suggests that the R of the distributed recommendation algorithm increases gradually with the increase of the number of nodes. It indicates that the scalability performance of the recommendation algorithm is better than that in the single machine environment. Figure 11 shows that the three distributed algorithms have obvious advantages over traditional algorithms in execution time. In the case of experimental data, with the increase of the number of clusters, the acceleration ratio of the three recommendation algorithms gradually increases, which shows that the distributed algorithm continues to run at this time. When the cluster size reaches 10, the execution efficiency of MR-UserCF algorithm is more than 7 times than that of single machine. In the best case, MR-CBR and MR-FP algorithms can reach 4 times and 3 times. However, due to the scale of experimental data, when the Hadoop cluster nodes are added after reaching the peak of acceleration ratio, the growth rate slows down with the increase of nodes, but the slope decreases, which indicates that the growth rate slows down with the increase of nodes. However, in general, the increase of the number of nodes can effectively guarantee the decrease of system running time.

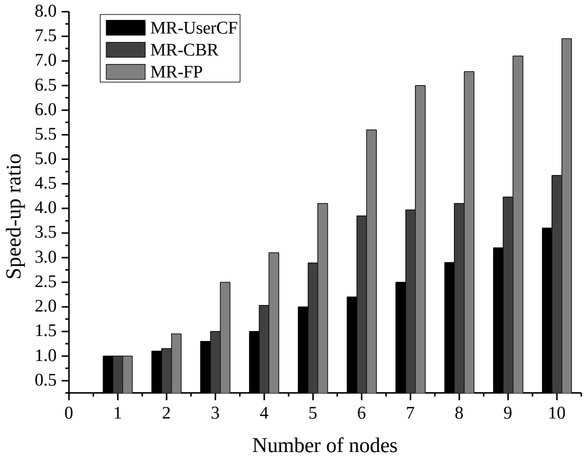


Fig. 11. R analysis chart of different recommendation algorithms

4. Discussion

Cloud computing technology and Hadoop platform are used to improve the traditional e-commerce recommendation algorithm. The recommendation speed of the algorithm significantly increases after the improvement (Figure 8). The main reason is that the more the nodes are in the distributed platform, the stronger the computing power of the algorithm is, the faster the implementation speed of the algorithm is. This is confirmed by the research in Zhu and Bai (2020). They took collaborative filtering recommendation algorithm as an example of data mining platform, and introduced the idea of weighted factor based on project popularity to improve the degree of personalized recommendation system. The improved algorithm improves the performance of personalized recommendation system [27]. Based on the open source cloud computing platform Hadoop, MapReduce parallel framework is used to process massive data sets. Distributed file system HDFS is adopted to store and manage super large files. Compared with the traditional recommendation algorithm, it has obvious improvement in recommendation speed, which can greatly improve the recommendation speed of the recommendation system and improve the user satisfaction (Figure 9). It has also been confirmed in the research of Tang and Cheng (2017). They constructed and replaced the rating matrix based on user preference characteristics. MapReduce parallel computing framework can improve computing efficiency and algorithm scalability [28]. The applicability and scalability of traditional recommendation algorithms for single machine computing in the face of large data sets is systematically analyzed, including the operation mechanism of divide and rule and the method of problem domain partition in MapReduce method of cloud computing. From the perspective of parallel decomposition of tasks and data, the design principles of recommendation algorithm based on cloud computing are proposed. It is found that simply increasing the number of nodes in the cluster cannot improve the performance of the algorithm infinitely, and the increase of the number of nodes can effectively guarantee the decrease of system running time. This is confirmed by Cao et al. (2018). They found that in the process of MapReduce parallelization, the data partition matrix is stored in line segments, the computing load is distributed in each node of the cluster, and the time consumption and partition matrix consumption of moving data matrix is calculated, which can reduce the calculation amount in the execution process, and greatly reduce the consumption of storage space, but the number of nodes cannot be increased too much [29]. Hadoop platform is used to build e-commerce recommendation system, and improve the parallel computing ability of the recommendation system. Therefore, this study is consistent with the existing research results, which shows that the research method is effective and feasible.

5. Conclusion

In the environment of big data, the traditional system filtering recommendation, content filtering, and association rule filtering algorithms are rebuilt on the Hadoop platform through cloud computing technology, and the parallel processing efficiency of several algorithms is improved by MapReduce. A comprehensive evaluation system is established. Its calculation and analysis as well as R prove that several improved distributed recommendation algorithms based on cloud computing are more efficient than those in the traditional single machine environment; moreover, the performance of the distributed recommendation algorithm in R comparison and analysis is excellent. It shows that the improved distributed recommendation system based on cloud computing improves the support of the recommendation algorithm, reduces the operating time of the algorithm, and improves the recommendation efficiency of the algorithm. The content filtering recommendation algorithm based on cloud computing has an excellent performance in precision, coverage, diversity, and novelty. The collaborative filtering recommendation system based on cloud computing has the best operation efficiency and good scalability.

The Hadoop platform is a relatively mature research platform at present. In the future development, the optimization and improvement of this platform is still a research focus, and it is hoped that more traditional algorithms can be improved through this platform to improve the performance of the algorithm; moreover, when cloud computing technology is used to improve the traditional algorithm, MapReduce improves the parallel processing efficiency of the algorithm; however, in coordinating the hybrid systems of different cloud platforms and coordinating the layout of datasets, further research is needed to improve the working performance of the cloud system.

6. References

1. Fan, W.; Xu, M.; Dong, X.; Wei, H.: Considerable environmental impact of the rapid development of China's express delivery industry. Resources, Conservation and Recycling, Vol. 126, 174-176. (2017)
2. Swar, B.; Hameed, T.; Reychav, I.: Information overload, psychological ill-being, and behavioral intention to continue online healthcare information search. Computers in Human Behavior, Vol. 70, 416-425. (2017)
3. Xiao, J.; Wang, M.; Jiang, B.; Li, J.: A personalized recommendation system with combinational algorithm for online learning. Journal of Ambient Intelligence and Humanized Computing, Vol. 9, 667-677. (2018)
4. Hwangbo, H.; Kim, Y.S.; Cha, K.J.: Recommendation system development for fashion retail e-commerce. Electronic Commerce Research and Applications, Vol. 28, 94-101. (2018)
5. Subramaniyaswamy, V.; Logesh, R.; Chandrashekhar, M.; Challa, A.; Vijayakumar, V.: A personalised movie recommendation system based on collaborative filtering. International Journal of High Performance Computing and Networking, Vol. 10, 54-63. (2017)
6. Kumar, P.; Thakur, R.S.: Recommendation system techniques and related issues: a survey. International Journal of Information Technology, Vol. 10, 495-501. (2018)
7. Tian, G.; Wang, J.: Recommendation algorithm for mobile E-commerce based on cone depth learning. International Journal of Computers and Applications, 1-6. (2019)
8. Xu, B.; Xu, L.; Cai, H.; Jiang, L.; Luo, Y.; Gu, Y.: The design of an m-Health monitoring system based on a cloud computing platform. Enterprise Information Systems, Vol. 11, 17-36. (2017)
9. Wang, L.; Ma, Y.; Yan, J.; Chang, V.; Zomaya, A.Y.: pipsCloud: High performance cloud computing for remote sensing big data management and processing. Future Generation Computer Systems, Vol. 78, 353-368. (2018)
10. Mezni, H.; Abdeljaoued, T. A cloud services recommendation system based on Fuzzy Formal Concept Analysis. Data & Knowledge Engineering, Vol. 116, 100-123. (2018)
11. Mahmood, A.; Shoaib, U.; Shahzad, S.: A Recommendation System for Cloud Services Selection Based on Intelligent Agents. Indian Journal of Science and Technology, Vol. 11, 1-6. (2018)
12. Jiang, L.; Cheng, Y.; Yang, L.; Li, J.; Yan, H.; Wang, X.: A trust-based collaborative filtering algorithm for E-commerce recommendation system. Journal of Ambient Intelligence and Humanized Computing, Vol. 10, 3023-3034. (2019)
13. Zhuo, R.; Bai, Z. Key technologies of cloud computing-based IoT data mining. International Journal of Computers and Applications, 1-8. (2020)
14. Alsbatin L, Öz G, Ulusoy A H.: Efficient virtual machine placement algorithms for consolidation in cloud data centers. Computer Science and Information Systems, Vol. 1, 36-36. (2019)
15. Tarus, J.K.; Niu, Z.; Mustafa, G.: Knowledge-based recommendation: A review of ontology-based recommender systems for e-learning. Artif. Intell. Rev, Vol. 50, 21-48. (2018)
16. Wu, D.; Xu, Y.; Zhou, Y. Study on the technical architecture of the computer data mining platform based on the hadoop framework. Boletin Tecnico/Technical Bulletin, Vol. 55, 275-281. (2017)
17. Kune, R.; Konugurthi, P.K.; Agarwal, A.; et al.: XHAMI - extended HDFS and MapReduce interface for Big Data image processing applications in cloud computing environments. Software, Vol. 47, 455-572. (2017)
18. Bi Z, Dou S, Liu Z, et al. A recommendations model with multiaspect awareness and hierarchical user-product attention mechanisms. Computer Science and Information Systems, Vol. 0, 24-24. (2020)
19. Yi, L.; Jun, F.; Tong-Tong, W.; et al. An Improved Collaborative Filtering Recommendation Algorithm. computer and modernization, Vol. 32, 204-208. (2017)

20. Li, C.; He, K.: CBMR: An optimized MapReduce for item-based collaborative filtering recommendation algorithm with empirical analysis. Concurrency & Computation Practice & Experience, Vol. 29, 4092-4106. (2017)
21. Gavalas, D.; Konstantopoulos, C.; Mastakas, K.; Pantziou, G.: Mobile recommender systems in tourism. J. Netw. Comput. Appl, Vol. 39, 319-333. (2014)
22. Son, J.; Kim, S.B.: Content-based filtering for recommendation systems using multiattribute networks. Expert Systems with Application, Vol. 89, 404-412. (2017)
23. Hang, L.; Kang, S.-H.; Jin, W.; Kim, D.-H. Design and Implementation of an Optimal Travel Route Recommender System on Big Data for Tourists in Jeju. Processes, Vol. 6, 133-143. (2018)
24. Li, C.; Liang, W.; Wu, Z.; et al.: [IEEE 2018 IEEE International Conference on Web Services (ICWS) - San Francisco, CA, USA (2018.7.2-2018.7.7)] 2018 IEEE International Conference on Web Services (ICWS)An Efficient Distributed-Computing Framework for Association-Rule-Based Recommendation, 339-342. (2018)
25. Bhatta R, Ezeife C I,.: Distributed Data Mining-Based E-Commerce Recommendation System. Computer Systems & Applications, Vol. 18, 33-37. (2009)
26. Kacprzyk, J.; Nurmi, H.; Sławomir, Z.: Towards a Comprehensive Similarity Analysis of Voting Procedures Using Rough Sets and Similarity Measures. Intelligent Systems Reference Library, Vol. 42, 359-380. (2013)
27. Zhuo, R.; Bai, Z.: Key technologies of cloud computing-based IoT data mining. International Journal of Computers and Applications, 1-8. (2020)
28. Tang, H.; Cheng, X. Personalized E-commerce recommendation system based on collaborative filtering under Hadoop. World, Vol. 1, 146-148. (2017)
29. Cao, Y.; Li, P.; Zhang, Y.: Parallel processing algorithm for railway signal fault diagnosis data based on cloud computing. Future Generation Computer Systems, Vol. 88, 179-283. (2018)

Yiman Zhang was born in Wenzhou, Zhejiang, China, in 1983. She received the Master degree from Kyoto College of Graduate Studies for Informatics, Japan. Now, she works at Oujiang College of Wenzhou University. Her research interests include Data mining and E-commerce.

Received: September 17, 2020; Accepted: February 19, 2021.

Computer Science and Information Systems 18(4):1379–1394 https://doi.org/10.2298/CSIS200901027L

Optimization of Intelligent Heating Ventilation Air Conditioning System in Urban Building based on BIM and Artificial Intelligence Technology

Zhonghui Liu[1,*] and Gongyi Jiang[2]

[1]School of Environmental Engineering, the University of Kitakyushu,
Kitakyushu 8080135, Japan,
z8dbb412@eng.kitakyu-u.ac.jp
[2]Tourism College of Zhejiang, Hangzhou 310000, China,
jgy1128@tourzj.edu.cn

Abstract. The study aims to effectively reduce building energy consumption, improve the utilization efficiency of building resources, reduce the emission of pollutants and greenhouse gases, and protect the ecological environment. A prediction model of heating ventilation air conditioning (HVAC) energy consumption is established by using back propagation neural network (BPNN) and adapted boosting (Adaboost) algorithm. Then, the HVAC system is optimized by building information modeling (BIM). Finally, the effectiveness of the urban intelligent HVAC optimization prediction model based on BIM and artificial intelligence (AI) is further verified by simulation experiments. The research shows that the error of the prediction model is reduced, the accuracy is higher after the Adaboost algorithm is added to BPNN, and the average prediction accuracy is 86%. When the BIM is combined with the prediction model, the HVAC programme of hybrid cooling beam + variable air volume reheating is taken as the optimal programme of HVAC system. The power consumption and gas consumption of the programme are the least, and the CO_2 emission is also the lowest. Programme 1 is compared with programme 3, and the cost is saved by 37% and 15%, respectively. Through the combination of BIM technology and AI technology, the energy consumption of HVAC is effectively reduced, and the resource utilization rate is significantly improved, which can provide theoretical basis for the research of energy-saving equipment.

Keywords: building information modeling, Adaboost-BP algorithm, heating ventilation air conditioning system, energy consumption prediction, simulation.

1. Introduction

As the economy and society develop fast, people's life is becoming more and more stable, and the economic benefits that people pursue at the cost of environment have been retaliated. The high frequency of global extreme climate has seriously threatened people's life safety [1]. With the consumption of resources, there are global energy shortage problems. How to improve the efficiency of energy utilization, reduce resource

[*] Corresponding author

consumption, reduce carbon dioxide emissions, and curb the pace of global greenhouse effect has become an important issue of widespread concern in the international community [2]. The current data show that China's existing building energy consumption accounts for about 33% of the total social energy consumption. Among the buildings owned, buildings with high energy consumption account for a large proportion. Due to the increase of such buildings, the crisis of energy shortage has been aggravated [3]. According to the existing data, China now has a construction area of nearly 40 billion square meters, 90% of which belong to high energy consumption buildings [4]. The energy consumption of heating ventilation air conditioning (HVAC) system accounts for more than half of all building energy consumptions. Although the heat given by heating per unit area of buildings in China is three times that of advanced national buildings, the degree of comfort brought to users during heating period cannot meet people's expectations [5]. Therefore, the optimization research of intelligent HVAC in urban buildings has become an urgent scientific problem to be solved in this field [6]. The optimization of HVAC system design can not only reduce the energy loss of HVAC system, but also provide happier living space for indoor users.

The data information in building information model (BIM) technology can provide data support and basis for judging green building performance and quality. BIM describes the characteristics of buildings in a data-based way, which can provide the data of buildings at all stages of the project [7]. As artificial intelligence technology develops rapidly, various algorithms can effectively optimize the system, and then achieve the energy consumption reduction [8]. Among them, Shalabi and Turkan (2017) improved the data required for air conditioning system maintenance by using the visualization and operability functions of BIM. The results showed that the system could effectively feedback the fault data of air conditioning equipment to the control platform, and reduce the increasing energy consumption cost caused by continuous operation of equipment due to fault [9]. Afram et al. (2017) designed the predictive control system based on the artificial neutral network (ANN) model, and found that the model could significantly reduce the operating cost of HVAC equipment without affecting the system performance [10]. Ghahramani et al. (2017) optimized the HVAC system in industrial buildings assisted by genetic algorithm. The optimized system ensured the daily minimum energy consumption and the thermal comfort of air conditioning [11]. Sporr et al. (2019) designed a HVAC control system based on the building information data in BIM. The system could improve the existing control system, thus optimizing the operation energy consumption of building air conditioning equipment [12]. Based on the above studies, it is found that if only the supply of the main power grid and gas network is relied on to meet the needs of users, for the supply side, the energy utilization rate is extremely low, and huge power supply pressure is caused during the peak period of energy consumption, which shortens the service life of the power grid and endangers the security of power supply. On the demand side, it will also increase the power consumption cost and cause certain losses to the electrical equipment.

In order to deal with the above problems, based on the in-depth analysis of the existing HVAC system problems, effective solutions are put forward. The integrated back propagation neural network (BPNN) and Adaboost algorithm are used for HVAC system research, and targeted energy saving and emission reduction is carried out. BIM technology is used to simulate the HVAC system of urban buildings, and construct the appropriate optimization programme. The design of HVAC system based on this programme can not only solve many problems existing in the current design process of traditional air conditioning system, but also improve the efficiency of HVAC specialty and other related specialties in the design process, so that HVAC plays a more important role in the exploration of energy-saving building design.

There are five sections in total. The first section is the introduction, which explains the advantages of the heating and ventilation system and the necessity of its research. It also discusses previous studies and clarifies the differences of this investigation. The second section introduces the research method, such as the energy consumption prediction algorithm of the HVAC system, the BIM technology, the HVAC system optimization design based on BIM technology, and the air conditioning system's energy consumption simulation process described by the Adaboost-BP algorithm. The third section presents the results, which explains the performance of the Adaboost-BP prediction algorithm and analyzes the energy consumption of the HVAC system. The fourth section is the discussion, which compares the obtained results with the state of the art algorithms in previous works and probes into the possible problems of the system. The fifth section is the conclusion, which explains the principal contributions and limitations in detail.

2. Method

2.1. Energy consumption prediction algorithm for HVAC system

(1) Adaboost algorithm: The integrated learning algorithm is a research focus in machine learning algorithms. In the integrated learning algorithm, the enhanced learning algorithm is one of the commonly used algorithms. Adaboost algorithm (adapted boosting) is the most popular one in the current meta-algorithm and widely used in various classification prediction problems [13]. Adaboost algorithm has a low generalization error rate, and can be implemented by coding, which is suitable for most classifiers without parameter adjustment. However, this algorithm has a running time field and is more sensitive to outliers. To some certain extent, the algorithm relies too much on the training data and the selection of weak classifiers, which means that when the training data can't satisfy the algorithm, the classification performance of the weak classifier is not high enough, and the classification effect of the algorithm will also become worse [14].

The Adaboost algorithm will first give the same initial weight value to each sample in the training dataset and convert the weight value into a vector. The training sample data are used to train the established weak classifier, the error rate of the classifier is obtained, and then the weak classifier is trained again through the training dataset. Based on the obtained error rate, the weight of the sample in the training data is adjusted to reduce the weight value of the accurate sample data of each classification, and the weight value of the wrong sample data is improved. The weight is adjusted to transform the weak classifier into a strong classifier and reduce the error rate of classification, and the algorithm is characterized by eliminating interference and high training pertinence.

However, due to the high dependence of the algorithm on the training data and the performance of the classifier, it is easy to fall into local minimum. While BPNN is a multi-layer feedforward neural network algorithm based on error back propagation and it has high nonlinear mapping ability, self-learning, and adaptive ability. Meanwhile, the generalization performance is also relatively excellent, which can meet the demand of Adaboost algorithm for weak classifier and improving BP algorithm is easy to fall into the disadvantage of local minimum [15].

(2) BPNN algorithm: Generally, BPNN mainly includes input layer, hidden layer, and output layer. The original data are input into the input layer of BPNN, and the calculation results are obtained through the calculation of the hidden layer and the output layer [16]. When BPNN is used for calculation, the weights and thresholds between levels are adjusted by means of back propagation of errors, and the calculation results are not output until the calculation results meet the error range or reach the maximum number of iterations. The number of nodes in each hierarchy of BPNN is multiple, among which the number of nodes in the input layer is determined by the data characteristic variables contained in the data sample, and the number of nodes in the output layer is determined by the number of sample classification. The number of hidden layers can have multiple layers, and each layer can contain one or more nodes. However, the number of layers and the number of nodes of this layer are determined through subjective experience, and then adjusted through the data calculated later. The calculation method is shown in (1).

$$m = 2n + 1 \tag{1}$$

In (1), n is the number of input nodes, and m is the number of nodes in the output layer.

With the increase of the number of network layers, the complexity of BPNN also increases, and meantime, it will also lead to a decrease in the accuracy and speed of the algorithm [17]. Therefore, the neural network algorithm with the total number of network layers of 3 layers is selected in experiment. The number of nodes in the hidden layer h of BPNN is as follows:

$$h_i = f_1(\sum w_{ij} + b_i) \tag{2}$$

Then, the calculation method of neuron node number y in the output layer is as follows.

$$y_k = f_2(\sum w_{jk} + b_k) \tag{3}$$

In equations (2) and (3), w_{ij} is the initial weight value between the input layer and the hidden layer, w_{jk} is the initial weight value between the hidden layer and the output layer, f_1 and f_2 refer to the transfer functions between the input layer and the hidden layer, and between the hidden layer and the output layer, and b_i and b_k are the thresholds of the hidden layer and the output layer, respectively.

(3) Adaboost-BP algorithm: BPNN has the characteristics of strong classification ability, which can meet the requirements of Adaboost algorithm with strong dependence on the classifier, and Adaboost algorithm can improve BPNN from falling into the local minimum. Therefore, Adaboost-BP algorithm is combined with these two algorithms to

predict the energy consumption of building HVAC system [18]. The prediction flow of the combined algorithm is shown in Figure 1 below.

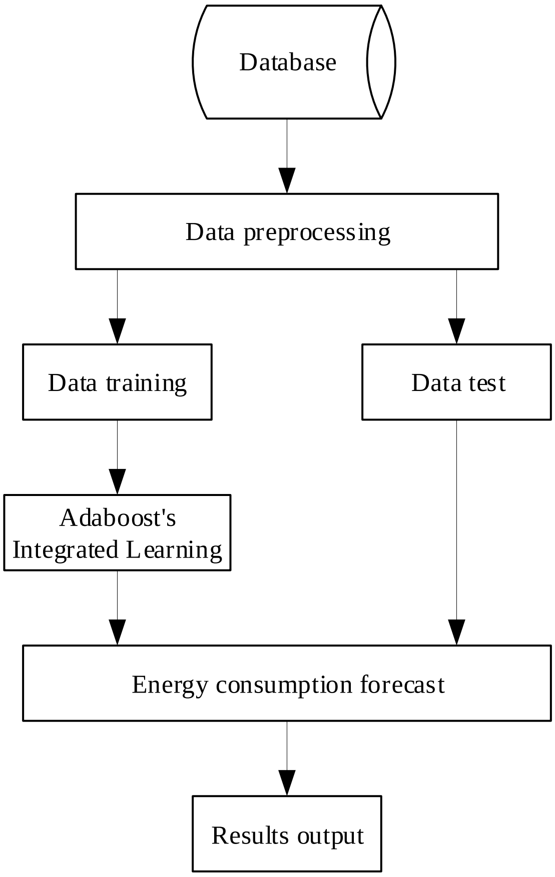


Fig. 1. Flow chart of system energy consumption prediction based on Adaboost-BP algorithm

When Adaboost-BP algorithm is used for energy consumption prediction, the data is prepared first. A total of S group training datasets is randomly extracted from the original sample datasets, and then the datasets give the initial weight values. The method is shown as (4).

$$Q_i = \frac{1}{S}, i = 1,2,\cdots s \tag{4}$$

In (4), *i* refers to the sample dataset, and Q is the initial weight value of the dataset.

According to the characteristic number (dimension) of the training data, BPNN is built, and the initial weight and threshold value of each node are given in the neural network.

It is also necessary to establish a weak classifier. The method is to establish a BPNN weak classifier of the corresponding value through the cycle coefficient N required by the algorithm. Next, the error rate of the established weak classifiers is calculated, and according to the error rate, the initial weight value of the weak classifier is calculated. The calculation method is shown as follows.

$$\alpha = \frac{1}{2}\ln(\frac{1-\varepsilon}{\varepsilon}) \tag{5}$$

After obtaining the initial weight value of each sample, it is necessary to adjust the weight value, reduce the weight value of the sample with correct classification, and increase the weight value of the sample with wrong classification.

After adjusting the weight value, Adaboost continues the next round of loop iteration. During the loop iteration, the training sample data is repeatedly trained, and the weight value of the sample is constantly adjusted until the classification error rate of the sample reaches the allowable range or reaches the maximum number of loop iterations. These weak classifiers are combined to form a strong classifier. The calculation method is shown as follows.

$$G(x) = sign(f(x)) = \left\{ \sum_{m=1}^{M} \alpha_m G_m(x) \right\} \tag{6}$$

In (6), m is the number of layer nodes, and Gm(X) is the weak classifier.

2.2. BIM tecnology

BIM technology uses all data parameters in the engineering project to build the physical model of the building, and converts the information related to the building into data for storage. It can store and simulate all the data of the project from the planning to the completion of the construction. It provides scientific basis for the decision-making of relevant personnel, changes the working form of the project construction, improves the construction efficiency of the construction project, and reduces the risk of the project construction [19]. With the further development of information technology, BIM technology has also updated drawing software tools and improves the connection and interoperability of data information during project construction. From the perspective of building standards, BIM technology digitally expresses all data related to the project and BIM technology has the characteristics of visualization, datalization, coordination, and simulation [20].

Visualization refers to the building entity model established by using BIM technology, which can directly express all the processes of the project from planning and design to construction, and then to operation and maintenance through simulation, showing the design effect diagram. Relevant personnel can discuss and modify the project design, construction, operation, and maintenance phase of the programme at any time. Datalization is mainly about the data management of the relevant information generated by the project. By using BIM technology, project data can be quickly and efficiently calculated and processed, the construction progress of the project can be accelerated, effective calculation and digital information can be provided for the project, and scientific and reasonable simulation analysis can be carried out for the project meantime. Therefore, the management of the construction project is more precise, scientific, and meticulous. Coordination means that BIM technology can be used for collaborative design. Before the project construction, the design programme involved in the project can be simulated through BIM and then integrated. It can find the contradictions in the construction process of the project as early as possible, help relevant personnel to find the problems fast, formulate reasonable solutions, and improve the work efficiency of the project. Simulation means that BIM technology can not only simulate and analyze project engineering, but also simulate application research to a certain extent. BIM can analyze and simulate the requirements of the whole project, such as energy conservation needs, security needs, and comfort requirements. It can make 3D simulation of the construction site, thus facilitating the planning and design of scientific, reasonable construction plans and guidance programmes. Additionally, it can also carry out 5D simulation analysis to help the relevant personnel to control the project cost [21].

BIM technology has professional value, information value, and management value. Professional value means that the technology can deepen and optimize the projects involved in construction projects, information value mainly means that the data generated in the project could be stored, analyzed and managed, and management value means that relevant personnel can optimize and control their professional value and information value to improve efficiency [22].

2.3. Optimization design of HVAC system based on BIM technology

To some extent, the optimization of HVAC system can promote the energy conservation of buildings and improve the design level of HVAC system. The simulation of HVAC system by using BIM technology can directly present the energy consumption of building HVAC system. Moreover, different data information can be used to carry out simulation and compare the energy consumption of the HVAC system under different conditions, thereby selecting the best design programme of the HVAC system. Furthermore, the technology has a good docking, and can be connected with a variety of software, and directly passes the software system data for simulation. The information data of the optimized air conditioning system can reflect that the system guides BIM to conduct modeling of the air conditioning system, and then the data of the optimization results and simulation results of the air conditioning system can be stored to facilitate the later invocation [23]. The optimization process of air conditioning system using BIM is shown in Figure 2 below.

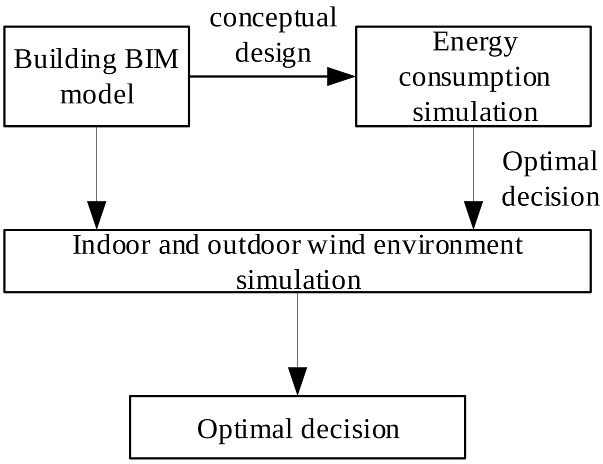

Building BIM model
conceptual design
Energy consumption simulation
Optimal decision
Indoor and outdoor wind environment simulation
Optimal decision

Fig. 2. Optimization process of air conditioning system based on BIM technology

The HVAC system is used to ensure that the indoor temperature is suitable by means of system cooling or heating. Meanwhile, the system can also improve the indoor air circulation, ensure the smooth exchange between the outside and the indoor air, improve the indoor air environment, and improve the air quality [24]. The energy consumption of HVAC system is mainly produced by three parts: heating system, ventilation system, and HVAC system energy consumption, and the energy consumption of HVAC system is mainly related to HVAC system. Energy consumption modeling using BIM technology is mainly based on the characteristics of the building, and the corresponding mathematical equation and computer software are used to simulate the energy consumption of the building.

2.4. Simulation of energy consumption of air conditioning system based on BIM combined with Adaboost-BP algorithm

BIM technology and Adaboost-BP algorithm are used to simulate the energy consumption of HVAC system in urban buildings in experiment. Common BIM design platforms include Autodesk Revit, ArchiCAD, and Bentley, and the most widely used Autodesk Revit design platform is selected in experiment. The design platform supports the creation of architectural models, structural models, and electromechanical system models [25].

In the design of HVAC system, parameters such as indoor and outdoor wind environment and sunshine environment should be considered to optimize the design of building HVAC system. During the design process, computational fluid dynamics (CFD) is also used to calculate the discrete distribution of the fluid flow field in the region and to simulate the fluid flow [26].

Taking commercial buildings as an example, in the system design, BIM technology is first used to model the buildings. All the data information of the construction project is imported into the BIM, the geographical environment, indoor and outdoor meteorological parameters will affect the energy consumption of HVAC system, and indirectly cause the final analysis error. Therefore, it is necessary to analyze outdoor meteorological parameters, including outdoor wind speed, relative temperature, and humidity [27]. In addition, it is necessary to simulate the annual cold and hot load of the building, calculate the maximum cold and hot load value, and then set the temperature of the building HVAC system according to the corresponding load value. The main purpose in experiment is to simulate the variable air volume (VAV) system. Three different VAV systems are shown in Table 1 below.

Table 1. Performance comparison of different VAV air conditioning systems

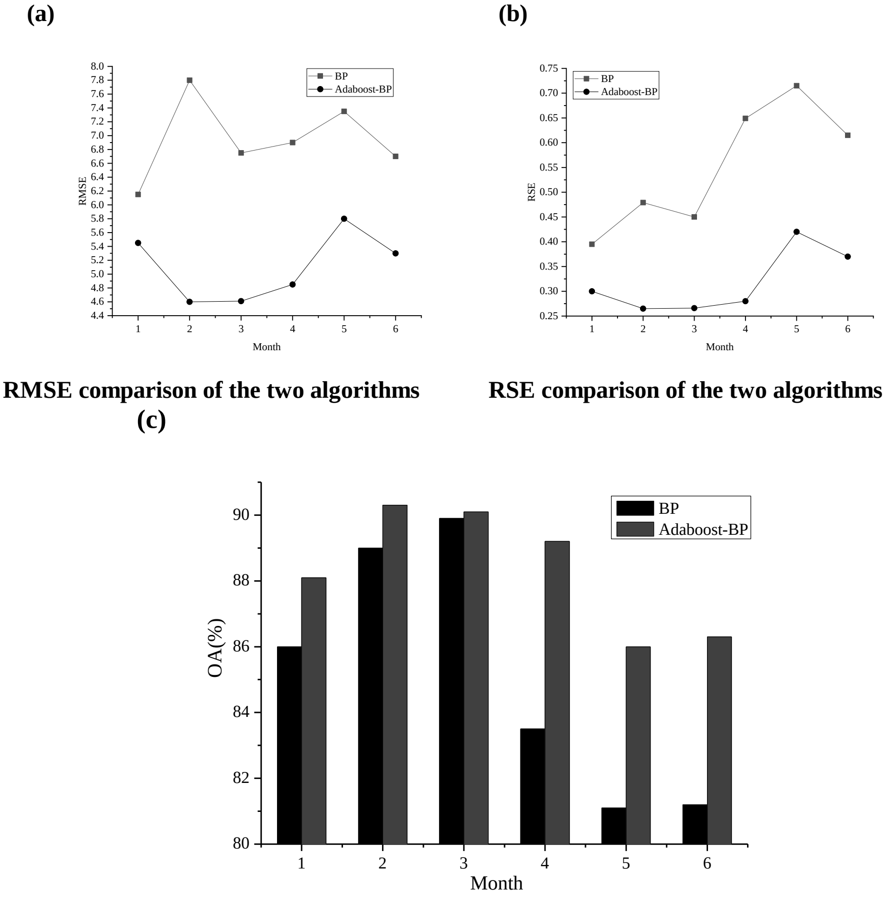
Programme	VAV system end unit	Characteristic	Advantage
1	VAV system in parallel fan power box	The end unit used electric, steam or hot water to heat the coil to gain additional heat	Good flexibility, quiet operation, waste heat recovery, night return operation
2	Mixed cooling beam + VAV reheating system	Air processor with external constant volume	Improving air quality
3	VAV system with reheating system	Heating occurred primarily or exclusively at the regional level, with cross flow control and additional heating and cooling to control the regional temperature	Reheating the end

In addition, building HVAC systems will also increase CO_2 emissions, which will lead to the occurrence of the global greenhouse effect. Therefore, the CO_2 emissions of these programmes should be simulated and analyzed to optimize the overall air conditioning system.

3. Results

3.1. Performance analysis of Adaboost-BP prediction algorithm

In order to verify the performance of the energy consumption prediction algorithm studied in experiment, the prediction accuracy of the traditional BPNN algorithm and Adaboost-BP algorithm is compared from the predicted root mean square error (RMES), relative square root error (RSE), and overall accuracy (OA). The comparison results are shown in Figure 3 below. OA of BPNN monthly energy consumption forecast is between 81.16% and 89.89%. The prediction effect of January is the best, which can reach 89.89%, while that of March is only 81.16%. However, the overall monthly accuracy of Adaboost-BP is improved to 85.91% - 90.31%. It shows that compared with the single BPNN, Adaboost-BP algorithm reflects a higher overall accuracy, and for the prediction model with low accuracy of weak classifier classification effect, such as February and March, it shows better improvement effect. This is directly related to the parameter setting characteristics of Adaboost ensemble learning algorithm. Adaboost algorithm has low requirements for weak classifier, and hardly needs to adjust the parameters of weak classification, which is another reason why Adaboost BP algorithm is widely used.



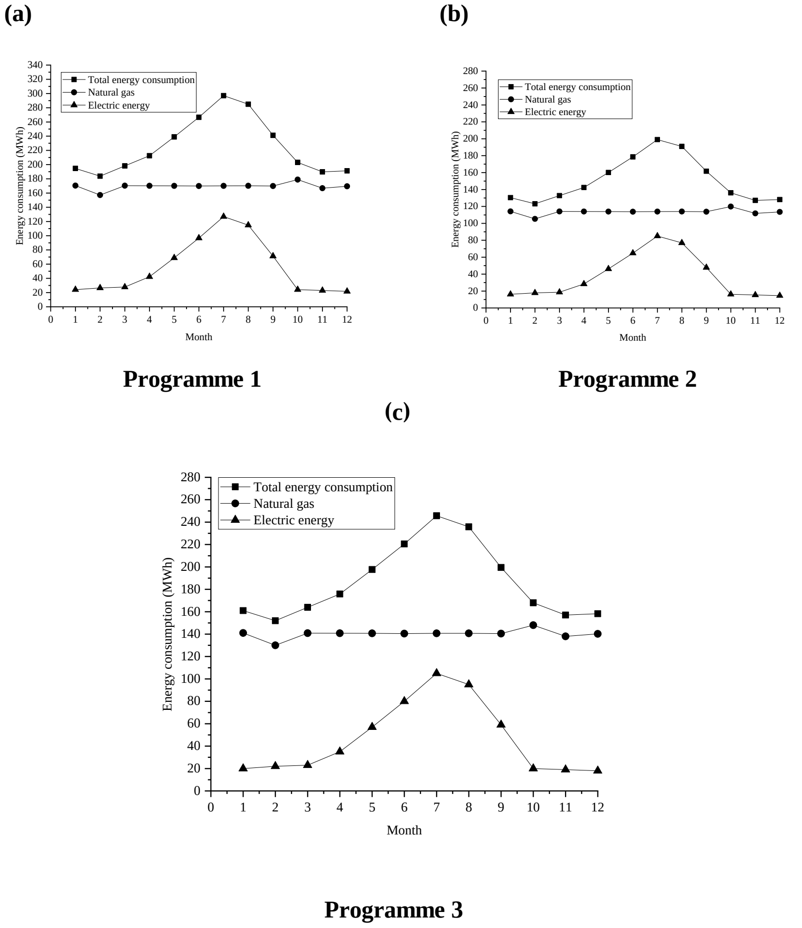
RMSE comparison of the two algorithms **RSE comparison of the two algorithms**

Overall accuracy comparison results of the two algorithms

Fig. 3. Comparison results of prediction error accuracy of the two algorithms

3.2. Energy consumption analysis of different HVAC systems

The total energy consumption, natural gas, electric energy and boiler energy consumption of three programmes of HVAC system are analyzed and calculated to save energy consumption and reduce energy waste, and the optimal programme is found out. The optimal design is carried out by using the programme system. The analysis results are shown in Figure 4 below. Figure 4 suggests that the power consumption and gas consumption of the programme 2 are the lowest. During the operation, the operation cost is lower than that of programme 1 and programme 3. Compared with programme 1 and programme 3, energy consumption is saved by 33% and 19%, respectively. Moreover, in terms of cost, the programme 2 also saves money. Compared with the programme 1 and programme 3, the cost is saved by 37% and 15%, respectively.



(a)

(b)

Programme 1

Programme 2

(c)

Programme 3

Fig. 4. Energy consumption analysis of different HVAC systems

When the system is designed, it not only considers the energy consumption of the system, but also considers the system cost and the emission of CO_2. The analysis and calculation results of energy consumption cost and CO_2 emission of each programme are shown in Figure 5 below. Figure 5 shows that the CO_2 emission of the programme 2 is the lowest, which is 60% lower than that of programme 1 and 31% lower than that of programme 3. The programme 2 is the best in terms of energy consumption and the impact on the environment, so the programme 2 is selected as the optimal programme, and the information of the selected air conditioning system is fed back to the Revit MEP model to establish the optimal air conditioning system model.

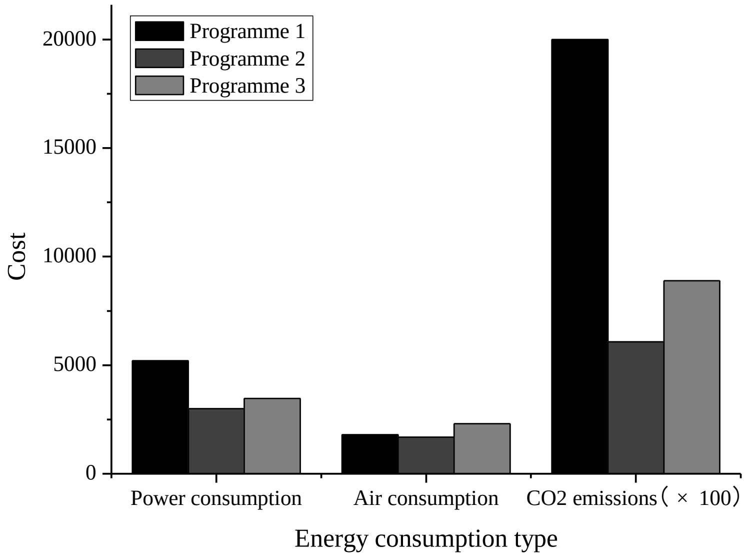


Fig. 5. Analysis results of energy consumption values and CO_2 emissions of different programmes

4. Discussion

First, based on BPNN algorithm, according to the previous research, it is found that the algorithm has some limitations, such as poor generalization ability and easy to fall into local minimum value. Therefore, a building integrated learning energy consumption prediction model based on Adaboost-BP algorithm is constructed by using Adaboost ensemble learning algorithm. It is found that the model has higher accuracy than the single BPNN, and it has better performance for the prediction model with low accuracy of weak classifier classification effect (Figure 3), which is consistent with the research results of Lu et al. (2017). Adaboost algorithm is used to improve BPNN. The life predicted by the improved method is compared with the traditional BPNN. Different power levels of light emitting diode (LED) lamps are compared. It is found that the average relative error of the improved method is reduced by 54%. However, the improved method takes 63.6% longer running time [28]. However, the measured operating time does not decrease much, which may be because the HVAC system has been in operation.

In addition, regarding the energy saving measures of HVAC system based on BIM technology, building envelope energy saving, air conditioning system scheme comparison and indoor and outdoor wind environment analysis, energy saving of HVAC system is studied. The programme 1 is to use water chilling unit and cooling tower in summer. In winter, the hot water boiler mode is used for heating, and the parallel fan power box mode of variable air volume system is adopted at the end. This mode provides great flexibility to meet the design requirements of various HVAC systems. This method has been reported in reference [29]. The programme 2 is to use the combination of water chilling unit and cooling tower in summer, hot water boiler mode in winter, and hybrid cooling beam + variable air volume reheating system at the end. The design of this system includes two external constant volume air processors to improve the quality, which has also been reported in reference [30]. In the programme 2, water cooling units and cooling towers are used in summer, and hot water boiler mode is used in winter. The variable air volume system with reheating system is used at the end of the system. Moreover, heating and cooling the air provided to the laboratory is to control the temperature and humidity, which is also more common [31]. BIM technology is used. Three programmes are dsicussed, the optimal HVAC programme is obtained and compared with other programmes, and the energy consumption of hybrid cooling beam + variable air volume reheating system is reduced by about 30%, which provides a new programme for HVAC system, and it is determined by energy consumption simulation, which is also mentioned in reference [32].

5. Conclusion

The energy consumption of building HVAC system is predicted by combining BPNN and Adaboost algorithm to form Adaboost-BP algorithm. Then, the relevant concepts of BIM technology are introduced, and the design of HVAC system is optimized based on BIM technology. Finally, the energy consumption of different HVAC systems is simulated by using BIM model combined with Adaboost-BP algorithm. Adaboost-BP algorithm can improve BPNN, which is easy to fall into the local minimum, and improve the classification effect of the weak classifier. The prediction error of this algorithm is smaller and the accuracy is higher. The BIM model combined with Adaboost-BP algorithm are used to simulate the energy consumption of three HVAC systems, and the results show that the HVAC system with mixed cooling beam and VAV reheating has the least power consumption and air consumption, and the CO_2 emission is reduced. Therefore, it is used as the optimization programme of HVAC system. The result shows that the Adaboost-BP algorithm can be used to predict the energy consumption of the air conditioning system. The HVAC system with mixed cooling beam and VAV reheating can reduce the energy consumption of the air conditioning system and the CO_2 emissions, so it can be used as the optimization programme of the HVAC system.

Only three different VAV systems are selected to conduct research when modeling air conditioning energy consumption using BIM technology, and the programmes are relatively few. Meanwhile, when modeling energy consumption, only some factors affecting building parameters are selected, which lack certain comprehensiveness. It is hoped that in the following work, factors causing interference to the building HVAC system can be fully considered for a comprehensive analysis, thus improving the comprehensiveness of the research. Due to the limitation of conditions, the results of the optimization programme are only based on theories, and lack certain practicality. It is hoped that it can be demonstrated in the subsequent research, and the feasibility of programme can be studied.

References

1. King, A.D.; Karoly, D.J.: Climate extremes in Europe at 1.5 and 2 degrees of global warming. Environmental Research Letters, Vol. 12, 114031-114039. (2017)
2. Carroll, J.; Aravena, C.; Denny, E.: Low energy efficiency in rental properties: Asymmetric information or low willingness-to-pay? Energy Policy, Vol. 96, 617-629. (2016)
3. Zhang, Y.; Yan, D.; Hu, S.; Guo, S.: Modelling of energy consumption and carbon emission from the building construction sector in China, a process-based LCA approach. Energy Policy, Vol. 134, 110949-110952. (2019)
4. Lin, B.; Tan, R.: Estimating energy conservation potential in China’s energy intensive industries with rebound effect. Journal of Cleaner Production, Vol. 156, 899-910. (2017)
5. Capozzoli, A.; Piscitelli, M.S.; Gorrino, A.; Ballarini, I.; Corrado, V.: Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings. Sustainable cities and society, Vol. 35, 191-208. (2017)
6. Chakraborty, N.; Mondal, A.; Mondal, S.: Multiobjective Optimal Scheduling Framework for HVAC Devices in Energy-Efficient Buildings. IEEE Systems Journal, Vol. 13, 4398-4409. (2019)
7. Huang, B.; Lei, J.; Ren, F.; Chen, Y.; Zhao, Q.; Li, S.; Lin, Y.: Contribution and obstacle analysis of applying BIM in promoting green buildings. Journal of Cleaner Production, 123946-123953. (2020)
8. Li, J.; Wang, R.; Wang, J.; Li, Y.: Analysis and forecasting of the oil consumption in China based on combination models optimized by artificial intelligence algorithms Energy, Vol. 144, 243-264. (2018)
9. Shalabi, F.; Turkan, Y.: IFC BIM-based facility management approach to optimize data collection for corrective maintenance. Journal of performance of constructed facilities, Vol. 31, 04016081-04016091. (2017)
10. Afram, A.; Janabi-Sharifi, F.; Fung, A.S.; Raahemifar, K.: Artificial neural network (ANN) based model predictive control (MPC) and optimization of HVAC systems: A state of the art review and case study of a residential HVAC system. Energy and Buildings, Vol. 141, 96-113. (2017)
11. Ghahramani, A.; Karvigh, S.A.; Becerik-Gerber, B.: HVAC system energy optimization using an adaptive hybrid metaheuristic. Energy and Buildings, Vol. 152, 149-161. (2017)
12. Sporr, A.; Zucker, G.; Hofmann, R.: Automated HVAC control creation based on building information modeling (BIM): Ventilation system. IEEE Access, Vol. 7, 74747-74758. (2019)
13. Ghimire, D.; Lee, J.: Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class Adaboost and Support Vector Machines. Sensors, Vol. 13, 7714-7734. (2013)
14. Bai, Q.; Jin, C.: Image Fusion and Recognition Based on Compressed Sensing Theory. Int. J. Smart Sens. Intell. Syst, Vol. 8, 159-180. (2015)
15. Khattak H A, Ameer Z, Din U I, et al. Cross-layer design and optimization techniques in wireless multimedia sensor networks for smart cities. Computer Science and Information Systems, Vol. 16, No. 1, 1-17. (2019)

16. Sun, L; Wei, Q; He, L.: The prediction of building heating and ventilation energy consumption base on Adaboost-bp algorithm. IOP Conference Series Materials ence and Engineering, Vol. 782, 032008-032016. (2020)
17. Zhang Y; Jia, Y; Wu, W; et al.: Research on Fault Diagnosis Method of Gearbox Based on SA and BP-Adaboost. IOP Conference Series Materials ence and Engineering, Vol. 793, 012009-012013. (2020)
18. Zheng, Y, L; Lin, P, J; Yu, J, L; et al.: A novel fault diagnosis method for photovoltaic array based on BP-Adaboost strong classifier. Iop Conference, Vol. 188, 012110-012126. (2018)
19. Chong, H.Y.; Lee, C.Y.; Wang, X.: A mixed review of the adoption of Building Information Modelling (BIM) for sustainability. J. Clean. Prod, Vol. 142, 4114-4126. (2017)
20. Katipamula, S.; Gowri, K.; Hernandez, G.: An open-source automated continuous condition-based maintenance platform for commercial buildings. Sci. Technol. Built Environ, Vol. 23, 546-556. (2017)
21. Dasović, B.; Galić, M.; Klanšek, U.: Active BIM Approach to Optimize Work Facilities and Tower Crane Locations on Construction Sites with Repetitive Operations. Buildings, Vol. 9, 21-36. (2019)
22. Parn, E.A.; Edwards, D.J.; Sing, M.C.P.: The building information modelling trajectory in facilities management: A review. Autom. Constr, Vol. 75, 45-55. (2017)
23. Zhu, J.; Wright, G.; Wang, J.; Wang, X.: A Critical Review of the Integration of Geographic Information System and Building Information Modelling at the Data Level. ISPRS Int. J. Geo-Inf, Vol. 7, 66-72. (2018)
24. Duong, M.Q.; Pham, T.D.; Nguyen, T.T.; Doan, A.T.; Tran, H.V.: Determination of Optimal Location and Sizing of Solar Photovoltaic Distribution Generation Units in Radial Distribution Systems. Energies, Vol. 12, 174-182. (2019)
25. Baik, A.: From point cloud to jeddah heritage BIM nasif historical house–case study. Digit. Appl. GG. Cult. Heritage, Vol. 4, 1-18. (2017)
26. Senturk A, Kara R, Ozcelik I. Fuzzy logic and image compression based energy efficient application layer algorithm for wireless multimedia sensor networks. Computer Science and Information Systems, Vol. 0, 8-8. (2020)
27. Mørck, O.C.: Energy saving concept development for the MORE-CONNECT pilot energy renovation of apartment blocks in Denmark. Energy Procedia, Vol. 140, 240-251. (2017)
28. Lu, K.; Zhang, W.; Sun, B.: Multidimensional data-driven life prediction method for white LEDs based on BP-NN and improved-Adaboost algorithm. Ieee Access, Vol. 5, 21660-21668. (2017)
29. Kaam, S.; Raftery, P.; Cheng, H.; Paliaga, G.: Time-averaged ventilation for optimized control of variable-air-volume systems. Energy and Buildings, Vol. 139, 465-475. (2017)
30. Alghamdi, K. Impact of Implementing a Dedicated Outdoor Air System in Parallel with a Multi-Zone Variable-Air Volume System on Energy Consumption, Thermal Comfort, and Life Cycle Cost, Vol. 23, 325-332. (2019)
31. Kim, D.; Cox, S.J.; Cho, H.; Im, P.: Evaluation of energy savings potential of variable refrigerant flow (VRF) from variable air volume (VAV) in the US climate locations. Energy Reports, Vol. 3, 85-93. (2017)
32. Abdelalim, A.; O'Brien, W.; Shi, Z.: Data visualization and analysis of energy flow on a multi-zone building scale. Automation in Construction, Vol. 84, 258-273. (2017)

Zhonghui Liu was born in Changchun, Jilin, China, in 1993. She received the bachelor‘s degree from Jinlin Jianzhu University, China. The Master’s degree from the University of Kitakyushu, Japan. Now, she studies in College of Environmental Engineering, the University of Kitakyushu. E-mail: z8dbb412@eng.kitakyu-u.ac.jp.

Gongyi Jiang was born in Hangzhou, Zhejiang,China, in 1983. She received the bachelor‘s degree from Zhejiang University City College, China, and the Master’s degree from the Zhejiang University, China. Now, she is working at Tourism College of Zhejiang, China, and studies in College of Environmental Engineering, the University of Kitakyushu. E-mail: jgy1128@tourzj.edu.cn.

Received: September 01, 2020; Accepted: March 25, 2021.

Computer Science and Information Systems 18(4):1395–1409 https://doi.org/10.2298/CSIS200922028F

Face Recognition Based on Full Convolutional Neural Network Based on Transfer Learning Model

Zhongkui Fan[1] and Ye-peng Guan[1, 2]

[1]School of Communication and Information Engineering,
Shanghai University, 200444 Shanghai, China
{fanzkui, ypguan}@shu.edu.cn
[2]Key Laboratory of Advanced Displays and System Application,
Ministry of Education, 200444 shanghai, China

Abstract: Deep learning has achieved a great success in face recognition (FR), however, little work has been done to apply deep learning for face photo-sketch recognition. This paper proposes an adaptive scale local binary pattern extraction method for optical face features. The extracted features are classified by Gaussian process. The most authoritative optical face test set LFW is used to train the trained model. Test, the test accuracy is 98.7%. Finally, the face features extracted by this method and the face features extracted from the convolutional neural network method are adapted to sketch faces through transfer learning, and the results of the adaptation are compared and analyzed. Finally, the paper tested the open-source sketch face data set CUHK Face Sketch database(CUFS) using the multimedia experiment of the Chinese University of Hong Kong. The test result was 97.4%. The result was compared with the test results of traditional sketch face recognition methods. It was found that the method recognized High efficiency, it is worth promoting.

Keywords: transfer learning, convolutional neural network, face recognition, adaptive scale, optical face features

1. Introduction

With the rapid development of Internet technology, information technology has been fully integrated into people's lives. How to accurately and effectively confirm identity has become an urgent issue in the field of information security. Face recognition [1] Compared with other biometric technologies, such as: vein recognition[2], voice recognition, fingerprint recognition, gene recognition, iris recognition, etc., because of its intuitive, convenient, non-contact and other excellent features, it makes it useful in people's daily life and society. In terms of security, such as access control, image retrieval, automatic login and face payment, and criminal investigation, it plays an important role. Therefore, it has been a research hotspot in the field of computer vision for decades. In recent years, with the development of deep learning the popularity of GPUs, and the open source of large-scale face databases, the maturity of optical face recognition technology has made it reach The practical application of standards has been applied to all walks of life, so that we have entered the era of face brushing.

There are various channels for obtaining faces. Under normal circumstances, it can be obtained through surveillance cameras or cameras to take pictures. However, in some cases, it is not possible to obtain face pictures through technical means. Only the sketches of human faces can be drawn by the artist based on the images of witnesses. This situation appeared more in the field of criminal investigation, and optical face recognition technology could not solve this problem, so sketch face recognition technology was produced [9]. This technology is a new type of face recognition technology developed on optical face recognition technology. It has very important application value in the field of criminal investigation. In recent years, it has begun to rise in the field of face recognition.

Sketch face recognition is a major branch of heterogeneous face recognition technology. Its role is to match a person's optical face with its corresponding sketch face. Its contribution is: in most criminal investigation cases, the police optical photos of the suspect were not available. At this time, the optical face recognition method has failed, but drawing a sketched face through the description of a witness is undoubtedly the most effective method to determine its identity. In view of this problem, the idea of extracting optical face features from CNN and then adapting them to sketch face features through transfer learning is proposed to solve the problem of insufficient sketch face training samples.

The proposed method achieved superior performance on CUFS [3] data-set and the contributions are summarized as follows:

(1) Adaptive scale feature extraction: An adaptive scale feature extraction method is proposed, which can adaptively adjust the feature extraction scale according to feature sensitivity.

(2) Establish a full convolutional neural network based on transfer learning model by analysing the role of each layer of AlexNet in image classification and using the VGGFace transfer learning model for face photo-sketch recognition.

2. Related Work

Sketch face recognition has been pioneered by Uhl and Lobo [4], to match sketch to photos, the proposed method uses Eigenface and Principle Component Analysis (PCA). Based on CUFS and CUFSF datasets, many state-of-the-art methods have been proposed by many researchers. Klare and Jain [5] proposed a method that extract the feature locally using a Scale Invariant Feature Transform (SIFT) descriptor. To improve the accuracy further, this method has been extended by fusing Multiscale Local Binary Pattern (MLBP) and the SIFT with Local Feature Discriminant Analysis (LFDA) [6]. Galoogahi and Sim [7] see that most of the research works use common features that are not meant for a cross-modality matching. Therefore a new face descriptor called Histogram of Averaged Oriented Gradients (HAOG) is proposed to extract modality-invariant features of salient facial components. The fact that facial shape is relatively invariant across modality, thus, Galoogahi and Sim [8] proposed a new face descriptor that is a shape-based and claimed to work on cross images. It is called Local Radon Binary Pattern (LRBP). The algorithm projects each non-overlapping patch on Radon space using Radon transform. Then, the features are extracted at every patch using Local

Binary Pattern (LBP). Recently, Difference of Gaussian Oriented Gradient Histogram (DoGOGH) has been demonstrated to be very effective in matching facial sketch to photo [9]. To cater for shape exaggerations effects, this method has been extended to Cascaded Static and Dynamic DoGOGH (C-DoGOGH) with intention to further improve the retrieval rate accuracy by catering the shape exaggerations effects [10]. It combines static and dynamic local featur extraction in a cascaded fashion.

Few works have considered the use of deep learning for face photo-sketch synthesis and recognition, most notable being the approaches in [11, 12, 13, 14]. However, these systems generally use relatively shallow networks or are primarily trained using images residing in a single modality (typically face photos).

Finally, few works consider the use of multiple sketches per subject. Most relevant to this letter is the work done in [15], [16], but the number of subjects and sketches used were both limited since the latter were manually created by employing several artists or software operators, making the process costly and time-consuming. These problems are critical, especially in the time-sensitive nature of real-world criminal investigations.

3. Adaptive Scale Feature Extraction

3.1. Feature Extraction

Extracting facial features needs to meet two basic requirements: first, to find the optimal features so that it can distinguish different faces to the greatest extent. Second, the extracted feature dimensions should be as small as possible, so that the training and testing speed can be improved. Before convolutional neural networks, the mainstream facial feature extraction algorithms were SIFT, LBP, HOG, and Gabor [17]. Theoretically speaking, the larger the extracted feature dimensions, the higher the accuracy. Literature pointed out that the use of multi-scale extraction of face features can improve the accuracy of the algorithm to a certain extent, but it will significantly increase the amount of calculation and is not conducive to engineering implementation. Based on this, an adaptive scale feature extraction method is proposed to reduce feature dimensions and improve the computing speed [18, 19]. Feature localization is the first step of feature extraction. It specifies the location of feature extraction, uses the face database with labels as training samples, learns the gradient direction of each label position, and then determines feature extraction based on the learned gradient direction position. Given an image $d \in R_{m\times 1}$ with m pixels, which contains *p* manually labeled points, as shown in Figure 1 (a), *h* is a feature extractor (herein specifically referred to as AMLBP), which uses labeled the human face is used as a training sample, and let it be x_*. When a human face is detected, an initialization mark x_0 is given as an average mark, as shown in Fig. 1 (b). Face feature localization can be achieved by minimizing the expression (1).

$$f\left(x_0+\Delta x\right)=\left\|h\left[d\left(x_0+\Delta x\right)\right]-\phi_*\right\|_2^2 \tag{1}$$

Where $\phi_* = h\left[d\left(x_*\right)\right]$ represents the AMLBP feature value at the face mark in the template library, and Δx represents the iteration step size. Training starts at x_0 and converges at x_*. To use the gradient method to differentiate, Taylor (1) is Taylor-expanded at x_0:

$$f\left(x_0 + \Delta x\right) \approx f\left(x_0\right) + J_f\left(x_0\right)^T + \frac{1}{2}\Delta x^T H\left(x_0\right)\Delta x \tag{2}$$

Where $J_f\left(x_0\right)$ and $H\left(x_0\right)$ are Jacobian and Hessian matrices. Find the partial derivative of Δx in Equation (2) and make its derivative zero, and then get the first update of x:

$$\Delta x_1 = -H^{-1}J_f = -2H^{-1}J_h^T\left(\phi_0 - \phi_*\right) \tag{3}$$

In the first gradient iteration process, the above formula is regarded as the projection of $\Delta\phi = \phi_0 - \phi_*$ on the matrix $R_0 = -2H^{-1}J_h^T$. In order to avoid calculating the Hessian matrix and the Jacobian matrix, take R_0 as the gradient direction, and R_0 can be directly obtained by learning the linear regression between $\Delta x_* = x_* - x_0$ and $\Delta\phi_0$. After simplifying equation (3), we can get:

$$\Delta x_1 = R_0\phi_0 + b_0 \tag{4}$$

Where b_0 is the offset. For a specific image, use Newton's method to update along the gradient direction:

$$x_k = x_{k-1} - 2H^{-1}J_h^T\left(\phi_{k-1} - \phi_*\right) \tag{5}$$

Equation (5) uses the R_{k-1} and b_{k-1} learned in the previous step to determine the new iteration position x_k:

$$x_k = x_{k-1} + R_{k-1}\phi_{k-1} + b_{k-1} \tag{6}$$

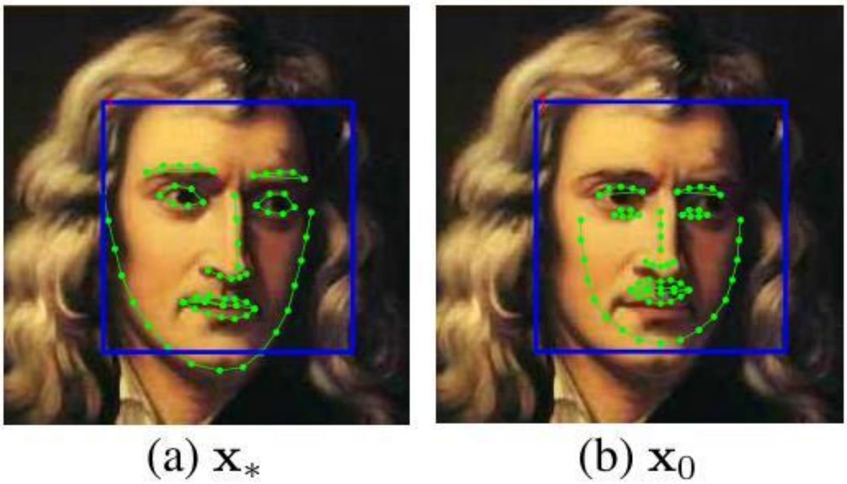

(a) $\mathbf{x}_*$ (b) $\mathbf{x}_0$

Fig. 1. a is a manually labelled face, b is a mark initialized using a face detector.

When a human face is detected, it can be positioned according to the gradient direction. The positioning effect is shown in Figure 2.

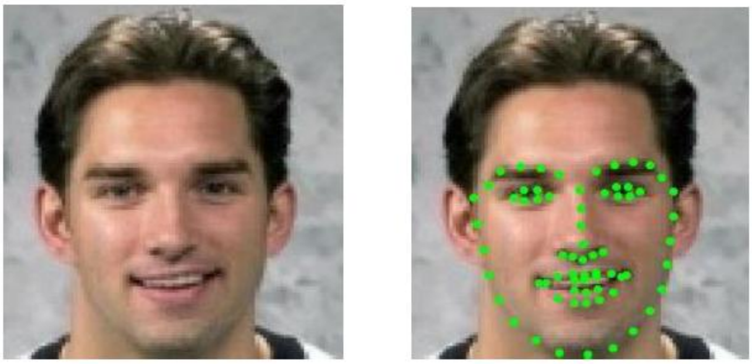


Fig. 2. a is the original face, b is the face after feature localization.

3.2. Adaptive Scale Feature Extraction

Taking into account the different amounts of information contained in different feature parts of the face, for example: the features contained in the eyes are the most sensitive and contribute the most to face recognition. Therefore, the feature extraction scale of this part should be the largest when feature extraction is performed; It is small, so its feature extraction scale should be appropriately reduced, so as to achieve the effect of reducing the feature dimension without reducing the amount of information. Based on this, this paper proposes an adaptive scale feature extraction method, which can adaptively adjust the feature extraction scale according to feature sensitivity.

In order to obtain the sensitivity of the facial features, the facial features trained by Adaboost [20] and the features extracted by the adaptive scale are mapped, and then based on the error rate of each weak classifier (this is regarded as sensitivity) to determine each feature extraction scale. Figure 3 shows the experimental results, from which it can be analyzed that the classification effect is best when the feature extraction scale is greater than 4.

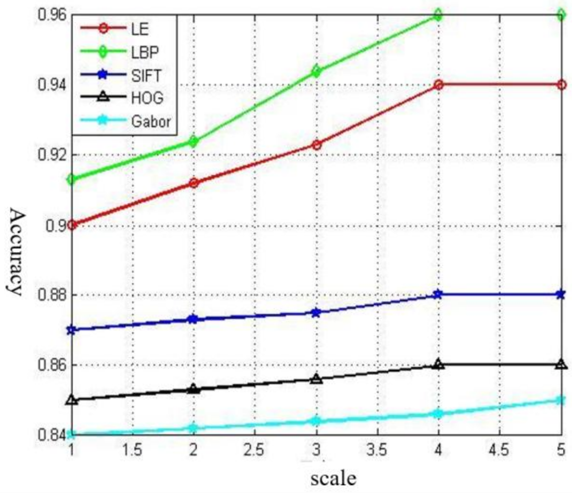


Fig. 3. Comparison of feature recognition results extracted at different scales.

Based on the experimental results in Figure 3, the maximum scale of adaptive feature extraction is set to 4, and the 200 facial features trained by the Adaboost algorithm are divided into four categories $K = \{k_1, k_2, k_3, k_4\}$, and the features are ranked from low to high according to the error rate. The features extracted by the adaptive scale are classified into four categories. The classification method is: use the Euclidean distance to search for features at the location of the feature to be classified, and the category of the feature closest to it is the category to which this feature belongs.

$$\min_{k(h(x))} D\left(p\{h(x)\}\right), p\{f(x)\}, k\{h(x)\} = k\{f(x)\} \qquad (7)$$

Face classification is performed using a Gaussian process combined with a spectral mixed kernel function. The feature extraction scales are adaptive scale and single scale. The results are shown in Figure 4, which can be analyzed from the graph. The accuracy will also increase, but its feature dimensions will also increase, and the computing efficiency will decrease. The classification accuracy obtained using adaptive scale is the same as that of $k = 4$ and $k = 5$. The average accuracy is 2.6, which significantly reduces the feature dimension. Table 1 is the adaptive scale mapping, in which the feature mark numbers correspond to the face feature mark numbers in FIG. 5.

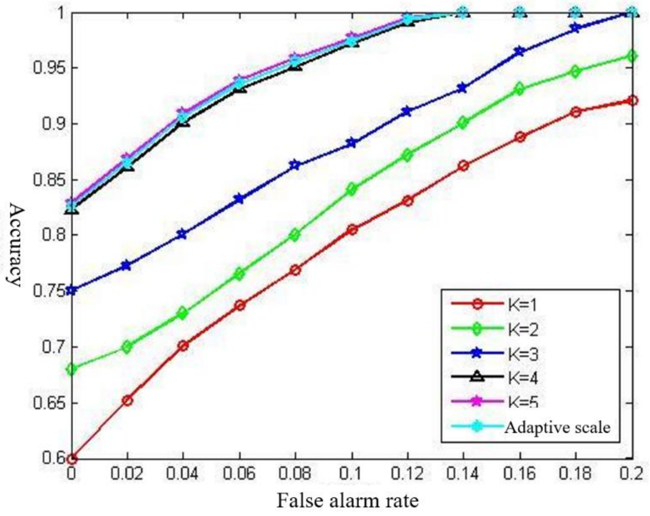


Fig. 4. Feature extraction effect of adaptive scale and single scale.

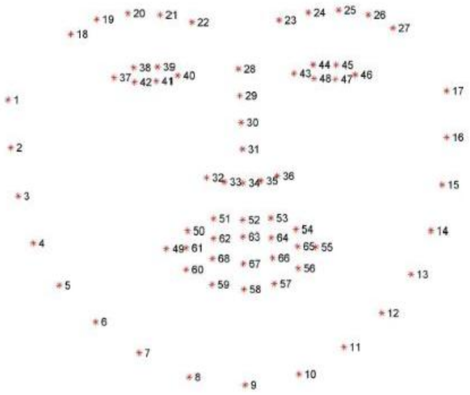


Fig. 5. Face feature labeling.

Table 1. Feature extraction scale mapping

Number of feature points	Feature label	scale
17	1-17	1
10	18-27	3
4	28-31	4
5	32-36	3
12	37-48	4
20	49-68	2

4. Convolutional Neural Network

The research on Convolutional Neural Networks (CNN) began in 1962. After Hubel and Wiesel proposed the receptive field by studying the visual cortex of cats, in 1984, Fukushima proposed the concept of cognitive machines based on receptive fields. Machine is the first application of receptive field on neural network, which can be regarded as the first implementation of convolutional neural network. After many scholars in the later period of innovation, convolutional neural network has become the most representative network in deep learning. Convolutional neural networks have achieved great success in the image field, and CNN models have achieved excellent results in visual competitions based on the ImageNet [21] dataset over the years. Compared with traditional vision algorithms, CNN has the advantage of directly extracting features from the original image without pre-processing the image [22], thereby avoiding the problem of image information loss during the pre-processing stage.

Figure 6 shows the convolutional neural network structure. The network structure has 5 convolutional layers and 3 fully-linked layers. The final output layer uses the Softmax function to output 1000 classes. Equation 8 is the Softmax principle, with K output categories. The calculation process is:

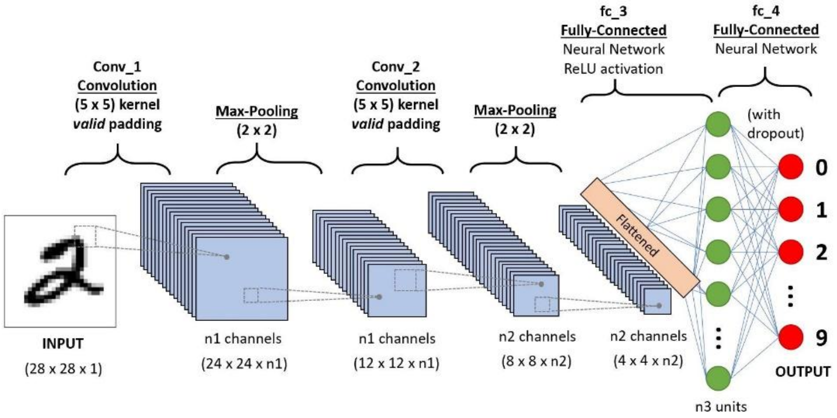


Fig. 6. Development stages of convolutional neural networks.

$$soft\max\left(a_i\right)=\frac{\exp\left(a_i\right)}{\sum_j \exp\left(a_j\right)},\ j=0,1,2,\ldots,k-1 \tag{8}$$

Softmax maps the output of multiple neurons to the [0,1] interval when performing multi-classification, so it can be understood as a probability. The output of Softmax is equivalent to the probability distribution of the picture classified into each category. This function is a monotonically increasing function. The larger the input value, the greater the probability that the input image belongs to the category label.

$$b_{x,y}^{i} = a_{x,y}^{i} / \left(k + a \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} \left(a_{x,y}^{j} \right)^{2} \right) \tag{9}$$

In the formula, $a_{x,y}^{i}$ is the i-th convolution of the feature at the (x, y) position in the input feature, and then the result is obtained through ReLU. $b_{x,y}^{i}$ is the corresponding normalized result. N is the total number of convolutions, and $k = 2, n = 5, \alpha = 10^{-4}, \beta = 0.75$ is the hyperparameter.

The network parameters are constantly adjusted when training the network. The changes in the network parameters at each layer will cause the input feature distribution of the latter layer to change. This phenomenon is called internal covariance change. However, it is necessary to adapt the parameters of each layer to the input when learning. Feature distribution. To this end, the data needs to be normalized. The purpose of normalization is to make the data mean 0 and unitized variance. The expression is as follows:

$$\widehat{x}^{(k)} = \frac{x^{(k)} - E\left(x^{(k)}\right)}{\sqrt{Var\left[x^{(k)}\right]}} \tag{10}$$

However, this method will reduce the expressive ability of the convolution layer. For example: using sigmoid as the activation function. This method will limit the data to around 0 mean. Then, only the linear part of the activation function is used and the function of the non-linear part is not used. Makes the network expression ability poor. Therefore, the normalized data needs to be further processed to maintain the expressive power of the model. Its formula is as follows:

$$y^{(k)} = \gamma^{(k)} \widehat{x}^{(k)} + \beta^{(k)} \tag{11}$$

In theory, all the data must be processed every time, but the amount of data processed by the convolutional neural network is huge. Each time all the data is processed will significantly increase the amount of calculation, so the data is processed in batches. Let each batch of input data be: $B = \{x_1, x_2, ..., x_m\}$. The output is: $y_i = BN_{\gamma,\beta}(x_i)$.

$$\begin{aligned} \mu_B &= \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x_i - \mu_B\right)^2 \\ \widehat{x}_i &= \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \qquad y_i = \gamma x_i + \beta = BN_{\gamma,\beta}(x_i) \end{aligned} \tag{12}$$

In order to extract signal features more comprehensively, the research direction of deep learning from 2014 has mainly focused on building deeper network structures, but increasing the model will reduce the network computing efficiency. Making use of computing performance has become the focus of research.

5. Face Recognition For Deep Learning Transfer Components

Since AlexNet won the championship in the ImageNet competition in 2012, deep learning has received unprecedented high attention in the field of machine learning. In recent years, a large number of papers on deep learning have emerged, but until now, deep learning is still a black box. You can't feel it, and you can't explain and deduce it with theory. Because CNN has a good hierarchical structure, this article analyzes its mobility by using CNN's feature extraction rules. The regularity of CNN's face feature extraction is shown in Figure 7.

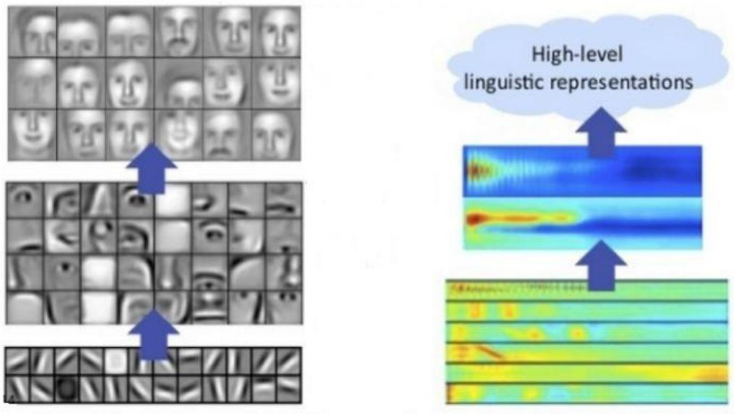


Fig. 7. Neural network for facial feature extraction.

This article trains the AlexNet network on the Pytorch framework, using the ImageNet training set, which has a total of 1000 classes. In the experiment, it was divided into A and B, each of 500 types. The AlexNet network has a total of 8 layers. Except that the 8th layer is a classification layer, this article analyzes the layers 1 to 7 layer by layer. The analysis method is: take the data of category B as the reference standard, fix the first *n* layers of network A, then initialize the remaining 8-*n* layers, and then classify B. This process is called AnB. The corresponding BnB is to fix the first n layers of the trained B network, initialize the remaining 8-n layers, and then classify the type B data. This process is called BnB. AnB and BnB network structure is shown in Figure 8.

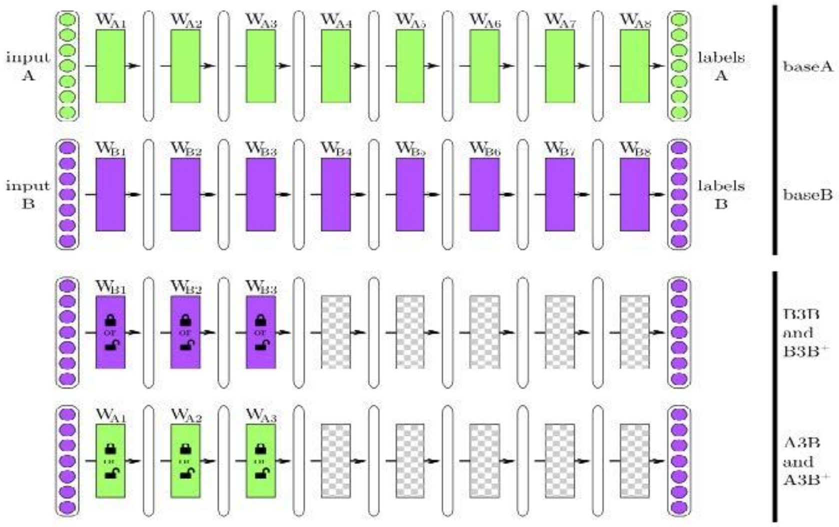


Fig. 8 AnB and BnB network structure.

Figure 9 shows the results of AnB and BnB experiments. From the figure, it can be analyzed that for BnB, the first 3 layers of the trained model are tested and there will be no loss of model accuracy. At the 4th and 5th layers, the accuracy is reduced but It's not too bad. The accuracy has obviously improved by the sixth and seventh layers. The reason is that the fourth and fifth layers are in the back part of the network. The extracted features are relatively specific, so the accuracy will decrease when the training samples change. Layers 6 and 7 learn again on the basis of abstract features to improve accuracy, which is exactly the effect required for transfer learning. For BnB + (BnB plus fine-tune), the whole result is not changed, which shows that fine-tune can promote the model well. The migration effect of AnB and AnB + is more convincing, because this is the migration of two different network-trained models. For AnB, the first 3 layers of network A are migrated to network B, but its accuracy is not affected. This proves once again that the first three layers of the network have learned abstract features, and the accuracy has begun to decline when it migrates to the fourth to fifth layers, which shows that the deeper the network, the more specific the extracted features. However, the accuracy of the 6th to 7th layers has decreased after a slight improvement. This is because the features of the 6th to 7th layers are not updated, the learning ability is poor, and the features extracted from these layers are the most specific. With the addition of fine-tune to the AnB + network, all layers performed very well and even exceeded baseB.

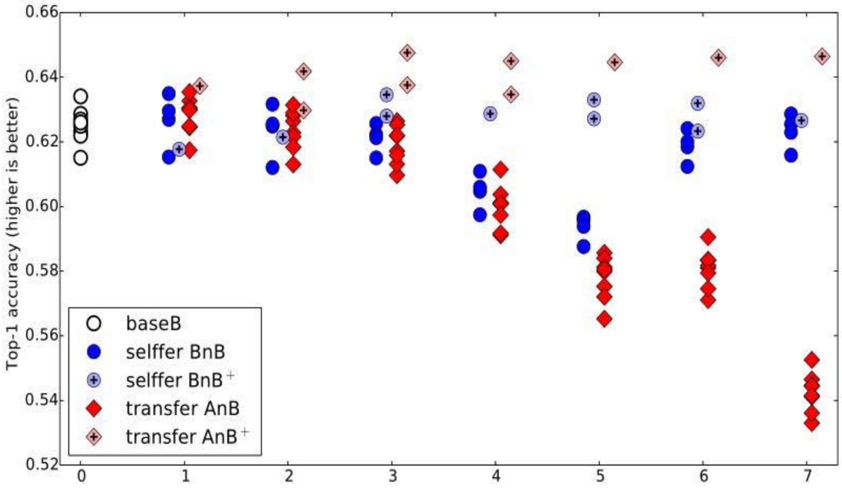


Fig. 9. AnB and BnB experimental results.

In order to exclude the accidental grouping of ImageNet data, that is, the data in the two groups are similar, for example, there are dogs in class A and dogs in class B. Then this will cause B to get better results when migrating A, Re-classify to ensure that there are no similar pictures in groups A and B, and repeat the above experiment to achieve the same results as the previous one. Figure 10 shows the experimental results.

It can be concluded from Figure 10 that the deeper the CNN network, the more specific the features learned, resulting in the performance of the model decreasing with the deepening of the network layer where the features are migrated, so the previous abstract feature layer is more suitable for migration. In addition, adding fine-tune to the transfer learning overcomes the differences between the data and can significantly improve the accuracy and help network migration.

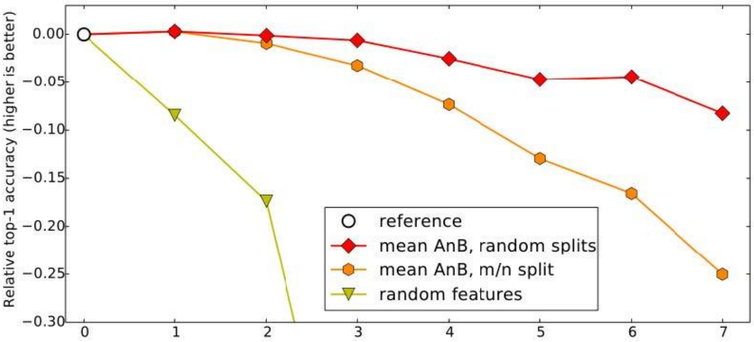


Fig. 10. AnB and BnB experimental results after removing similarity of samples.

Through experiments comparing the effects of network migration in the middle of the three frames, it was found that the optical face features extracted on the VGGFace framework and the sketch face features are best integrated. Therefore, this article divides the VGG16 network into the first, middle, and last three parts. The first part is the first to fourth layers of the network, the middle part is the fifth to tenth layers of the network, and the second part is the eleventh to sixteenth layers of the network. The facial features correspond to sketch faces, and finally they are adapted using JDA. Finally, the sketch faces and corresponding optical faces are used to train and test the model.

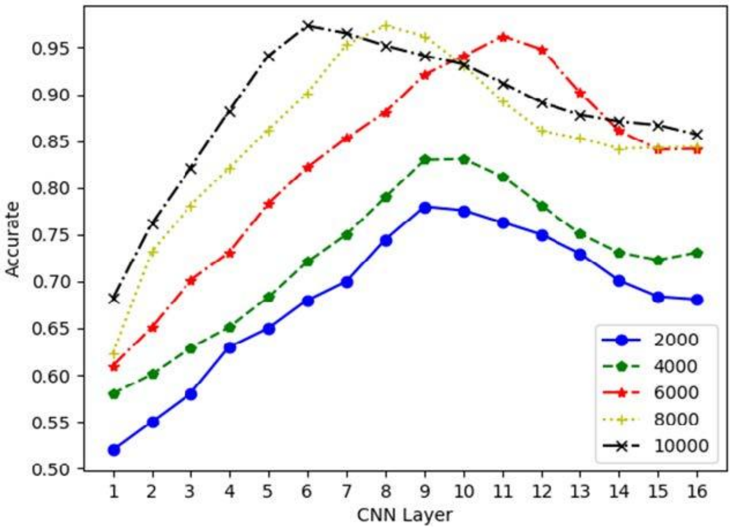


Fig. 11. The effect of the number of sketch face samples on the accuracy of the model.

From the experimental results shown in Figure 11, it can be seen that the features extracted in the middle part of the network are best suited for sketching the face. It is verified again that the face features extracted in the middle part of the CNN are most suitable for migration. The main role of this figure is that it proves that By extracting the optical face features from the CNN and adapting them to the sketch faces, it can effectively reduce the training samples of the sketch faces. Comparing Fig. 11, it can be seen that when the optical face features are not extracted from the CNN, 10,000 sketch faces are trained with the highest accuracy. It can reach 78.2%. After extracting the optical face features from CNN to match the sketch face, the accuracy of 2,000 sketch faces can reach 75.1%, and the accuracy can reach more than 82.3% when the number of sketch faces is 4,000. When the number of sketched faces reaches 6,000, the accuracy can reach 95.2%. When the number of sketched faces continues to increase to 8,000 and 10,000, the accuracy increase is not obvious, and the highest can only reach 97.4%, which indicates the success of this paper. The problem that the accuracy of the sketch face cannot be increased due to insufficient training samples for the sketch face is solved.

6. Conclusion

In this paper, an adaptive scale feature extraction method is proposed to build the sketch face training sample. Optical face features are extracted from the central network layer of CNN. Good results have been achieved through JDA [23] and sketch face adaptation. The performance of this method is analyzed through test results. It is found that the use of JDA + CNN can make sketch face recognition accuracy. It reaches about 97.4%, and the accuracy has been slightly improved compared with the traditional sketch face recognition algorithm. Its outstanding performance is reflected in the fact that it can effectively reduce the sketch face training samples. By analyzing the role of each layer of the convolutional neural network, reveal the basic principle of convolutional neural network, point out the direction for transfer learning.

Acknowledgment. This research work was supported in part by National Key R&D Program of China (2019YFC1520500,2020YFC1523004).

References

1. Masi, Iacopo, et al. Deep face recognition: A survey.2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). IEEE, 471-478. (2018)
2. Yang, Lu, et al. "Finger vein recognition with anatomy structure analysis." IEEE Transactions on Circuits and Systems for Video Technology, 1892-1905. (2017)
3. X. Tang and X. Wang, Face sketch synthesis and recognition, Ninth IEEE International Conference on Computer Vision, 687–694, (2003).
4. R. U. Jr and N. d. V. Lobo, A framework for recognizing a facial image from a police sketch, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 586–593, (1996).
5. B. Klare and A. K. Jain, "Sketch to Photo Matching: A Feature-based Approach, " Proc. SPIE Conference on Biometric Technology for Human Identification VII, (2010)
6. B. Klare, Z. Li, and A. K. Jain, Matching forensic sketches to mug shot photos, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 3, 639–646,(2011).
7. H. K. Galoogahi and T. Sim, Inter-modality face sketch recognition, IEEE International Conference on Multimedia and Expo, 224–229,(2012).
8. H. Kiani Galoogahi and T. Sim, Face sketch recognition by Local Radon Binary Pattern: LRBP, Proceedings - International Conference on Image Processing, ICIP, 1837–1840, (2012).
9. S. Setumin and S. A. Suandi, "Difference of gaussian oriented gradient histogram for face sketch to photo matching, " IEEE Access, vol. 6,39344–39352, (2018).
10. S. Setumin and S. A. Suandi, Cascaded static and dynamic local feature extractions for face sketch to photo matching, IEEE Access, vol. 7, 27135–27145, (2019).
11. P. Mittal, M. Vatsa, and R. Singh, Composite sketch recognition via deep network – A transfer learning approach, in Proc. Int. Conf. Biometrics, 251–256. (2015).
12. L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang, End-to-end photo-sketch generation via fully convolutional representation learning, in Proc. ACM Int. Conf. Multimedia Retrieval, New York, NY, USA. 627–634.(2015)
13. S. Saxena and J. Verbeek, Heterogeneous face recognition with CNNs, in Proc. Eur. Conf. Comput. Vis. 483–491. (2016)

14. D. Zhang, L. Lin, T. Chen, X. Wu, W. Tan, and E. Izquierdo, Content adaptive sketch portrait generation by decompositional representation learning, IEEE Trans. Image Process., vol. 26, no. 1, 328–339, (2017) .
15. C. Peng, N. Wang, X. Gao, and J. Li, Face Recognition from Multiple Stylistic Sketches: Scenarios, Datasets, and Evaluation. Cham, Germany:Springer, 3–18. (2016)
16. X. Gao, N. Wang, D. Tao, and X. Li,Face sketch-photo synthesis and retrieval using sparse representation IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 8, 1213–1226, (2018).
17. Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." computer vision and pattern recognition: 886-893. (2005)
18. Zhang, C., Zhao, X., Cai, M., Wang, D., Cao, L.: A New Model for Predicting the Attributes of Suspects. Computer Science and Information Systems, Vol. 17, No. 3, 705-715. (2020),.
19. Chen, H., Dai, Y., Gao, H., Han, D., Li, S.: Classification and Analysis of MOOCs Learner's State: The Study of Hidden Markov Model. Computer Science and Information Systems, Vol. 16, No. 3, 849–865. (2019),
20. Shuai Di, Honggang Zhang, Chun-Guang Li, Xue Mei, Danil Prokhorov, & Haibin Ling. (2018) 'Cross-domain traffic scene understanding: a dense correspondence-based transfer learning approach.' IEEE Transactions on Intelligent Transportation Systems, Vol. 19 No.3, pp.45-757. (2017)
21. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., & Ma, S., et al.. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252. (2015)
22. Zhang, D., et al., The generative adversarial networks and its application in machine vision. Enterprise Information Systems,: p. 1-21. (2019)
23. Chen, D., Ren, S., Wei, Y., Cao, X., & Sun, J.. Joint Cascade Face Detection and Alignment. European Conference on Computer Vision. 109-122. (2014)

Zhongkui Fan receive the BS degree in Computer engineering from Huanghe science and technology College, and the MS degree in Computer Engineering from Jiangxi University of Science and Technology. He is currently working toward the PhD degree in the School of Communication and Information Engineering with Shanghai University. His research interests include computer vision and machine learning.

Ye-peng Guan was born in Xiaogan, Hubei Province, China, in 1967. He received the B.S. and M.S. degrees in physical geography from the Central South University, Changsha, China, in 1990, 2006, respectively, and the Ph.D. degree in geodetection and information technology from the Central South University, Changsha, China, in 2000. From 2001 to 2002, he did his first postdoctoral research at Southeast University in electronic science and technology. From 2003 to 2004, he did his second postdoctoral research at Zhejiang University in communication engineering, and he had been an Assistant Professor with the Department of Information and Electronics Engineering, Zhejiang University. Since 2007, he has been a Professor with School of Communication and Information Engineering, Shanghai University. He is the author of more than 120 articles, and more than 20 patents. His research interests include intelligent information perception, digital image processing, computer vision, and security surveillance and guard.

Received: September 22, 2020; Accepted: April 04, 2021.

Computer Science and Information Systems 18(4):1411–1426 https://doi.org/10.2298/CSIS200930029W

Background Modeling from Video Sequences via Online Motion-Aware RPCA

Xu Weiyao[1], Xia Ting[1], and Jing Changqiang[2]*

[1] Zaozhuang University
Zaozhuang 277160, Shandong Province, China
{xuweiyao_2008,xiayuxue121}@126.com
[2] Linyi University
Linyi 276000, Shandong Province, China
jingchangqiang@lyu.edu.cn

Abstract. Background modeling of video frame sequences is a prerequisite for computer vision applications. Robust principal component analysis(RPCA), which aims to recover low rank matrix in applications of data mining and machine learning, has shown improved background modeling performance. Unfortunately, The traditional RPCA method considers the batch recovery of low rank matrix of all samples, which leads to higher storage cost. This paper proposes a novel online motion-aware RPCA algorithm, named OM-RPCAT, which adopt truncated nuclear norm regularization as an approximation method for of low rank constraint. And then, Two methods are employed to obtain the motion estimation matrix, the optical flow and the frame selection, which are merged into the data items to separate the foreground and background. Finally, an efficient alternating optimization algorithm is designed in an online manner. Experimental evaluations of challenging sequences demonstrate promising results over state-of-the-art methods in online application.

Keywords: Computer vision, Background modeling, Online RPCA, Truncated nuclear norm.

1. Introduction

Background modeling aims to extract foreground objects from videos, which has been widely used in many fields, such as object detection [46][29][18], object localization [33], and image alignment [28]. It aims to initialize efficient and accurate background modeling from a series of video frames.

Many background modeling methods have been proposed in the past few years. Recently, RPCA [2][17][4]has attracted wide attention in the fields of video surveillance and computer vision, which are based on decomposition of the matrix into sparse and low rank components, and has shown improved performance in background modeling[43]. RPCA decomposes the observed video frame matrix into background and foreground[7][16].

Let $Z \in R^{m \times n}$ be the observational data, which can be represented in matrix form. RPCA attempts to decompose Z as the sum of a sparse matrix F and a low rank matrix B.

$$\min\ \mathrm{rank(B)}+\lambda||\mathrm{F}||_0 \quad \mathrm{s.t.} \quad \mathrm{Z} = \mathrm{B} + \mathrm{F} \tag{1}$$

* Corresponding author

where $|| * ||_0$ denotes l_0 norm, i.e., the number of nonzero elements, which is the regularization term for promoting sparsity. λ is a regularization parameter. Unfortunately, due to the discontinuity of the rank function and the nonconvexity of l_0 norm , the optimization problem (1) is ordinarily NP-hard. Many researchers are seeking suitable alternatives to rank functions[8][6]. In addition, l_0 norm can convert problem (1) into a convex optimization problem. Wright et al. [2] proved that while the sparse matrix F is sufficiently sparse, the low-rank matrix Z can be recovered by solving the following problem:

$$\min \ ||B||_* + \lambda ||F||_1 \quad \text{s.t.} \quad \mathrm{Z = B + F} \tag{2}$$

where $||B||_*$ is the nuclear norm of B.

The problem (2) is a convex optimization problem, and many algorithms have been proposed for this problem. Unfortunately, most of the algorithms solve this problem in a batch manner. Because all samples loaded in memory during the optimization procedure may have high storage costs, it is especially unacceptable for large-scale sample sets. In addition, while all the samples are collected by the streamlined way, these algorithms cannot update the low dimensional subspace efficiently while a new sample is adding. Each iteration needs to optimize every frame, which may seriously limit the scalability of streaming video.

To solve these problems, the online method of RPCA has recently been proposed. Many online mode algorithms for background modeling have been pursued. Shen et al. [30] adopted max norm as a substitute of the rank function in problem (2) to solve the RPCA problem in an online method.

However, the background in dynamic sequences may include multiple motions, which makes the accurate modeling more challenging. In most previous methods, objects are not moving. Smearing artifacts would be introduced while dealing with slow motion and stationary foregrounds. To be aware of motion, when the background scenes change gradually, many RPCA methods exhibit degraded performance, such as under changing lighting conditions. Moreover, existing methods such as Zhou et al. [46] produced an overwhelming outlier in the low-rank component, when the background was heavily occluded by foreground objects. Javed et al. [15] created a background model by using a modified version of RPCA to generate a low-rank matrix from a set of matrices.

Recently, X. Ye [42] proposed a motion-assisted matrix restoration (RMAMR) model for the separation of background objects from the foreground objects, in which the dense motion fields were incorporated into the framework of the RPCA. J. Yang [40] proposed an online motion-assisted RPCA model for back ground recovery from video sequences. This method is more efficient for memory and is scalable for the long video sequences, which are weighted by motion information.

In Hu [11], a novel norm called the truncated nuclear norm was proposed. The new nuclear norm is subtracted from the sum of several maximal singular values, achieving better rank approximation. Based on this method, F. Cao [3] proposed a new algorithm which was called low-rank and sparse decomposition based on the truncated nuclear norm(LRSD-TNN). B. Hong [10] proposed a novel and online robust principal analysis algorithm via truncated nuclear norm regularization and designed an online optimization scheme in which the matrices were updated alternately.

W. Hu [12] analyzed the problem of mocap data completion based on the truncated nuclear norm. In order to reduce the redundant frames, a simple joint motion detection and frame selection operation was adopted[31]. Unfortunately, these methods deal with batch data only. In order to meet the needs of dynamic subspace, Hu et al.[5] Proposed an online optimization method to 60 deal with the static camera background scene in the video sequence of each sampling point.

Pan et al.[27] proposed a motion-assisted RPCA model based on matrix factorization, and designed an effective linear alternating direction multiplier method and matrix factorization algorithm to solve the proposed FM-RPCA 65 model. Hu et al.[13] Proposed a non-convex rank approximation RPCA model based on segmentation constraints. Firstly, the original video sequence is divided into three parts by low rank matrix decomposition. Then, a new non-convex function is proposed to constrain the low rank feature[9].

In this paper, we propose an online motion-aware RPCA with a truncated nuclear norm regularization (OM-RPCAT) framework for background modeling. The key idea is to extend the motion-aware low rank matrix approximation methods into an online model. In addition, to get better effect, we added a joint method of frame selection. Motion information from the video sequence is estimated by two methods in this paper. This method replaces the objective of rank minimization by minimizing the truncated kernel norm, which can be represented in a matrix factorization form. Then, we designed an efficient iterative optimization method for implementation [39][14].

The rest of this paper is organized as follows: In Section II, we present the OM-RPCAT scheme and two methods for obtaining the motion estimation matrix. In Section III, we design an efficient optimization algorithm to solve the optimization function. In Section IV, our algorithm is evaluated by experiments. Conclusions are made in Section V.

2. Background Modeling via OM-RPCAT

In this section, the OM-RPCAT model is proposed for background recovery. The main idea of our method is to make a more rigorous approximation to the rank operator and to exploit an online method for solving the optimization problem. Moreover, the background of recovery always suffers from smearing artifacts in areas covered by slow-moving objects. In order to overcome this defect, we combine the motion information and frame selection into the framework to separate the background from the moving objects[19][22][21].

2.1. The Proposed OM-RPCAT Model

In particular, let $Z \in R^{\mathrm{mxn}}$ be the observational data, with Z= $(\mathrm{z}_1,...,\mathrm{z}_n)$, and each z_i expresses a sample. Our goal is to decompose the matrix Z into the low-rank matrix and the sparse matrix. The traditional methods of recovering the two components B and F is solved by solving the following equation:

$$\min \frac{1}{2}||Z - B - F||_F^2 + \lambda_1||B||_* + \lambda_2||F||_1 \tag{3}$$

The truncated nuclear norm minimization is adopted as a more rigorous low-rank constraint on B. Therefore, the objective function for this method becomes:

$$\min \frac{1}{2}||Z - B - F||_F^2 + \lambda_1||B||_T + \lambda_2||F||_1 \tag{4}$$

where $|| * ||_F$ is the Frobenius norm, $|| * ||_T$ is the truncated nuclear norm, $|| * ||_1$ is the l_1 norm, λ_1 and λ_2 are the regularization parameters.

To solve the problem in (4), we usually use iterative optimization methods, such as the augmented Lagrangian multiplier [20] or the accelerated proximal gradient. Unfortunately, these methods are implemented in batch processing. Hence, huge data storage costs are incurred when large data are solved. To overcome this problem, we factorize B as $B=$ LR^T.

Given a matrix B, the relationship [11] between $||B||_T$ and $||B||_*$ is:

$$\begin{aligned} ||B||_T = ||B||_* - \max Tr(UBV^T) \\ ,UU^T = I, VV^T = I \end{aligned} \tag{5}$$

where $Tr(*)$ denotes the trace of the matrix and I stands for the identical matrix.

Then, the nuclear norm can be factorized as follows [30] :

$$||\mathrm{B}||_* = \min_{B=LR^T} \tfrac{1}{2}(||L||_F^2 + ||R||_F^2) \tag{6}$$

where $L \in R^{\mathrm{mx}d}$, $R \in R^{\mathrm{nx}d}$.

From this paper B. Hong [10] , we can obtain the following relationship:

$$\begin{aligned} ||B||_T &= ||B||_* - \sum_{i=1}^{T} \sigma_i(B) \\ &= \frac{1}{2}||L||_F^2 + \frac{1}{2}||R||_F^2 - Tr(UBV^T) \end{aligned} \tag{7}$$

Thus, the problem(4) can be transformed into the following constrained problem:

$$\begin{aligned} &\min \tfrac{1}{2}||Z - LR^T - F||_F^2 + \\ &\quad \lambda_1(\frac{1}{2}||L||_F^2 + \frac{1}{2}||R||_F^2 - Tr(ULR^TV^T)) + \lambda_2||F||_1 \\ &\quad \text{s.t.} UU^T = I, VV^T = I \end{aligned} \tag{8}$$

Our OM-RPCAT model is proposed to separate the foreground and background by joining motion information in the video sequences. We introduce the matrix W to represent motion information, $W \in [0, 1]$. The elements in the matrix W indicate whether the pixels in Z belong to the background. Therefore, the final form of the problem is

$$\begin{aligned} &\min \tfrac{1}{2}||W \circ (Z - LR^T - F)||_F^2 + \\ &\quad \lambda_1(\frac{1}{2}||L||_F^2 + \frac{1}{2}||R||_F^2 - Tr(MR^T)) + \lambda_2||F||_1 \\ &\quad \text{s.t.} UU^T = I, VV^T = I \end{aligned} \tag{9}$$

where $\circ$ denotes the Hadamard product, which is known as the element-wise product. Let M=V$^T UL$, and we use the fact that Tr(XYZ)=Tr(ZXY):

$$Tr(ULR^TV^T) = Tr(V^TULR^T) \tag{10}$$

For each sample z_i, we can get the approximate value of each sample under dictionary L through Lr_i + f_i. The problem (9) can be decomposed into the sample form:

$$\begin{aligned} &\min\tfrac{1}{2}||w_i \circ (z_i - Lr_i - f_i)||_2^2 + \\ &\lambda_1(\frac{1}{2}||L||_F^2 + \frac{1}{2}\sum_{i=1}^{n}||r_i||_2^2 - \sum_{i=1}^{n} m_i^T r_i) + \lambda_2 \sum_{i=1}^{n}||f_i||_1 \\ &\text{s.t.} UU^T = I, VV^T = I \end{aligned} \tag{11}$$

To simplify this problem, we introduce the $l(z_i, L)$ function:

$$\begin{aligned} l(z_i, L) = \min\frac{1}{2}||w_i \circ (z_i - Lr_i - f_i)||_2^2 + \\ \frac{\lambda_1}{2}||r_i||_2^2 - \lambda_1 m_i^T r_i + \lambda_2||f_i||_1 \end{aligned} \tag{12}$$

The problem (11) can be solved by minimizing the following loss function:

$$f_n(L) = \frac{1}{n}\sum_{k=1}^{n} l(z_i, L) + \frac{\lambda_1}{2n}||L||_F^2 \tag{13}$$

where z_i is the current frame, r_i represents the coefficient under the dictionary L, and w_i is the ith row of W. $l(z_i, L)$ is the loss function for each sample.

2.2. Motion Estimation

In this section, the motion matrix W is generated by computing the video sequence Z. We use two methods to construct the weighting matrix W. The first method is to adopt the optical flow [1], and the second method involves joining the motion detection and frame selection [31][34][35][36].

A. The first method involves using optical flow to obtain the motion estimation matrix.

Assume that z_i and z_{i+1} are the two consecutive frames of a sequence Z. Then, the horizontal V^x and vertical V^y motion vector components can be obtained by estimating the optical flow field. The motion map W is constructed as follows[44][45]:

$$w_{i,j} = \begin{cases} 0, if \sqrt{(v_{i,j}^x)^2 + (v_{i,j}^y)^2} \geq \tau \\ 1, otherwise \end{cases} \tag{14}$$

where τ is the threshold of motion magnitude according to the average intensity of the motion field determined experimentally. $v_{i,j}^y$ and $v_{i,j}^x$ are entries of V^y and V^x in the vertical motion fields and horizontal motion fields respectively.

B. The second method is to add the frame selection.

This scheme aims to reduce the number of redundant frames [31] . The index of the relevant frames is given as follows:

$$y_i = \begin{cases} 1, if |\hat{z} - \hat{\mu}| \geq \tau \\ 0, otherwise \end{cases} \tag{15}$$

where $\hat{\mu}$ denotes the mean value of the vector $\hat{z}$ which is gradiented and regularized for each frame [31]. τ controls the threshold. When the label y_i is 1, the corresponding frame will be selected. Due to the removal of invalid data frames, this method can restore the background of the video more accurately[23][24][26].

While the frame selection process is complete, the motion estimation of the selected frames is determined by:

$$w_k(i,j) = \begin{cases} 0, if \frac{1}{2}(D_k(i,j))^2 \geq \beta \\ 1, \quad\quad otherwise \end{cases} \tag{16}$$

Differing from this method [31] , β is the thresholding parameter, D_k is the difference between the two consecutive frames, and z_t denotes the t-th frame:

$$D_t = \sqrt{(Z_t - Z_{t-1})^2} \tag{17}$$

3. Optimization Method.

We propose an algorithm to minimize the average cost, which could solve our OM-RPCAT model in detail. The coefficient r_t, sparse error f_t, motion estimation matrix w_t, and basis L are optimized using alternative methods. The optimization procedure is divided into two steps[37][38][41]:

In the first step, we optimize coefficient r_t, sparse error f_t and motion estimation matrix w_t. The w_t has been illustrated in Section 2.2.

$$\begin{aligned} \{r_t, f_t\} = \arg\min \tfrac{1}{2} ||w_i \circ (z_t - L_{t-1} r - f)||_2^2 + \\ \lambda_1 (\frac{1}{2} ||r||_2^2 - m_t^T r) + \lambda_2 ||f||_1 \end{aligned} \tag{18}$$

where w_i denotes the motion map for the current frame z_t.

Update r_t: Given the current basis L, the coefficient r_t can be obtained by the following formula:

$$f(r) = \tfrac{1}{2} ||w_i \circ (z_t - L_{t-1} r - f_t^k)||_2^2 + \lambda_1 (\tfrac{1}{2} ||r||_2^2 - m_t^T r) \tag{19}$$

Let $\partial f / \partial r = 0$, and we can obtain the following solution:

$$\begin{aligned} & r_t^{k+1} = \\ & (L_{t-1}^T \hat{W}_t^T \hat{W}_t L_{t-1} + \lambda_1 I)^{-1} (L_{t-1}^T \hat{W}_t^T \hat{W}_t (z_t - f_t^k) + \lambda_1 m_t) \end{aligned} \tag{20}$$

where $\hat{W}$ is a diagonal matrix formed by placing the elements on the diagonal.

Update f_t: The objective formula to optimize f_t induced from problem(18) is:

$$g(f) = \tfrac{1}{2} ||w_i \circ (z_t - L_{t-1} r_t^{k+1} - f)||_2^2 + \lambda_2 ||f||_1 \tag{21}$$

We used the common approach presented in E. J. Candes [28] to solve this problem and obtained the following closed formula:

$$f_t^{k+1} = S_{\lambda_2} [w_i \circ (z_t - L_{t-1} r_t^{k+1})] \tag{22}$$

where $S_\tau[*]$ is a shrinkage operator, which is defined as

$$S_\tau[x] = sign(x)\max(|x| - \tau, 0) \tag{23}$$

In the second step, we optimize the basis matrices L_t, V_t, and U_t under the previously obtained r_k, f_k, and w_k. Based on problem(9), the objective function is derived as follows:

$$\begin{aligned} \{L_t, U_t, V_t\} = & \min \tfrac{1}{2}||W \circ (Z - LR^T - F)||_F^2 + \\ & \lambda_1(\frac{1}{2}||L||_F^2 - Tr(MR^T)) \end{aligned} \tag{24}$$

Because of the Hadamard product, it is difficult to calculate the matrix multiplication for this formula. We introduce an additional variable Y:

$$Y_k = Z_k - L_k R_k^T - F_k \tag{25}$$

The above equation is converted to the following equation:

$$\begin{cases} Y_k^{l+1} = \arg\min\limits_{Y} \frac{1}{2}||W_k \circ Y_k^l||_F^2 + \lambda_1(\frac{1}{2}||L_k^l||_F^2 - \\ Tr(M_k R_k^T)) + \frac{\lambda_3}{2}||Y_k^l - Z_k + L_k^l R_k^T + F_k||_F^2 \\ L_k^{l+1} = \arg\min\limits_{L} \frac{1}{2}||W_k \circ Y_k^{l+1}||_F^2 + \lambda_1(\frac{1}{2}||L_k^l||_F^2 - \\ Tr(M_k R_k^T)) + \frac{\lambda_3}{2}||Y_k^{l+1} - Z_k + L_k^l R_k^T + F_k||_F^2 \end{cases} \tag{26}$$

where λ_3 is a constrained parameter. After removing the irrelevant items, we can get the following:

Update Y:

$$\begin{aligned} Y_k^{l+1} = & \arg\min\limits_{Y} \tfrac{1}{2}||W_k \circ Y_k^l||_F^2 + \\ & \tfrac{\lambda_3}{2}||Y_k^l - Z_k + L_k^l R_k^T + F_k||_F^2 \end{aligned} \tag{27}$$

We can compute Y in a pixel-wise manner:

$$y_{ik} = \frac{\lambda_3}{\lambda_3 + w_{ik}^2}(z_{ik} - f_{ik} - l_i r_k) \tag{28}$$

where l_i denotes the ith row of matrix L.

Update L:

$$\begin{aligned} L_k^{l+1} = & \arg\min\limits_{L} \lambda_1(||\tfrac{1}{2}L_k^l||_F^2 - Tr(M_k R_k^T)) + \\ & \tfrac{\lambda_3}{2}||Y_k^{l+1} - Z_k + L_k^l R_k^T + F_k||_F^2 \end{aligned} \tag{29}$$

where $M = V^T UL \in R^{n \times d}$.

The matrix L can be updated via the closed-form solution of the least square problem in Eq.(29):

$$L_k = (\lambda_3(Z - F - Y)R + U^T VR)(\lambda_1 I + \lambda_3 R^T R) \tag{30}$$

U_t is optimized by the following formula [10]:

$$U_t = \arg\ \max \mathrm{Tr}(\mathrm{UL}_t \mathrm{R}_t^T \mathrm{V}_{t-1}^T) \text{ s.t. } \mathrm{UU}^T = I \tag{31}$$

Similarly, V_t is optimized by the following formula [10]:

$$V_t = \arg\ \max \mathrm{Tr}(\mathrm{VR}_t\mathrm{L}_t^T\mathrm{U}_t^T) \text{ s.t. } \mathrm{VV}^T = I \tag{32}$$

The algorithm is shown as Algorithm 1.

Algorithm 1: OM-RPCAT Algorithm

```
Input:
The observed data:$Z = [z_1, z_2, ..., z_n] \in R^{m \times n}$
matrix: $L_0 \in R^{m \times d}$, $U_0 \in R^{s \times m}$, $V_0 \in R^{s \times n}$
regularization parameters:$\lambda_1, \lambda_2, \lambda_3 \in$ R
motion matrix:$W \in R^{m \times n}$
number of frames: t
Output: $L_n, R_n$
for t = 1 to n do
    coefficient $r_t$ = 0
    sparse error $f_t$ = 0
    $m_t$ is is the t-th row of $V_{t-1}^T U_{t-1} L_{t-1}$
    compute motion matrix $w_t$ by Eq.(14) or Eq.(16)
    Stage 1: compute $r_t$ and $f_t$
    while not converged do
        Update the coefficient $r_t$ by Eq.(20).
        Update the sparse error $f_t$ by Eq.(22).
    end
    Stage 2: compute $L_t$, $U_t$ and $V_t$
    while not converged do
        Update the additional matrix $Y_t$ by Eq.(28).
        Update the matrix $L_t$ by Eq.(30).
        Update the $U_t$ by Eq.(31).
        Update the $V_t$ by Eq.(32).
    end
end
```

4. Experimental Results and Discussions

In this section, we investigate the performance of the proposed method. All the experiments were implemented on a PC with an Intel Core i5 CPU at 2.4 GHz and with 16 GB of memory. Simulations were performed using MATLAB 2014a.

Experiments are performed on scene background initialization (SBI) dataset [25][3]. The SBI dataset, consisting of fourteen image sequences with ground truth images and a set of commonly adopted metrics, with a wide range of complex backgrounds and different situations, including a variety of sequences, such as HallMonitor, Board, CAVIAR1, CaVignal and HumanBody2, and it has been adopted by many existing and new background initialization methods. For the convenience of comparison, the ground truth images of the background are also provided.

[3] http://sbmi2015.na.icar.cnr.it/

For detailed qualitative comparisons, the proposed method is evaluated by comparing it with the online RPCA method [32] , OMA-RPCA method [40] , and OTNNR method [10] . These methods are state-of-the-art techniques for online background modeling.

We implemented the following settings: $\lambda_1 = 0.1, \lambda_2 = 1, \lambda_3 = 0.1$, the convergence error is set to 10e-4, the rank of the low dimensional subspace is 5, the threshold τ is set by the average intensity of the optical flow. We implemented these four algorithms on the more challenging moving background video sequences, i.e., HallMonitor, Board, CAVIAR1, CaVignal and HumanBody2. Table 1 describes of the information of datasets.

Table 1. Brief introduction of the datasets

Dataset	Size	Number of frames
HallMonitor	352*240	299
Board	200*164	227
CAVIAR1	384*256	609
CaVignal	200*136	257
HumanBody2	320*240	740

4.1. Qualitative Results and Comparison

In this paper, we use two methods to get the motion matrix. The first method is to adopt the optical flow [1] , and the second method is to apply joint motion detection and frame selection [31]. For these two methods of obtaining the motion matrix, the results of our proposed method and the existing methods are shown in Fig. 1 and Fig. 2 respectively.

Due to the slow moving foreground takes up many areas of the frame, it used to be severely tailed by using the previous methods to recover the true background, such as online RPCA, OMA-RPCA and OTNNR. The slow moving foreground may be considered as part of the background. From the results as shown in Fig. 1 and Fig. 2, we can see that our method is the closest to the real background and is superior to other algorithms. Fig. 1 and Fig. 2 show the background modeling results of frame selection mode and optical flow mode respectively. Because of abandoning the redundant frames in the frame selection technology, the effect of background modeling is much better than that of optical flow methods. But for video object with long static foreground, the effect of the two methods are both not very good.

4.2. Quantitative Evaluations and Analysis

In order to quantitatively analyze the performance of the algorithm and verify the rationality of the algorithm, we measured these methods with the PSNR(Peak Signal-to-Noise Ratio) and RRE(Relative Reconstruction Error), which is defined as follows:

$$RRE = ||\hat{A} - A||_F / ||A||_F \tag{33}$$

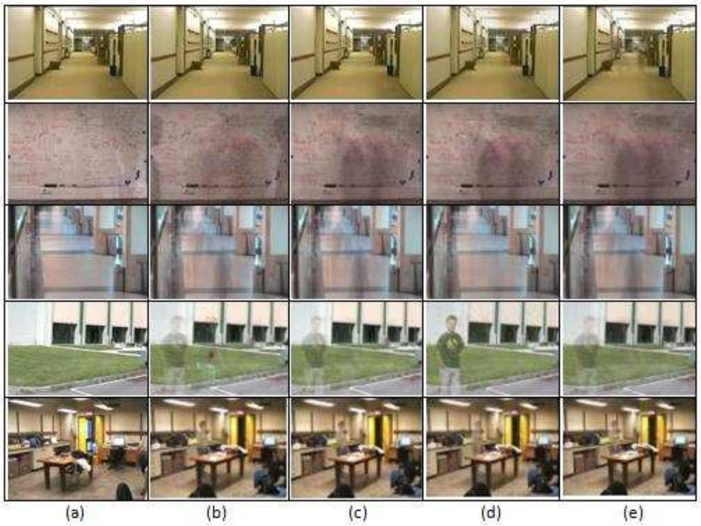


Fig. 1. Background modeling results by frame selection method: (a) true backgrounds, (b) OM-RPCAT(Ours), (c) OMA-RPCA, (d) online RPCA, (e) OTNNR. From top to bottom, the recovered backgrounds for HallMonitor, Board, CAVIAR1, CaVignal and HumanBody2 are presented respectively.

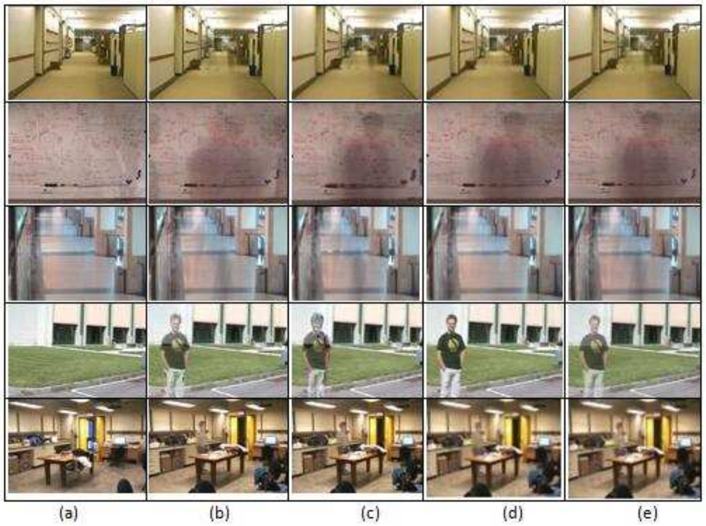


Fig. 2. Background modeling results by optical flow method:(a) true backgrounds, (b) OM-RPCAT(Ours), (c) OMA-RPCA, (d) online RPCA, (e) OTNNR. From top to bottom, the recovered backgrounds for HallMonitor, Board, CAVIAR1, CaVignal and HumanBody2 are presented, respectively.

$$PSNR = 10 * \log_{10}(255^2/MSE) \tag{34}$$

where the MSE is the mean square error of the real background and the reconstruction background.

Table 2. Quantitative background modeling results by frame selection

	PSNR	RRE
OM-RPCAT(OURS)	83.6	0.014
OMA-RPCA[40]	83.4	0.015
OTNNR[10]	75.8	0.021
Online RPCA[32]	76.1	0.019

Table 3. Quantitative background modeling results by optical flow

	PSNR	RRE
OM-RPCAT(OURS)	79.3	0.024
OMA-RPCA[40]	74.1	0.023
OTNNR[10]	75.7	0.027
Online RPCA[32]	76.3	0.020

In order to ensure the fairness of verification, we all use the HallMonitor dataset. Based on the results of Table 2 and Table 3, our method can get the best PSNR. In addition, the frame selection method obtain greater performance than optical flow.

4.3. Implementation Details and Computational Time

The solution of the proposed models require a set of parameters including λ_1, λ_2, λ_3, τ and d. The d represents the rank of the low dimensional subspace. In order to update the model quickly, in the previous experiment, we set the rank d to 5. Now we analyze the performance impact of rank d on background modeling, HallMonitor dataset was used in this experiment. The frame selection method is selected for the motion estimation.

It can be seen from Fig. 3 that with the change of rank d, the PSNR also changes. When the d is set to 5, the performance of background modeling is the best.

In the experiment, we also study the time complexity problem. To compare the overall computational time, we selected a short sequence named HallMonitor. For a fair comparison with the other methods, the time is recorded in seconds. The frame selection method is selected for the motion estimation. Fig. 4 presents the performance in terms of computational time. Compared with the previous algorithms, although our method does not achieve the most promising results, considering the final performance, our algorithm is better.

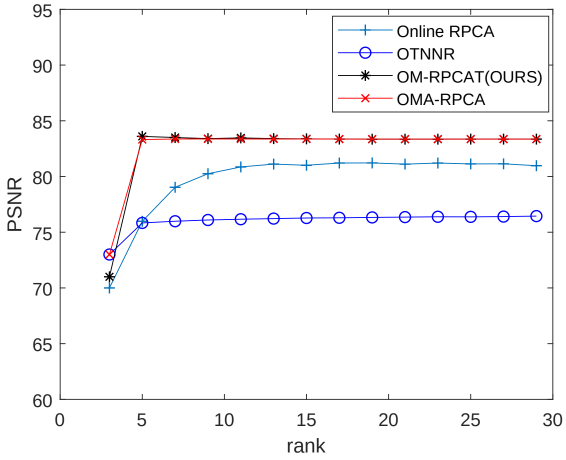


Fig. 3. PSNR analysis of the rank of the low dimensional subspace d

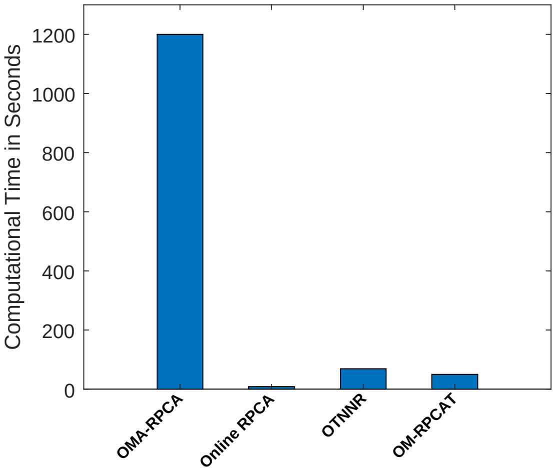


Fig. 4. Comparison of computational time in seconds

5. Conclusions

In this paper, we propose a novel model for background modeling from a given sequence of video frames and adopt the truncated nuclear norm as the convex optimization framework. In order to perceive motion information in video, we use motion information as the weighting matrix. In addition, the low-rank approximation is optimized to online mode, which is more efficient for the online video. Additionally, to achieve better effects and improve the processing efficiency, we introduced two methods to get the motion estimation matrix. The first method is to adopt the optical flow, and the second method involves joining the motion detection and frame selection which can use the frame selection technique to abandon the redundant frames. We further designed an online optimization scheme to solve the matrix decomposition problem with weighted matrix. Experimental results demonstrate that the proposed algorithm outperforms existing online RPCA algorithms significantly. Considering the limitations of RPCA, our future work will focus on designing more efficient background modeling algorithms.

Acknowledgments. This work is supported by the Foundation of Zaozhuang University (No. 102061808 and No. 102061804), in part by Key Research and Development Project of Shandong Province (No. 2019JZZY010134) and National Natural Science Foundation of China (No. 61901206).

References

1. Brox, T., Malik, J., Fellow, IEEE: Large displacement optical flow: Descriptor matching in variational motion estimation. IEEE Transactions on Pattern Analysis & Machine Intelligence 33(3), 500–513 (2011)
2. Candes, E.J., Xiaodong, L.I., Yl, M.A., Wright, J.: Robust principal component analysis? Journal of the ACM (JACM) 58(3) (2011)
3. Cao, F., Chen, J., Ye, H., Zhao, J., Zhou, Z.: Recovering low-rank and sparse matrix based on the truncated nuclear norm. Neural Networks 85 (2017)
4. Fu, H., Gao, Z., Liu, H.Z.: Fast robust pca on background modeling. In: Chinese Intelligent Systems Conference (2018)
5. Fu, H., Wang, B., Liu, H.Z.: Online RPCA on Background Modeling: Volume II (2019)
6. Garg, K., Ramakrishnan, N., Prakash, A., Srikanthan, T.: Rapid and robust background modeling technique for low-cost road traffic surveillance systems. IEEE Transactions on Intelligent Transportation Systems PP(99), 1–12 (2019)
7. Guangming, S., Tao, H., Weisheng, D., Jinjian, W., Xuemei, X.: Robust foreground estimation via structured gaussian scale mixture modeling. IEEE Transactions on Image Processing pp. 1–1 (2018)
8. Guian, Zhang, Zhiyong, Yuan, Qianqian, Tong, Qiong, Wang: A novel and practical scheme for resolving the quality of samples in background modeling. Sensors (2019)
9. Guo, Wenzhong, Zhang, Qishan, Chen, Yuzhong, Kun, Qiu, Qirong: Community discovery by propagating local and global information based on the mapreduce model. Information Sciences: An International Journal (2015)
10. Hong, B., Wei, L., Hu, Y., Cai, D., He, X.: Online robust principal component analysis via truncated nuclear norm regularization. Neurocomputing 175(JAN.29PT.A), 216–222 (2016)
11. Hu, Yao, Zhang, Debing, Ye, Jieping, Li, Xuelong: Fast and accurate matrix completion via truncated nuclear norm regularization. IEEE Transactions on Pattern Analysis & Machine Intelligence 35(9), 2117–2130 (2013)

12. Hu, W., Wang, Z., Liu, S., Yang, X., Yu, G., Zhang, J.J.: Motion capture data completion via truncated nuclear norm regularization. IEEE Signal Processing Letters pp. 1–1 (2017)
13. Hu, Z., Wang, Y., Su, R., Bian, X., Wei, H., He, G.: Moving object detection based on non-convex rpca with segmentation constraint. IEEE Access 8, 41026–41036 (2020)
14. Huang, Z., Yu, Y., Gu, J., Liu, H.: An efficient method for traffic sign recognition based on extreme learning machine. IEEE Transactions on Cybernetics 47(4), 920–933 (2016)
15. Javed, S., Mahmood, A., Bouwmans, T., Jung, S.K.: Motion-aware graph regularized rpca for background modeling of complex scenes. In: 2016 23rd International Conference on Pattern Recognition (ICPR) (2017)
16. Le, H., Wei, W., Li, G., Mathematics, S.O.: Compressed video background/foreground recovery and separation based on ptv-tv tensor modeling. Journal of South China University of Technology(Natural Science Edition) (2019)
17. Li, H., Miao, Z., Li, Y., Wang, J., Zhang, Y.: Background subtraction via online box constrained rpca. In: 2018 International Conference (2018)
18. Li, Y., Liu, G., Liu, Q., Sun, Y., Chen, S.: Moving object detection via segmentation and saliency constrained rpca. Neurocomputing 323 (2018)
19. Lin, B., Guo, W., Lin, X.: Online optimization scheduling for scientific workflows with deadline constraint on hybrid clouds. Concurrency and Computation: Practice & Experience (2016)
20. Lin, Z., Chen, M., Ma, Y.: The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices (2010)
21. Liu, GG, Huang, Guo, WZ, Niu, YZ, Chen, GL: Multilayer obstacle-avoiding x-architecture steiner minimal tree construction based on particle swarm optimization. IEEE T CYBERNETICS 2015,45(5)(-), 989–1002 (2015)
22. Liu, G., Guo, W., Niu, Y., Chen, G., Huang, X.: A pso-based timing-driven octilinear steiner tree algorithm forvlsi routing considering bend reduction. Soft Computing 19(5), 1153–1169 (2015)
23. Luo, F., Guo, W., Yu, Y., Chen, G.: A multi-label classificationalgorithm based on kernel extreme learning machine. Neurocomputing 260(oct.18), 313–320 (2017)
24. Ma, T., Liu, Q., Cao, J., Tian, Y., Al-Rodhaan, M.: Lgiem: Global and local node influence based community detection. Future Generation Computer Systems 105 (2019)
25. Maddalena, L., Petrosino, A.: Towards benchmarking scene background initialization. In: International Conference on Image Analysis and Processing (2015)
26. Niu, Y., Chen, J., Guo, W.: Meta-metric for saliency detection evaluation metrics based on application preference. Multimedia Tools & Applications 77(20), 1–19 (2018)
27. Pan, Peng, Wang, Yongli, Zhou, Mingyuan, Sun, Zhipeng, Guoping: Background recovery via motion-based robust principal component analysis with matrix factorization. Journal of Electronic Imaging 27(2), 23034.1 (2018)
28. Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. IEEE Transactions on Pattern Analysis & Machine Intelligence 34(11), 2233–46 (2012)
29. Sajid, Javed, Arif, Mahmood, Somaya, Al-Maadeed, Thierry, Bouwmans, Ki, S., Jung: Moving object detection in complex scene using spatiotemporal structured-sparse rpca. IEEE transactions on image processing : a publication of the IEEE Signal Processing Society (2018)
30. Shen, J., Xu, H., Li, P.: Online optimization for max-norm regularization. Machine Learning 2(3), 419–457 (2017)
31. Sobral, A., Bouwmans, T., Zahzah, E.H.: Comparison of matrix completion algorithms for background initialization in videos. In: Scene Background Modeling and Initialization (SBMI) Workshop in conjunction with ICIAP 2015 (2015)
32. Song, W., Zhu, J., Li, Y., Chen, C.: Image alignment by online robust pca via stochastic gradient descent. IEEE Transactions on Circuits & Systems for Video Technology 26(7), 1241–1250 (2016)

33. Stauffer, C.: Adaptive background mixture model for real-time tracking. Proc Cvpr 2, 2246 (1998)
34. Wang, J., Zhang, X.M., Lin, Y., Ge, X., Han, Q.L.: Event-triggered dissipative control for networked stochastic systems under non-uniform sampling. Information Sciences p. S0020025518301749 (2018)
35. Wang, S., Guo, W.: Robust co-clustering via dual local learning and high-order matrix factorization. Knowledge-Based Systems 138 (2017)
36. Wu, T.Y., Chen, C.M., Wang, K.H., Meng, C., Wang, E.K.: A provably secure certificateless public key encryption with keyword search. Journal of the Chinese Institute of Engineers pp. 1–9 (2019)
37. Xia, Y., Chen, T., Shan, J.: A novel iterative method for computing generalized inverse. Neural Computation 26(2), 449C465 (2014)
38. Xia, Y., Wang, J.: Low-dimensional recurrent neural network-based kalman filter for speech enhancement. Neural Networks 67, 131–139 (2015)
39. Xing, H., Guo, W., Liu, G., Chen, G.: Fh-oaos: A fast four-step heuristic for obstacle-avoiding octilinear steiner tree construction. Acm Transactions on Design Automation of Electronic Systems 21(3), 1–31 (2016)
40. Yang, J., Yang, J., Yang, X., Yue, H.: Background recovery from video sequences via online motion-assisted rpca. In: 2016 Visual Communications and Image Processing (VCIP) (2017)
41. Yang, L.H., Yang, L. H., K.S.: Multi-attribute search framework for optimizing extended belief rule-based systems. Information Sciences (2016)
42. Ye, X., Yang, J., Sun, X., Li, K., Hou, C., Wang, Y.: Foreground-background separation from video clips via motion-assisted matrix restoration. IEEE Transactions on Circuits & Systems for Video Technology 25(11), 1721–1734 (2015)
43. Yinxiao, Zhan, Ting, Liu: Weighted rpca based background subtraction for automatic berthing (2019)
44. Zhang, S., Xia, Y.: Two fast complex-valued algorithms for solving complex quadratic programming problems. IEEE Transactions on Cybernetics 46(12), 2837–2847 (2015)
45. Zhang, S., Xia, Y., Wang, J.: A complex-valued projection neural network for constrained optimization of real functions in complex variables. IEEE Transactions on Neural Networks & Learning Systems 26(12), 3227–3238 (2015)
46. Zhou, X., Yang, C., Yu, W.: Moving object detection by detecting contiguous outliers in the low-rank representation. IEEE Transactions on Pattern Analysis & Machine Intelligence 35(3), 597–610 (2013)

Xu Weiyao received the master's degree in communication and information systems from Nanjing University of Posts and Telecommunications. He is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications. His current research interests include around machine learning and human action recognition.

Xia Ting received the master's degree in Communication and Information Systems from Hangzhou University of Electronic Science and Technology. She is currently a lecturer at Zaozhuang University. Her current research interests include machine learning and human action recognition.

Jing Changqiang received Ph.D. degrees from Kwangwoon University, Korea in 2015. His research interests include computer vision, and wireless sensor network.

Received: September 30, 2020; Accepted: April 18, 2021.

Computer Science and Information Systems 18(4):1427–1444 https://doi.org/10.2298/CSIS200909030C

A Novel Network Aligner for the Analysis of Multiple Protein-protein Interaction Networks

Jing Chen[1,2,*] and Jia Huang[1]

[1] School of Artificial Intelligence and Computer Science,
Jiangnan University Wuxi, China
chenjing@jiangnan.edu.cn
[2] Jiangsu Provincial Engineering Laboratory of Pattern
Recognition and Computing Intelligence, Jiangnan University, Wuxi, China
6181914004@stu.jiangnan.edu.cn

Abstract. The analysis of protein-protein interaction networks can transfer the knowledge of well-studied biological functions to functions that are not yet adequately investigated by constructing networks and extracting similar network structures in different species. Multiple network alignment can be used to find similar regions among multiple networks. In this paper, we introduce Accurate Combined Clustering Multiple Network Alignment (ACCMNA), which is a new and accurate multiple network alignment algorithm. It uses both topology and sequence similarity information. First, the importance of all the nodes is calculated according to the network structures. Second, the seed-and-extend framework is used to conduct an iterative search. In each iteration, a clustering method is combined to generate the alignment. Extensive experimental results show that ACCMNA outperformed the state-of-the-art algorithms in producing functionally consistent and topological conservation alignments within an acceptable running time.

Keywords: graph data analysis, big data, protein-protein interaction network, network clustering, seed-and-extend strategy.

1. Introduction

Great progress has been made in constructing large amounts of biological networks of different species using high-throughput experimental techniques and computational predictions. In recent years, obtaining information on cell composition and function by analyzing network data has gradually become a popular research topic. In protein-protein interaction networks (PPINs), proteins are represented by nodes and interactions between two proteins by edges between two nodes. The alignment is usually generated according to the topological structure and sequence similarity information of the protein-protein interaction network. The topology of the network can extract much of the hidden information in the network [23], [26], which can be used for network research on different data. Functionally homogeneous proteins and protein complexes in different species can be discovered through network alignment, which is divided into pairwise network alignment (PNA) and multiple (*i.e.*, three or more) network alignment (MNA) according to the number of aligned networks. The purpose of PNA is the creation of node mappings

[*] Corresponding author

between two networks. In addition to finding a mapping between multiple networks, the MNA can obtain correlation information of different species simultaneously. Therefore, a well-studied MNA can provide deeper network insight. With respect to the mapping types, network alignment algorithms are divided into one-to-one, one-to-many and many-to-many alignment algorithms. In one-to-one alignment, there is exactly one node from each aligned network, and not every node is required to be mapped; in one-to-many alignment, which is usually used in metabolic network alignments, one metabolic path can be mapped to another subset; in many-to-many alignment, there can be one or more nodes from the same network in each alignment cluster. Network alignment types are classified into local network alignment (LNA) and global network alignment (GNA). The purpose of LNA is to find highly conserved, unrelated sub-networks with a highly similar structure among the input network. However, LNAs only consider the similarity of local structures, which may lead to conflicts or ambiguities. On the other hand, GNA aims to construct node mappings between the overall nodes from the input networks with the cost of sub-optimal conservation in the local area and finally obtain a network with a larger coverage, which can produce a more consistent alignment compared to LNA.

Network alignment can be regarded as a subgraph isomorphism problem; however, subgraph isomorphism is an NP-hard complete problem [7],which makes it very difficult to find the network alignment solution. Heuristic algorithms are usually used for the solution of NP-hard problems. They have the ability of intelligence, generality and global search, which make them applicable in many fields, such as the cutting problem [44], image analysis [5], the graph matching problem and so on. Therefore, heuristics alignment algorithms are used to address the issue that the computational difficulty of network alignment increases exponentially with the increase of input network size.

The two most important aspects in evaluating network alignment results are network topology and biological consistency. Nevertheless, achieving high topological conservation while obtaining biological significance is often contradictory in present literature, even though they are both vital goals of network alignment [12]. The present study attempted to solve the problem of balancing network topological conservation and functional consistency. Moreover, the ACCMNA algorithm proposed in this paper introduces a network clustering method as a solution to the problem of network alignment. The MNA algorithm gathers many similar nodes in the same cluster and is, therefore, similar to a clustering algorithm.

The multiple network aligner ACCMNA is proposed in this paper, which could match as many consistent proteins together as possible and outperformed other state-of-the-art algorithms in real and synthetic network experiments. The ACCMNA algorithm is based on a seed-and-extend schema, inspired by the backbone extraction from the BEAMS algorithm [1]. To begin, the calculation of node weights is included in the initialization process and the topology and sequence information of the network are also considered. Second, a clustering method finds the maximum edge weighted cluster, and an expansion method is used so that similar proteins can be put into a cluster as much as possible.

In this section, we introduce network alignment, and the remainder of this paper is organized as follows: In Section 2, we introduce the current state of network alignment research. In Section 3, we describe the definition of the problem, the details of the algorithm, and the two important innovations of our algorithm. In Section 4, we present the experimental results of the algorithm on different data sets and analyze the time complexity. Finally, the concluding remarks are discussed in Section 5.

2. Related Work

Network alignment algorithms have been widely studied in recent years. Research on pairwise network alignment was quite popular in the early years. Consequently, many excellent pairwise network alignment algorithms have been developed. The GRAAL family of algorithms consists of GRAAL [23], H-GRAAL [29], MI-GRAAL [24], C-GRAAL [28] and L-GRAAL [27],which use graphlet degree signatures and sequence similarity information to calculate the similarity between two networks. MAGNA [39] is optimized by a genetic algorithm to obtain the results, IsoRank [40] uses a method similar to Google's PageRank algorithm to calculate the similarity of nodes in a different network to find the alignment, NETAL [30] calculates the similarity score and obtains the alignment through the greedy search algorithm, and PINALOG [34] uses community detection to improve the alignment algorithm result. The above algorithms are compared and analyzed in detail in prior work [9]. With the increasing availability of PPI networks, the need for simultaneous alignment of multiple networks is growing, and the study of MNA algorithms has become increasingly popular. Multiple network alignment is different from the pairwise network alignment in that multiple networks can be aligned simultaneously,however the time complexity and computational difficulty of the algorithm become higher. Alignment algorithms that have been proposed include Graemlin [11], an early two-phase local MNA algorithm that learns the score function to optimize the vector of features while continuously iterating to produce the final alignment. However, Graemlin requires additional phylogenetic information as input. Both Graemlin1.0 and Graemlin2.0 [10] have been developed as local aligner and global aligner, respectively. IsoRankN [25] uses spectral graph theory to calculate similarity scores of nodes between any two networks. SMETANA [37] calculates the similarity score matrix based on a semi-Markov random walk model and uses probabilistic consistency transformations to enhance the similarity score matrix. The final alignment result is generated by a greedy searching method. BEAMS [1] establishes the alignment by generating the maximum edge weighted cliques. Then, the backbone extraction and merge strategy are used to produces alignment results of high biological consistency was proposed. CSRW [18] is an improved version of SMETANA in that it establishes a score matrix by using a context-sensitive random walk model. NetCoffee [16] is an extension of the T-Coffee algorithm [31], which uses a triplet approach that combines the third network information. Subsequently, the similarity score between any two networks is calculated and a simulated annealing method is used for continuous iteration until the final alignment is produced. Due to the limitation of the triplet method, NetCoffee can align only three or more networks. To address this issue, NetCofee2 [15] was proposed, which is based on graph feature vectors and is more accurate and efficient for aligning two or more networks. The MAPPIN [8] algorithm is an improved version of NetCoffee. MAPPIN can align two or more networks as well and combines the GO annotation information of proteins with topology and sequence similarity to calculate the similarity of nodes. Node Handprinting (NH) [35]is a global MNA, which solves the weighted bipartite graph matching problem by using the progressive alignment strategy to obtain the final optimal alignment. MultiMAGNA++ [42] is a global MNA designed to maximize the optimization objective function using genetic algorithms and is an extension of MAGNA and MAGNA++ [43], which are pairwise network alignment algorithms. FUSE [13] calculates similarity scores by using the non-negative matrix tri-factorization method and the k-partite matching algorithm to obtain the one-to-one alignment results. MPGM [20] generates seeds through sequence similarity and then obtains the final many-to-many alignment results through the percolation-based, graph-matching algorithm.

3. Method

3.1. Problem Definition

Let $G_1(V_1, E_1), G_2(V_2, E_2), \ldots, G_k(V_k, E_k)$ denote the k initial input PPI networks. Here, G_i represents the ith input network. V_i, E_i represent the nodes, that is, the proteins and edges (interactions), of the set of the ith input network, respectively. S represents the complete k-partite similarity graph of the weighted edge, where S has the same nodes as the input networks. The edges represent the interrelationship of proteins among different species. The value of the edge weight represents the sequence similarity score, where the value of weight is the bit score value between u and v, which are two nodes from different networks obtained through Basic Local Alignment Search Tool (BLAST), which was proposed in prior work [2]. S_β represents the filtered version of similarity graph S, which is a subgraph of S with some edges removed. If unfiltered sequence similarity data are used, then the computational complexity increases exponentially with the size of S and some similarity data may lead to incorrect alignment due to incompleteness in sequence similarity information. To avoid this, the S graph is filtered using beta, which is a user-defined threshold for each edge (x, y). If $w(x, y) < \beta \times max(x, y)$, then edge (x, y) in the similarity graph S is deleted. Here, $max(x, y)$ denotes the maximum value of the weight of an edge associated with x or y in S.

Assume that $A = \{Cl_1, Cl_2, \ldots, Cl_n\}$ is an alignment result of the input network, and alignment $A \in E$, where E is the edge set of all the networks mentioned above. In many-to-many network alignment A, for any cluster $Cl_i = \{V_{1,i}, V_{2,i}, \ldots, V_{k,i}\}$, $V_{c,i}$ is the node set of the ith cluster and nodes come from the cth network. $V_{c,i} \cap V_{c,j} = \emptyset, \forall i \neq j$, that is, a node belongs exclusively to one cluster. For a given network, the quality of alignment A is unknown and needs to be measured. Therefore, we quote the method in BEAMS [1] as the objective function of alignment A. Here, Formula 1 can be used as the objective function of the algorithm as follows:

$$AS(A) = \alpha \times CIQ(A) + (1 - \alpha) \times ICQ(A)\,, \tag{1}$$

where α is a real number from 0 to 1 that balances the contribution weight of topology and sequence scores.

$$CIQ(A) = \frac{\sum_{\forall Cl_m, Cl_n} |E_{Cl_m, Cl_n}| \times cs(m, n)}{\sum_{\forall Cl_m, Cl_n} |E_{Cl_m, Cl_n}|} \tag{2}$$

In the equation above, $CIQ(A)$ stands for the cluster interaction quality and is a score function that measures the quality of the conservative edge between clusters; E_{Cl_m, Cl_n} represents the set of edges whose vertices are in the distinct cluster Cl_m, Cl_n; $cs(m, n)$

represents any two clusters Cl_m, Cl_n and the proportion of the conserved edge network calculated by the formula $cs(m,n) = c'_{m,n}/c_{m,n}$; $c_{m,n}$ represents the number of networks with nodes in both clusters Cl_m and Cl_n; and $c'_{m,n}$ is the number of networks in which the vertices of the edges in E_{Cl_m,Cl_n} are in different networks. We assign $cs(m,n) = 0$ if $c'_{m,n} = 1$, which indicates that there is no conservative edge between clusters Cl_m and Cl_n.

$$ICQ(A) = \frac{\sum_{Cl_i \in A} ICQ(Cl_i)}{|A|} \tag{3}$$

$$ICQ(Cl_i) = \frac{\sum_{\forall(u,v)\in E(Cl_i)} \sqrt{\frac{w(u,v)^2}{w_{max}(u)\times w_{max}(v)}}}{|E(Cl_i)|} \tag{4}$$

Here, $ICQ(A)$ stands for the internal cluster quality and is defined as a measure for the sequence similarity score between aligned nodes, expressed in Formula 3, where A denotes alignment result; $ICQ(Cl_i)$ represents the sequence similarity measure of Cl_i, expressed in Formula 4, where $w(u,v)$denotes the bit score of sequence similarity information between node u and node v; $w_{max}(u)$denotes the maximum value of the edge attached to the node u in S_β; and $E(Cl_i)$ is the number of edges from the S_β incident on nodes in cluster Cl_i.

3.2. Algorithm

Inspired by the backbone extraction of the BEAMS algorithm [1], the whole framework of the ACCMNA algorithm is a seed-and-extend strategy. The method is used in combination with clustering to generate the alignment cluster. The clusters are generated in each iteration as seeds and the seeds are expanded in the input network to generate new clusters. The pseudo-code for the ACCMNA algorithm is demonstrated in the following Algorithm description. The algorithm is initialized, while both the alignment set and the candidate set are empty. To begin, the weight of each node is calculated based on the topology and sequence information of the network, then the first candidate C_0 is generated by searching the cluster in the graph through the function $Generate\_Candidate(S_\beta)$. This function searches for the node with the largest weight and its neighbor nodes in the weighted graph S_β to generate a subgraph of S_β through these nodes, while a cluster is generated in this subgraph. The main part of the algorithm is the repeat loop. The first step involves selecting the candidate with the highest AS score in candidate set C as the new alignment A_{new} in this loop, adding A_{new} to the alignment set A, and deleting the nodes contained in A_{new} in S_β. The second step is generating the neighborhood node set in the PPIN according to the nodes in A_{new} and establishing the subgraph $N_{S_\beta}(A_{new})$ of this neighborhood node set. If $N_{S_\beta}(A_{new})$ contains only isolated nodes, then C_{new2} is empty; otherwise, a new prospective candidate $C_{new1}$is generated in the graph. In the third step, if C_{new1} contains nodes in each input network, then it is not extended; otherwise, candidate C_{new2} is generated by extending C_{new1} in the S_β. In the fourth step, if there is overlap between the newly generated cluster and the candidate set, then the candidate needs to be updated and the above four steps are repeated until the candidate set is empty.

```
Algorithm description
  Input: $S_\beta$, $G_1$, $G_2$, ..., $G_k$, $\alpha$
```

```
    Output: Set of cluster A
    $C = \emptyset$;  $A = \emptyset$;
    //Initial
    Calculate node score NodeWeight;
    $C_0$=Generate initial candidates in $S_\beta$;
    $C = C \cup \{C_0\}$;
    repeat
        Select a cluster $A_{new}$ from candidate set C;
        $A = A \cup \{A_{new}\}$;
        remove $A_{new}$ from $S_\beta$;
        generate new candidates $C_{new1}$ in $A_{new}$'s neighbors;
        expand new Candidate of $C_{new1}$ in $S_\beta$;
        $C = C \cup \{C_{new2}\}$;
        for $all$ $C_i \in C$ do
            if $C_i \cap A_{new} \neq \emptyset$ then
                if i==0 then
                    $C_0$=Generate initial candidates in $S_\beta$;
                else:
                    generate new candidates $C_i$;
                end if
            end if
        end for
    until no more Candidate
  end.
```

Calculation of the Node Weight. When the initial candidate is generated in S_β, the node with the highest weight needs to be located, and then the cluster is searched with this node as the center. Inspired by the HubAlign algorithm for computing the node similarity function, the HubAlign algorithm is used for pairwise network aligners. HubAlign uses a minimum-degree heuristic method to measure the role of nodes in the network and preferentially aligns the more important nodes [14]. The use of the HubAlign algorithm is extended to the calculation of the similarity of nodes among multiple networks. Using a similar approach to HubAlign, the topological importance score for all nodes in the network is calculated to begin with, and then the score of all nodes is calculated by combining the sequence similarity information between pairs of nodes from different networks. The degree is one of the properties that can reflect the importance of nodes, given that the importance of nodes in the network can be determined by the global topological property. The weights of nodes and edges are calculated by traversing from the node with the smallest degree to the node with a degree of 10. The weight of the node with a small degree is transferred to the node or edge with a larger degree at the adjacent node so that higher weight scores are assigned to nodes with a higher degree in the network and there is a greater weight of the edge connected with the node. The network nodes and edge weights are initialized as follows (see Fig. 1 for a simplified example):

$$we(u,v) = \begin{cases} 1, \text{(u, v)} \in E \\ \\ 0, \text{otherwise} \end{cases}, wn(u) = 0, \forall u \in V\,, \tag{5}$$

where $wn(u)$ represents the weight of the nodes in V; $we(u, v)$ represents the weight of the edges between nodes u and v in PPINs; for a particular node u in a network, let $deg(u)$ be the degree of node u; $N(u)$ denotes the neighbor nodes set of node u; and $|N(u)|$ is the number of neighbors of node u and also the degree of node u. The node weight update starts from the node with the lowest degree and the topology information of the node is gradually transferred to the neighbors with the higher degree. Nodes with zero degrees are generally ignored. For a given node u, $\forall v \in N(u)$, $wn(v) = wn(v) + wn(u) + we(u, v)$, if $deg(u) = 1$. If $deg(u) > 1$ and $\forall v_1, v_2 \in N(u)$, then

$$we(v_1, v_2) = we(v_1, v_2) + \frac{wn(u) + \sum_{v \in N(u)} we(u, v)}{\frac{|N(u)||N(u)-1|}{2}}\,, \tag{6}$$

Following the weight calculation in Formula 6, the importance score of the node is calculated by combining the weight of the node with the weight of the edge, as follows:

$$importance(u) = wn(u) + \gamma \sum_{v \in V} we(u, v)\,. \tag{7}$$

where $importance(u)$ is the importance score of node u, and γ is set $\gamma = 0.2$ and controls the contribution of the node related edge weights. The importance score obtained from the network topology information is combined with the sequence similarity information to obtain the final node weight. Nodes with zero degrees are generally ignored again. Formula 8 is proposed to calculate the sequence similarity score of node u in S_β. The formula is as follows:

$$B(u) = \frac{\sum_{v \in N_S(u)} B(u, v)}{|N_S(u)|}\,, \tag{8}$$

where $B(u, v)$ represents the sequence similarity information between nodes u and v, which in this paper was calculated by the BLAST bit score; and $B(u)$ represents the average value of sequence similarity values related to node u. Finally, the weight of each node in PPIN is obtained by combining the topology importance and sequence similarity score. The final node weight is calculated as follows:

$$Weight(u) = \alpha \times importance(u) + (1 - \alpha) \times B(u)\,, \tag{9}$$

where α is a balancing parameter, see the definition of Equation 1.

Cluster Searching. For a graph with a given non-negative weight edge, the candidate is generated by searching similar nodes in the search graph according to the edge weight. Thus, nodes with high sequence similarity are clustered. The clustering method is combined with the network alignment and similar nodes are gathered to generate clusters through the clustering method in the search graph. Inspired by the clustering algorithm SPICi [19], a clustering method based on the seed-and-extend approach is adopted. The data of sequence similarity is incomplete, which may lead to similar nodes with no sequence similarity value between them and contribute to the incomplete alignment. The alignment is constructed by improving this method with the inclusion of similar nodes in the same cluster as much as possible.

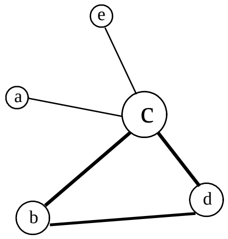


Fig. 1. An example to illustrate the calculation of node weights. This example network has five nodes. The thickness of an edge shows its weight, and the size of a node shows its weight. For example, for node a with degree one, N(a)={c}, wn(c)=wn(c)+wn(a)+we(a, c); however, for node b with degree greater than one, N(b)={c, d}, we(c, d)=we(c, d)+(wn(b)+we(b, c)+we(b, d))/(2(2-1)/2)=we(c, d)+wn(b)+we(b, c)+we(b, d)

The weights of all the nodes in the search graph are calculated according to the following formula:

$$deg_w(u) = \sum_{v \in N_S(u)} B(u, v)\,, \tag{10}$$

where $N_S(u)$ represents the set of neighbor nodes of node u in the filtered similarity graph S_β; and $B(u, v)$ represents the sequence similarity score between node u and node v in the S_β mentioned above. In each graph with weighted edges, the node with the highest weighted degree is selected as the first seed. The higher the weight of a node, the higher its importance in the graph, and it can be used as a meaningful seed node. A higher weight between nodes indicates a higher sequence correlation between two nodes, and, therefore, the weight of the neighbor of the first seed node is normalized and the nodes are divided into five bins according to the normalized weight between them. In this study, they were as follows: (0,0.2],(0.2,0.4],(0.4,0.6],(0.6,0.8],(0.8,1). Searching started from the bin with the highest node weight, that is, (0.8, 1], to the bin with the lowest node weight, that is, (0, 0.2]. If the bin being searched is not empty, then the node with the highest node degree weight in the current bin is the second seed; otherwise, searching continues in the next bin. The neighbor node that is most similar to the seed node is also an important node in the network.

After the initial seed node pair is obtained, the graph is extended through two seed nodes. First, S represents the nodes already included in the current cluster, and S contains only two seed nodes at the beginning of the seed extension. The search node set that could be added to S is composed of the neighbor nodes of the nodes in S. The node with the maximum value of $support(u, S)$ in the search set is selected in each iteration of the extension. The score of $support(u, S)$ is the sum of the weight of the edges in S related to u, indicating the correlation between node u and the node in S. Two constraint conditions decide whether to add node u to S. Node u is only added to S only if Formula 11 was satisfied; otherwise, the search loop is terminated.

$$\begin{cases} density\{S \cup \{u\}\} > T_d \\ \frac{|E_s(u)|}{|S| \times density\{S \cup \{u\}\}} \geq T_s \end{cases} \tag{11}$$

Here, $density\{S \cup \{u\}\}$ denotes the density of graph S after adding node u, and it reflects how close the current graph S is to clique; $|E_S(u)|$ denotes the number of edges related to u in S; and $|S|$ is the number of nodes in graph S. The Values for T_s, T_d were set to 0.5 in here.

After generating the prospective candidate in the neighborhood graph, the generated candidate cluster is extended only when the number of networks in the cluster is less than that of the input networks. The basic process of expansion is the same as that of the above search process. Here, S is the newly generated prospective candidate, and the search nodes are the neighbor nodes of the nodes in S. When the current node meets the two constraints above, the node is added. However, since there is no direct correlation between the extended search set and the nodes in the original alignment cluster, stricter constraints should be set. The values of T_s, T_d were set to be higher, namely 0.7 in the synthetic network and 0.9 in the real networks in here.

4. Results and Discussion

4.1. Datasets

The ACCMNA algorithm was compared with IsoRankN, SMETANA and BEAMS. IsorankN is the first global MNA. As one of the most popular two-phase alignment algorithms, many alignment algorithms have been compared to it. SMETANA is a multiple network aligner based on semi-Markov random walk and probabilistic consistency transformations. Several studies in the literature have proved that SMETANA can produce comparative results with relative topological significance. BEAMS is a heuristic algorithm that searches for the weighted maximum cluster, and the experimental results in many previous reports indicate that BEAMS can produce alignments with good functional consistency.

Table 1. The number of proteins and interactions of five eukaryotic species

	Node	Edge
S. cerevisiae	6659	82932
C. elegans	19756	4884
D. melanogaster	14098	25054
H. sapiens	22369	55168
M. musculus	24855	592

We used real and synthetic networks for the verification of our algorithm. Five eukaryotic network databases derived from the IsoBase [32]were used, S. cerevisiae, C. elegans, D. melanogaster, H. sapiens and M. musculus, which are consistent with the data used in SMETANA [37], IsoRankN [25] and BEAMS [1]. The PPINs data were constructed by combining data from BIOGRID [4], DIP [38], HPRD [21], IntAct [3] and MINT [6]. The node and edge data for each network are presented in Table 1. The sequence homology information of the network corresponded to the BLAST bit score retrieved from Ensembl [17].

The synthetic network used data provided by Network Alignment Performance Assessment Benchmark (NAPAbench) [36] and there were three different network growth models: crystal growth (CG) model [22], duplication-mutation-complementation (DMC) model [41] and duplication with random mutation (DMR) model [33]. Each model contained eight networks. Each network of the CG model contained 1000 nodes and 3985 edges. Each network of the DMC model consisted of 1000 nodes and the number of edges of each network was 1919, 1853, 1923, 1840, 1867, 1848, 1818 and 1867, respectively. The number of nodes in the DMR model network was also 1000 and the number of network edges was 2031, 2092, 1967, 1977, 1959, 1998, 2030 and 2056, respectively.

Table 2. Experimental results on a synthetic network CG model. best performance is shown in bold

	CIQ	SPE	Sen	MNE	nGOC
SMETANA	0.812	0.906	0.573	**0.071**	0.907
IsoRankN	0.692	0.620	0.679	0.276	0.575
BEAMS	0.702	0.879	0.588	0.112	0.910
ACCMNA	**0.892**	**0.920**	**0.713**	**0.071**	**0.947**

Table 3. Experimental results on a synthetic network DMC model. best performance is shown in bold

	CIQ	SPE	Sen	MNE	nGOC
SMETANA	0.754	**0.869**	0.631	**0.106**	**0.865**
IsoRankN	0.573	0.618	0.518	0.294	0.546
BEAMS	0.507	0.806	0.553	0.182	0.833
ACCMNA	**0.791**	0.858	**0.755**	0.119	0.850

Table 4. Experimental results on a synthetic network DMR model. best performance is shown in bold

	CIQ	SPE	Sen	MNE	nGOC
SMETANA	0.689	**0.872**	0.573	**0.106**	**0.873**
IsoRankN	0.545	0.607	0.566	0.304	0.544
BEAMS	0.640	0.815	0.558	0.181	0.841
ACCMNA	**0.748**	0.861	**0.714**	0.119	0.845

In the comparison experiment of the synthetic network, the parameters of our algorithm α and β were set as 0.5 and 0.2, respectively. The values of α and β in our algorithm were 0.5 and 0.3 on the real networks, respectively. The parameters of other compared algorithms were set as the recommended parameters from the literature. The parameters

Table 5. Performance of different algorithms on real networks. best performance is shown in bold

	CIQ	SPE	Sen	MNE	nGOC
SMETANA	**0.054**	0.724	0.360	1.394	0.247
IsoRankN	0.027	0.733	0.303	1.437	0.248
BEAMS	0.035	0.798	**0.379**	1.290	0.309
ACCMNA	0.041	**0.813**	0.345	**1.218**	**0.331**

of the BEAMS algorithm synthetic network were set as the same as our algorithm. The parameters α and β of the BEAMS algorithm real networks were set to 0.5 and 0.2, respectively. Parameter α of the IsoRankN algorithm was set to 0.6, and parameters $\alpha$and β of SMETANA were set to 0.9 and 0.8, respectively, and $n_{max} = 10$.

4.2. Analysis of the Alignment Result

The alignment results of the above algorithms were all many-to-many alignment, which indicated that, for each cluster, multiple nodes from the same network may exist. Protein coverage showed the total number of aligned nodes. The nodes were classified in each cluster according to their source network. The node k-coverage denotes the number of nodes that belong to clusters that contain nodes from k networks. To measure the biological significance of the alignment, GO annotation was used to evaluate the consistency of aligned proteins. If at least two proteins in a cluster were annotated by the GO category, then the whole cluster was considered to be annotated, and if all proteins in an annotated cluster shared the same GO category, then the whole cluster was considered to be consistent. The k-coverage of the consistent nodes denotes the number of consistent proteins present in clusters that contain proteins from k networks. As shown in Fig. 2, the total number of proteins aligned by ACCMNA, SMETANA and BEAMS was very close. IsoRankN aligned the least number of proteins. In general, each cluster is expected to contain proteins from as many species. The alignment generated by the ACCMNA algorithm had the largest number of 8-coverage of nodes. This indicates that the ACCMNA algorithm produced more high-quality clusters. From the consistent protein results in Fig. 2(b), the results of the ACCMNA algorithm indicate that its performance was the best among several aligners and that most of the consistent nodes belonged to clusters from k=8 species. Figure 2 shows that our algorithm produced the cluster that contained the highest number of proteins and consistent proteins from eight species. This can also demonstrate that our algorithm discovered more meaningful information and was more biologically consistent. The alignment results on real networks are displayed in Fig. 3, where A represents the protein coverage of alignment, and B represents the consistent protein coverage. There was little difference between the ACCMNA algorithm results and the BEAMS and SMETANA results, all of which were higher than IsoRankN. The coverage of proteins and consistent proteins on both real networks and synthetic networks revealed that the ACCMNA algorithm outperformed the other algorithms.

The alignment performance was measured using metrics established in the literature. The alignment measurement scores on the synthetic network and the real networks are displayed in Tables 2, 3, 4 and 5. CIQ has been proposed as a measurement for conserved edges between clusters and used in previous literature for result comparison [42],

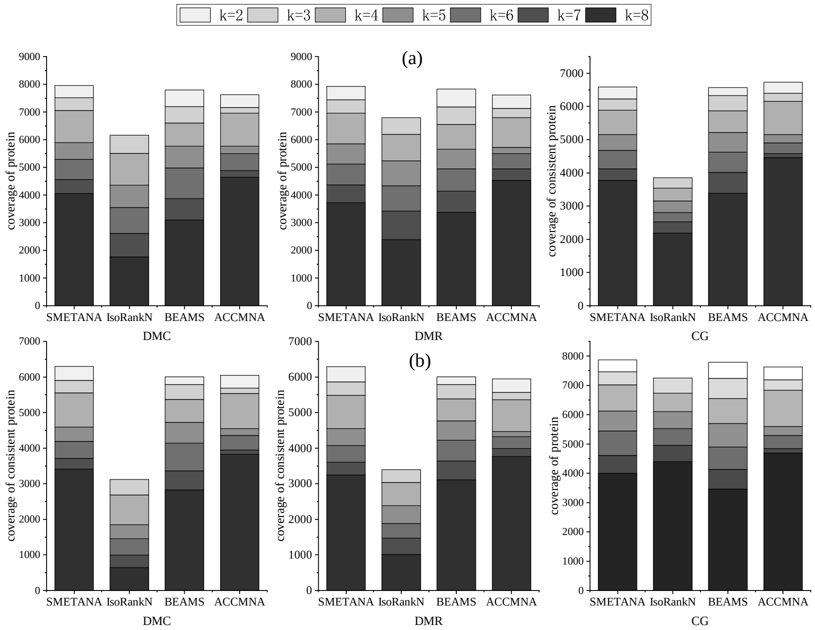


Fig. 2. Performance of various network alignment algorithms in the synthetic network. From left to right, respectively, are the results under the CG, DMC and DMR network model. (a) Node k-Coverage, where k denotes the number of input networks; (b) Consistent Node k-Coverage

[20]. Results on synthetic networks from Tables 2, 3 and 4 showed that our algorithm had the highest CIQ score on the three network sets CG, DMC and DMR, which suggests that our algorithm contained the highest proportion of conservation interaction. SPE stands for Specificity, which was proposed in prior work [37], and it is the proportion of the number of consistent clusters in the number of annotated clusters. Our algorithm had much higher SPE scores than IsoRankN and BEAMS in the three network sets and the highest score in the CG network. The other two network sets ranked second and were very close to the results of the first SMETANA. Sen represents Sensitivity, defined in previous literature [10], which indicates the sensitivity of the alignment. ACCMNA had the highest Sen score among the three network sets. MNE represents the Mean Normalized Entropy, which is an approach to measure the consistency of the alignment. The Mean Normalized Entropy was the average normalized entropy of all the clusters defined by prior work [25]. For a given cluster Cl_i, the normalized entropy is defined as $NE(Cl_i) = -\frac{1}{\log d} \times \sum_{i=1}^{d} p_i \times \log p_i$, where d denotes the number of different GO categories in cluster Cl_i, and p_i represents the proportion of proteins annotated by GO_i in cluster Cl_i. The biological consistency of the alignments increased with lower MNE values. Like the SPE results, the MNE value of our algorithm was the lowest in the CG model network, and our algorithm ranked second on the DMC and DMR datasets; however, the score was very close to that of the first SMETANA. nGOC has been also proposed for the measurement of the alignment consistency by prior researchers [1]. nGOC is an extension of GO Consistency(GOC), and the measurement used in one-to-one pairwise network alignment was extended to measure many-to-many alignment. nGOC is the average value of $nGOC(Cl_i)$ of all the clusters. For a given cluster Cl_i, nGOC is defined as $nGOC(Cl_i) = \frac{|GO_{int}|}{|GO_{uni}|} \times c$, where GO_{int} and GO_{uni} represent the intersection and union of the GO annotation items of proteins in cluster Cl_i, respectively, and c is the number of annotated proteins in cluster Cl_i. The consistency of alignment results increases with higher nGOC values. The ranking of nGOC for ACCMNA was the same as that of SPE and MNE. The main reason for this result may be that the number of nodes and edges of the eight networks in the CG model were the same, which indicated that our algorithm could get a better alignment in the case of a similar network size. However, the alignment generated by our algorithm was more consistent and specific. The result in Table 5 shows that the alignment generated by ACCMNA on real networks had the highest SPE, MNE and nGOC score, which also shows that our algorithm was more specific and consistent. ACCMNA scored second in the CIQ score, and it was only slightly lower than SMETANA. The SMETANA algorithm places high importance on the topology information of the network; therefore, the alignment on the real networks had a high topology score, but several biological scores were low. We believe that SMETANA performed well in the synthetic network, mainly because of its special network characteristics, namely, a relatively ideal network situation, which can explain the result on the synthetic network being slightly higher than ACCMNA. However, the alignment on the real networks was worse than ACCMNA.

To prove that the alignment generated by ACCMNA can perform well both in topological conservation and functional consistency, the product of CIQ and nGOC was plotted for all the algorithms and networks sets. This amplifies the advantages of the ACCMNA algorithm. CIQ is a measure to calculate the proportion of conservative edges between clusters, while nGOC measures the biological consistency of alignment. These are depicted in Fig. 4. Although some measures of the ACCMNA algorithm in Tables 2, 3, 4 and 5 were worse than SMETANA, the ACCMNA algorithm received the highest score among all the algorithms when the product of CIQ and nGOC was calculated. This proves that our algorithm can get a good result in both topology and biological consistency.

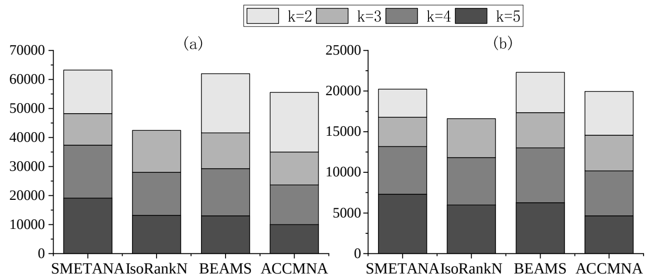


Fig. 3. Node and Consistent Node Coverage of different algorithms: (a)Node k-Coverage; (b) Consistent Node k-Coverage

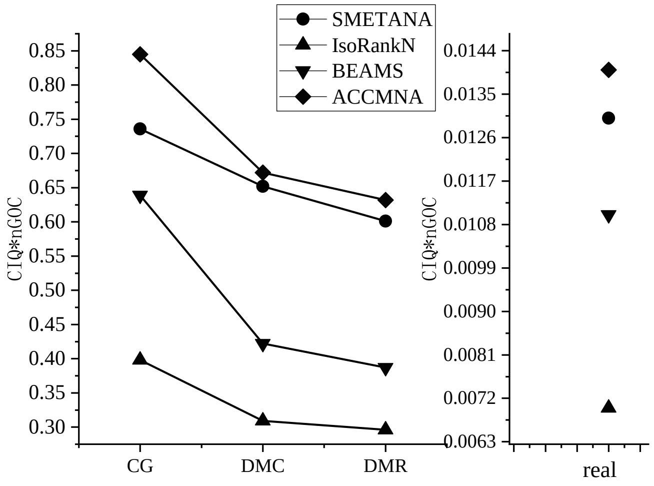


Fig. 4. The product CIQ and nGOC for all the algorithms. The figure on the left shows the scores on the three network models of the synthetic network, and the figure on the right shows the results on the real networks

4.3. Analysis of the Time Complexity

Let V be the set $V_1 \cup \ldots V_k$, and $n = max\{|V_1|, \ldots, |V_k|\}$. As we mentioned before, it takes $O(|V|)$ to calculate the NodeWeight of each node. Thus, the running time of

ACCMNA is mainly determined by the time spent in the main repeat loop. The number of iterations of the loop is $O(|V|)$, and, since the maximum number of output clusters can be $|V|$at most, each iteration finds a new cluster, and the iterations continue until no new clusters remain. The function Select_Candidate requires $O(|V|k^2\varDelta_{max})$, where k is the number of PPINs and $\varDelta_{max}$ the maximum degree in V. The function Generate_Candidate is made on the neighborhood graph of the new cluster. The total running time required by function Generate_Candidate is $O(\varDelta(k\varDelta_{max}))$, where $\varDelta$ is the maximum degree in S_β. Function expand_Candidate requires $O(k\varDelta)$. Note that the function Generate_Candidate is executed only once in the for-loop, but the functions Generate_Candidate and expand_Candidate in the for-loop are executed $O(|V|)$ times since the number of candidates at a specific iteration can be at most $|V|$. Thus, the overall time complexity of our algorithm is $O(|V|^2\varDelta(k\varDelta_{max}) + |V|^2k\varDelta + |V|^2\varDelta) = O(|V|^2k\varDelta^2)$.

4.4. Discussion of the Alignment Result

In this section, we discuss the alignment results of the ACCMNA algorithm on the real networks and the synthetic networks along with the comparison experiments with other state-of-the-art algorithms. The above experimental results show that the algorithm proposed in this paper could obtain better alignment results than other state-of-the-art algorithms. The node coverage shows that the ACCMNA algorithm could produce more node coverage with a larger k, indicating its ability to produce higher quality alignment and more useful biological information. Moreover, the measurement results of the biological consistency, specificity and sensitivity show that the scores of our algorithm ranked high among several algorithms, which indicated that the alignment results produced by ACCMNA had good biological significance. When topological and biological consistency scores are combined, the alignment results of the ACCMNA algorithm can reach the balance between topological and biological consistency.

5. Conclusion

To solve the NP-hard problem of network alignment and the computational complexity of MNA gradually increasing with the increase of network size, a new and efficient ACCMNA aligner was proposed in this paper, which combines topology and sequence similarity information for alignment generation. ACCMNA is an aligner that utilizes the importance of nodes and combines clustering methods to produce better alignment results. The basic framework of ACCMNA is the seed-and-extend search method. The algorithm utilizes the degree and neighbors of nodes to calculate the node weight, which aims to reduce the complexity of alignment and make as many similar nodes as possible that can be successfully mapped by combining the clustering method to search the alignment. The ACCMNA algorithm was compared against excellent and representative MNA algorithms on both real and synthetic networks. Extensive evaluations showed that the ACCMNA algorithm performed well both in topological conservation and functional consistency. The superior experimental results also reflected that the ACCMNA algorithm is an efficient and accurate aligner that can be applied to PPINs of various sizes within an acceptable running time. In addition to proving the effectiveness of the method proposed in this paper, the alignment results generated by ACCMNA are of reference significance for the study of real networks. Moreover, it has the potential to be extended to other types of complex networks in the future, rather than remain limited to PPINs.

References

1. Alkan, F., Erten, C.: Beams: Backbone extraction and merge strategy for the global many-to-many alignment of multiple ppi networks. Bioinformatics 30(4), 531–539 (2013)
2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology 215(3), 403–410 (1990)
3. Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J.: The intact molecular interaction database in 2010. Nucleic Acids Research 38(suppl_1), 525–531 (2010)
4. Bobby-Joe, B., Chris, S., Teresa, R., Lorrie, B., Ashton, B., Michael, L., Rose, O., Lackner, D.H., Jürg, B., Valerie, W.: The biogrid interaction database: 2008 update. Nucleic Acids Research 36(suppl_1), 637–640 (2008)
5. Braovic, M., Stipanicev, D., Seric, L.: Retinal blood vessel segmentation based on heuristic image analysis. Computer Science and Information Systems 16(1), 227–245 (2019)
6. Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., Cesareni, G.: Mint, the molecular interaction database: 2009 update. Nucleic Acids Research 38(5), 32–39 (2010)
7. Cook, S.: The complexity of theorem-proving procedures. In: Proc Acm Symposium on the Theory of Computation. pp. 151–158 (1971)
8. Djeddi, W.E., Yahia, S.B., Nguifo, E.M.: A novel computational approach for global alignment for multiple biological networks. IEEE/ACM Transactions on Computational Biology Bioinformatics 15(6), 2060–2066 (2018)
9. Elmsallati, A., Clark, C., Kalita, J.: Global alignment of protein-protein interaction networks: A survey. IEEE/ACM Transactions on Computational Biology Bioinformatics 13(4), 689–705 (2016)
10. Flannick, J., Novak, A., Do, C.B., Srinivasan, B.S., Batzoglou, S.: Automatic parameter learning for multiple local network alignment. Journal of Computational Biology 16(8), 1001–1022 (2009)
11. Flannick, J., Novak, A., Srinivasan, B.S., McAdams, H.H., Batzoglou, S.: Graemlin: general and robust alignment of multiple large interaction networks. Genome Research 16(9), 1169–1181 (2006)
12. Gao, J., Song, B., Ke, W., Hu, X.: Balanceali: multiple ppi network alignment with balanced high coverage and consistency. IEEE Transactions on Nanobioscience 16(5), 333–340 (2017)
13. Gligorijević, V., Malod-Dognin, N., Pržulj, N.: Fuse: multiple network alignment via data fusion. Bioinformatics 32(8), 1195–1203 (2016)
14. Hashemifar, S., Xu, J.: Hubalign: an accurate and efficient method for global alignment of protein–protein interaction networks. Bioinformatics 30(17), i438–i444 (2014)
15. Hu, J., He, J., Gao, Y., Zheng, Y., Shang, X.: Netcoffee2: A novel global alignment algorithm for multiple ppi networks based on graph feature vectors. In: International Conference on Intelligent Computing. pp. 241–246 (2018)
16. Hu, J., Kehr, B., Reinert, K.: Netcoffee: a fast and accurate global alignment approach to identify functionally conserved proteins in multiple networks. Bioinformatics 30(4), 540–548 (2014)
17. Hubbard, T.J., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., et al.: Ensembl 2009. Nucleic Acids Research 37(suppl_1), D690–D697 (2009)

18. Jeong, H., Yoon, B.J.: Accurate multiple network alignment through context-sensitive random walk. In: BMC Systems Biology. vol. 9, pp. 1–12. Springer (2015)
19. Jiang, P., Singh, M.: Spici: a fast clustering algorithm for large biological networks. Bioinformatics 26(8), 1105–1111 (2010)
20. Kazemi, E., Grossglauser, M.: Mpgm: Scalable and accurate multiple network alignment. IEEE/ACM Transactions on Computational Biology and Bioinformatics 17(6), 2040–2052 (2019)
21. Keshava Prasad, T.t., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al.: Human protein reference database—2009 update. Nucleic Acids Research 37(suppl_1), D767–D772 (2009)
22. Kim, W.K., Marcotte, E.M.: Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. PLoS Comput Biol 4(11), e1000232 (2008)
23. Kuchaiev, O., Milenković, T., Memišević, V., Hayes, W., Pržulj, N.: Topological network alignment uncovers biological function and phylogeny. Journal of the Royal Society Interface 7(50), 1341–1354 (2010)
24. Kuchaiev, O., Pržulj, N.: Integrative network alignment reveals large regions of global network similarity in yeast and human. Bioinformatics 27(10), 1390–1396 (2011)
25. Liao, C.S., Lu, K., Baym, M., Singh, R., Berger, B.: Isorankn: spectral methods for global alignment of multiple protein networks. Bioinformatics 25(12), i253–i258 (2009)
26. Liu, X., Zhuang, C., Murata, T., Kim, K.S., Kertkeidkachorn, N.: How much topological structure is preserved by graph embeddings? Computer Science and Information Systems 16(2), 597–614 (2019)
27. Malod-Dognin, N., Pržulj, N.: L-graal: Lagrangian graphlet-based network aligner. Bioinformatics 31(13), 2182–2189 (2015)
28. Memišević, V., Pržulj, N.: C-graal: Common-neighbors-based global graph alignment of biological networks. Integrative Biology 4(7), 734–743 (2012)
29. Milenković, T., Ng, W.L., Hayes, W., Pržulj, N.: Optimal network alignment with graphlet degree vectors. Cancer Informatics 9, CIN–S4744 (2010)
30. Neyshabur, B., Khadem, A., Hashemifar, S., Arab, S.S.: Netal: a new graph-based method for global alignment of protein–protein interaction networks. Bioinformatics 29(13), 1654–1662 (2013)
31. Notredame, C., Higgins, D.G., Heringa, J.: T-coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology 302(1), 205–217 (2000)
32. Park, D., Singh, R., Baym, M., Liao, C.S., Berger, B.: Isobase: a database of functionally related proteins across ppi networks. Nucleic Acids Research 39(suppl_1), D295–D300 (2010)
33. Pastor-Satorras, R., Smith, E., Solé, R.V.: Evolving protein interaction networks through gene duplication. Journal of Theoretical biology 222(2), 199–210 (2003)
34. Phan, H.T., Sternberg, M.J.: Pinalog: a novel approach to align protein interaction networks—implications for complex detection and function prediction. Bioinformatics 28(9), 1239–1245 (2012)
35. Radu, A., Charleston, M.: Node handprinting: a scalable and accurate algorithm for aligning multiple biological networks. Journal of Computational Biology 22(7), 687–697 (2015)
36. Sahraeian, S.M.E., Yoon, B.J.: A network synthesis model for generating protein interaction network families. PloS One 7(8), e41474 (2012)
37. Sahraeian, S.M.E., Yoon, B.J.: Smetana: accurate and scalable algorithm for probabilistic alignment of large-scale biological networks. PloS One 8(7), e67995 (2013)
38. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D.: The database of interacting proteins: 2004 update. Nucleic Acids Research 32(suppl_1), D449–D451 (2004)
39. Saraph, V., Milenković, T.: Magna: maximizing accuracy in global network alignment. Bioinformatics 30(20), 2931–2940 (2014)

40. Singh, R., Xu, J., Berger, B.: Global alignment of multiple protein interaction networks with application to functional orthology detection. Proceedings of the National Academy of Sciences 105(35), 12763–12768 (2008)
41. Vázquez, A., Flammini, A., Maritan, A., Vespignani, A.: Modeling of protein interaction networks. Complexus 1(1), 38–44 (2003)
42. Vijayan, V., Milenković, T.: Multiple network alignment via multimagna++. IEEE/ACM Transactions on Computational Biology and Bioinformatics 15(5), 1669–1682 (2017)
43. Vijayan, V., Saraph, V., Milenković, T.: Magna++: Maximizing accuracy in global network alignment via both node and edge conservation. Bioinformatics 31(14), 2409–2411 (2015)
44. Yin, A., Chen, C., Hu, D., Huang, J., Yang, F.: An improved heuristic-dynamic programming algorithm for rectangular cutting problem. Computer Science and Information Systems 17(3), 717–735 (2020)

Jing Chen, born in 1977. Ph. D., associate professor, her research interests include complex networks, indoor positioning, etc.

Jia Huang, born in 1996. Master degree candidate, her research interests include complex networks.

Received: September 09, 2020; Accepted: April 06, 2021.

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

004

COMPUTER Science and Information
Systems : the International journal /
Editor-in-Chief Mirjana Ivanović. – Vol. 18,
No 4 (2021) - . – Novi Sad (Trg D. Obradovića 3):
ComSIS Consortium, 2021 - (Belgrade
: Sigra star). –30 cm

Polugodišnje. – Tekst na engleskom jeziku

ISSN 1820-0214 (Print) 2406-1018 (Online) = Computer
Science and Information Systems
COBISS.SR-ID 112261644

Cover design: V. Štavljanin
Printed by: Sigra star, Belgrade