Computer Science and Information Systems 23(2):775–800 https://doi.org/10.2298/CSIS251030023C

Research on Key Technology of Network Information Extraction Oriented to Web Topic Detection for Big Data

Mo Chen

Business College of Beijing Union University,
A3, Yanjingdongli, Chaoyang District, Beijing, 100025, P.R. China
mo.chen@buu.edu.cn

Abstract. In the context of today's big data and numerical intelligence era, this study explores an incremental network information extraction technology for Web topic detection characterized by the semi-structured or unstructured big data as important research object to promote network information detection application. This study takes Web big data as the main research object and proposes an incremental network information extraction idea for Web topic detection. In this idea, the designed algorithm of theme similarity measurement for incremental network information extraction can extract Web instances related to theme, and calculate importance of Web instances related to theme, furthermore, the designed algorithm of incremental instance extraction for Web topic detection can analyze Pattern and BasePattern according to extracted Web instance URL, and conduct segmentation for Web instance title and text content, extract keywords, which are capable of describing Web topic. Experimental results demonstrate that the framework, method, and algorithm proposed in this paper significantly outperform traditional methods in network information extraction. Particularly, the accuracy rate of extracted Web instances that are similar to the theme can reach 0.833, the F-Measure value of extracted Web instances that are similar to the theme under different threshold adjustment is close to 0.83, the accuracy rate of topic detection under the condition of determining the number of Web news instances extracted, the threshold and the parameter value is close to 0.82. The study concludes that the incremental network information extraction idea proposed in this paper is feasible, verifiable, and superior, and can play an important role in reconfiguring numerical intelligence warehouses for detecting Web topic, inferring the Web hierarchical big data propagation path.

Keywords: Incremental Network Information Extraction, Big Data, Web Topic Detection.

1. Introduction

Against the backdrop of the development of big data era, scholars are still exploring and innovating in the face of the constantly emerging heterogeneous big data and diverse demands for numerical intelligence applications [1,2,3,4]. At this stage, the network has been providing the most valuable information to various users, while the amount of network data is also growing at an astonishing rate [5,6,7]. Therefore, scholars urgently need to think about how to extract more accurate knowledge from complex big data. In this exploration process, the incremental network information extraction for Web topic detection can be an important research direction.

In the network heterogeneous big data, Internet news, as a streaming resource, has the characteristics of real-time update, wide dissemination, and high interaction from the perspective of application [8,9,10]. With the continuous occurrence of various events, the number of Internet news is showing an explosive growth trend. From a scientific perspective, it has shown the 5V characteristics of volume, variety, value, velocity and veracity for big data [11,12,13,14]. Based on the above characteristics, how to study the incremental network information extraction method for Web topic detection, it has become an urgent problem to build a numerical intelligence warehouses and provide a real-time big data source for the network information detection application.

To establish a strong foundation for the research of this paper, the next section will review existing methods related to network information extraction, and highlight their strengths and limitations. Based on the study of literature, this paper will propose an idea for in-depth exploration of the process of network information extraction, and elaborate on the main implementation methods and algorithms for Web topic detection.

This paper will make three key contributions: (1) it introduces a novel network information extraction method for Web topic detection based on big data, (2) it optimizes the process for theme similarity measurement, instance information extraction, link importance calculation, filter mode analysis and so on, (3) it provides extensive experimental validation and demonstrates the effectiveness of the proposed approach.

2. Related works

Oriented to Web application for topic detection, it needs a lot of real corpus as support. However, at present, the published Web news is showing a massive increase trend facing frequent social events [15,16,17]. So, this paper intends to use Web news source material as the research object, and discuss the effective process of extracting Web news. Some scholars have done certain research about the technology of network information extraction shown in Table 1, and these research results will be a foundation for continuing to research key technology of network information extraction for Web topic detection based on big data.

An effective paradigm is provided by the pseudo-labeling based semi-supervised learning algorithm [18], in order to alleviate the reliance on labeled data by leveraging unlabeled data. However, in the task for key information extraction, the main challenges for this algorithm are as follows, the context dependency of key information extraction results in incorrect pseudo-labels, and the intra-class variance is high, the inter-class variation is low. To this end, a similarity matrix pseudo-label bias rectification semi-supervised method is also proposed for key information extraction task, which improves the quality of pseudo-labels on key information extraction benchmarks with rare labels. More specifically, the similarity matrix bias rectification module is designed, which utilizes the contextual information of key information extraction data through the analysis of similarity between labeled and unlabeled data, in order to improve the quality of pseudo-labels. Moreover, a dual branch adaptive alignment mechanism is also designed, in order to adaptively align intra-class variance and alleviate inter-class variation on key information extraction benchmarks, which is composed of two adaptive alignment ways, one is the intra-class alignment branch, which is designed to adaptively align intra-class variance, the other one is the inter-class alignment branch, which is developed to adaptively

alleviate inter-class variance changes on the representation level. The whole research process does main contribution in the area of key information extraction, and the extensive experiment results on two benchmarks demonstrate that the pseudo-label bias rectification achieves state-of-the-art performance and its performance surpasses the previous research by 0.0211.

A solution of keyphrase extraction is proposed during learning process for personalized recommendation [19], in order to represent multi-view knowledge. In this solution, the structural features and section texts obtained from the section structure information are utilized, in order to extract keyphrase. The approach proposed consists of two main parts, one part explores the effect of structural features on keyphrase extraction models, and the other part integrates the extraction results from all section texts used as input corpora for keyphrase extraction models via a keyphrase integration algorithm to obtain the keyphrase integration result. Furthermore, the effect is also examined for the classification quality of section structure on keyphrase extraction performance. The approach proposed is different from the traditional information extraction method, the results show that incorporating structural features improves keyphrase extraction performance, though different features have varying effects on model efficacy, and the keyphrase integration approach yields the best performance, the classification quality of section structure can affect keyphrase extraction performance, then these findings indicate that using the section structure information contributes to effective keyphrase extraction. So, the whole research result does main contribution in the area of information extraction, and addresses that the length of the research objective content is limited, the noise information is introduced for the research objective content, the keyphrase extraction performance diminishes.

A method for joint entity and relation extraction is proposed [20], in this method, the tasks of entity extraction and relation classification are integrated by sharing the encoding layer. However, this method faces challenges due to incongruities in the contextual information captured by these subtasks, resulting in potential feature conflicts and adverse effects on model performance. To address this problem, a novel joint entity and relation extraction method is introduced that incorporates multi-module feature information enhancement, and a relation awareness enhancement module is employed for the entity extraction task. In this module, the model's focus is directed towards extracting entities closely related to potential relations using a potential relation extraction technology and an attention mechanism. For the relation extraction task, an entity information enhancement module is implemented that uses entity extraction results to augment the original feature information through a gating mechanism, thereby enhancing relation classification performance. The whole research process does main contribution in the method of information extraction for entity and relation, and the experiments on the multi-source datasets demonstrate that the method proposed performs well. Compared to the state-of-the-art method, the F1 score improves by 0.007.

A model is constructed for investigating how to adapt BERT to Chinese text research on reducing data volume process oriented to text classification [21], during the construction process for this model based on deep learning framework and dataset, Firstly, the Chinese corpus is compiled from various sources, the Web pages are searched through Google and Baidu search engine API belonging to the Chinese text research domain, the articles are collected through the API library provided by Wikipedia, the scientific literature written in Chinese are downloaded from CNKI, WanfangDATA, and CQVIP based

on defined query keywords, a large number of documents for archived files are stored. Next, the domain Chinese word segmentation is finished based on construction of domain pre-training corpus, a recent work informask is proposed, which optimizes the masking strategy. Next, the Pre-trained Language Model is developed, in order to further improve the performance of downstream tasks and maximize the value of a large amount of unlabeled domain data. The whole research process does main contribution in the method of text mining and information extraction for domain knowledge, and the effectiveness of the model is validated using different downstream tasks such as named entity recognition, relation extraction, and event extraction, which can perform better than general models and promote information extraction and knowledge discovery from Chinese text.

Inspired by human reasoning, a graph-based multitask information extraction framework is presented that facilitates the interaction between several information extraction tasks capable of capturing both local and global information [22]. In this framework, the graphs are constructed by selecting the most confident entity spans and coupling them with a confidence-weighted relation type and a confidence-weighted coreference. Additionally, a dynamic span graph approach is employed, where span updates are propagated across both the coreference and the relation graph, which allows useful information to be learned from a broader context by enhancing interaction across different information extraction tasks. The input data are globally shared, and the interaction between subtasks is fully exploited, in order to avoid cascading errors. The whole research process does main contribution in the method of information extraction, and the experiments demonstrate that the proposed multitask information extraction framework outperforms the state-of-the-art in multiple information extraction tasks spanning a variety of datasets. The framework proposed is different from the traditional information extraction framework, which is shown to achieve state-of-the-art results on multiple information extraction tasks across various domains and the framework's ability to enhance interaction across tasks allows it to learn valuable information from a broader context.

Based on the above research status for the information extraction area, it can be summed up that most studies have adopted following methods including the induction way based on the wrapper, the extraction way based on the Web query, the extraction way based on the Ontology, the processing way based on the natural language, and the extraction way based on the HTML structure and so on. However, the above process cannot fully consider how to design a wrapper set, in order to extract instances of different categories or theme, how to express universal and effective extraction rules, and how to iteratively learn the structure of extraction target. If above details can be studied in depth, the complexity for network information extraction process can be reduced, the extraction accuracy for network information can be improved, the research result can play an important role in describing Web topic, reconfiguring the Web topic corpus, inferring Web hierarchical big data propagation path, and providing an intelligent big data warehouse for network information detection application. To solve the research problem, this paper will propose innovative incremental element extraction method based on the theme similarity measurement, in order to describe Web topic, the next section will complete the problem definition and highlight the problem research boundaries.

Table 1. The related research status

The research method	The research limitation	The research disadvantage	The proposed methodology
The induction way based on the wrapper [18]	A wrapper can only handle instances of one category, it is necessary for a wrapper set, in order to extract instances of different categories.	The scalability is lack	The algorithm of theme similarity measurement
The extraction way based on the Web query [19]	The extraction rules need to be expressed in the XSLT way	The complexity is high for the process of the algorithm implementation, and its application is lack.	The algorithm of incremental instance extraction for Web topic detection
The extraction way based on the Ontology [20]	There are specific requirements for the structure of extraction target	The scalability is lack	The algorithm of theme similarity measurement and incremental instance extraction for Web topic detection
The processing way based on the natural language [21]	The massive instances need to be learned, in order to get effective extraction rules.	The process is difficult for automatic extraction	The algorithm of incremental instance extraction for Web topic detection
The extraction way based on the HTML structure [22]	The hyperlinks are unable to be processed, so the information is only able to be extracted with obvious range structure.	The generality is lack	The algorithm of theme similarity measurement and incremental instance extraction for Web topic detection

3. Problem definition in incremental network information extraction

From a global perspective, every Web news report can be viewed an instance node in the authoritative news network. This instance node can also link multiple related instance nodes, therefore, social events supported by a set of instance nodes can be considered as a theme. This theme belongs to column node of Web news network again, so different dimension can link theme and multiple Web news instances, and it can be regarded as a hierarchical node. However, from a local perspective, when analysing structural characteristics of a Web news instance separately, it can be found that it usually contains two parts. One part is text information related to Web news reports, and the other part is noise information which is not related to Web news reports. Therefore, if Web news instances and their relationships can be deployed in tree structure, and noise information which is not related to Web news content can be filtered in process of extracting Web news information, it will provide a hierarchical and high-quality Web news corpus for continuing to analyse Web news.

In text messages related to Web news reports, its content has presented unstructured characteristics, and it has a certain degree of difficulty for information extraction process. By analysing release template used in Web news, it can be seen that there are two parts in ¡title¿ element tag. One part is headline of Web news, another part is name of issuing organization, and two parts are separated with underline. The headings of Web news are also included in ¡h1¿ element tag, Web news release time and source are included in ¡div¿ element tag, and the text of Web news is included in ¡p¿ element tag. When analysing release template used in Web news, it can be found that the location of Web news instance data item can be determined from perspective of perceived styling features for Web news content. For example, the important content that needs to be highlighted is usually controlled by tagging elements of ¡strong¿ and ¡h1¿ in Web news. When analysing template used to publish multiple Web news on different websites, some templates can be found to have something in common. For example, the text of Web news content can be deployed in element tags of ¡p¿ and ¡div¿, and it has a certain length. Therefore, if unstructured content of Web news can be stored with semi-structured content in process of extracting Web news information, it will provide a semi-structured and high quality Web news corpus for continuing to analyse Web news.

This paper primarily defines data structures for NewsSet, HyperLinkSet, UrlSet, TopKeyWordSet, InitialUrlQueue and WaitingUrlQueue, as shown in Table 2. In addition to the defined data structures, this paper also defines Pattern and BasePattern. Pattern is a filter mode, which can filter Web news URL of sublayer. BasePattern is a base filter mode, which can filter Web news URL of brotherhood. Effective sublevel link groups can be presented with $UV = \{uv_1, uv_2, \ldots, uv_n\}$, invalid sublevel link groups can be presented with $UN = \{un_1, un_2, \ldots, un_n\}$ corresponding to toplevel link groups extracted from UrlSet set, which is $U = \{u_1, u_2, \ldots, u_n\}$. If filter mode Pattern exists, UV and UN can be detected, and UN can be filtered. If filter mode $BasePattern \in uv_i$ $(i = 1, 2, \ldots, n)$ exists, and there is no its submode that belongs to BasePattern or uv_i corresponding to valid sublayer link groups filtered from toplevel link groups, BasePattern is base filter mode of sublayer link groups.

The problem that need to be solved by the incremental network information extraction method for Web topic detection is as follows, the instances are extracted from mas-

Table 2. The data structure definition

The data structure name	The data structure description	The data structure representation
NewsSet	A URL set of Web news The seed big data source extracted from Web news information The massive and authoritative URL in Web news network	{ns1, ns2, ns3, …, nsi-1, nsi, nsi+1, …, nsn}
HyperLinkSet	The hyperlinks of massive instances for Web news contained in NewsSet	{hlsi1, hlsi2, hlsi3, …, hlsi(j-1), hlsij, hlsi(j+1), …, hlsim}
UrlSet	The big data source extracted from Web news information The massive and authoritative instances in Web news network	{us1, us2, us3, …, usi-1, usi, usi+1, …, usn}
TopKeyWordSet	A theme set of Web news The theme extracted from Web news information The keywords in social events that occur	{tkws1, tkws2, tkws3, …, tkwsi-1, tkwsi, tkwsi+1, …, tkwsn}
InitialUrlQueue	An initial queue for storing Web news URL	{iuq1, iuq2, iuq3, …, iuqi-1, iuqi, iuqi+1, …, iuqn}
WaitingUrlQueue	A pending queue for storing Web news URL	{wuq1, wuq2, wuq3, …, wuqi-1, wuqi, wuqi+1, …, wuqn}

sive Web news related theme, and the filtering mode from parent layer to sub layer and base filtering mode in brotherhood layer can be detected from these instances, in order to transfer tree structure from graph structure for interlacing complex URL network system. In this process, when the specific Web news instances belonging to certain topics are unknown, manually filtering each instance will inevitably increase computational complexity. Therefore, the problem that needs to be solved reflects uncertainty. If the set of Web news instances to be extracted is ns1, ns2, ns3, ..., nsi-1, nsi, nsi+1, ..., nsn, then each Web news instance can be considered as a node in the set. If the extraction reduction is designed in the path of extracting objects, the length of the extraction path can be shortened. In the case where each URL of a Web news instance can uniquely match the node nsi in the set of Web news instances, there is no polynomial time complexity algorithm to solve the proposed problem. The result can be described with TopWebNews=twn1, twn2, twn3, …, twni-1, twni, twni+1, …, twnk, the range for parameter i value is from one to k. The twni.url saves address for Web news, twni.title saves title for Web news, twni.pubtime saves release time for Web news, twni.pubsource saves release source for Web news, twni.content saves text for Web news, twni.dividedtitle saves segmentation result for Web news headline, twni.dividedcontent saves segmentation result for Web news text, twni.contentkeyword saves keywords for Web news content, twni.relativityvalue saves similarity between Web news content and theme, twni.parenturl saves address of parent node for Web news instances, twni.pattern saves description for Web news instance filtering mode, twni.systemtime saves system time extracted for instances.

Given the problem definition for incremental network information extraction in Section 3, the next section will introduce the framework, method, and algorithm of incremen-

tal network information extraction for Web topic detection based on big data, which aims to improve computational efficiency and accuracy.

4. Proposed Methodology: The incremental network information extraction for Web topic detection

In view of frequent events in society, the released Web news has reached at least NB level, and has shown characteristics of 5V big data [23,24,25]. Based on above problem definition, this paper proposes an incremental network information extraction method for Web topic detection, as shown in Fig. 1.

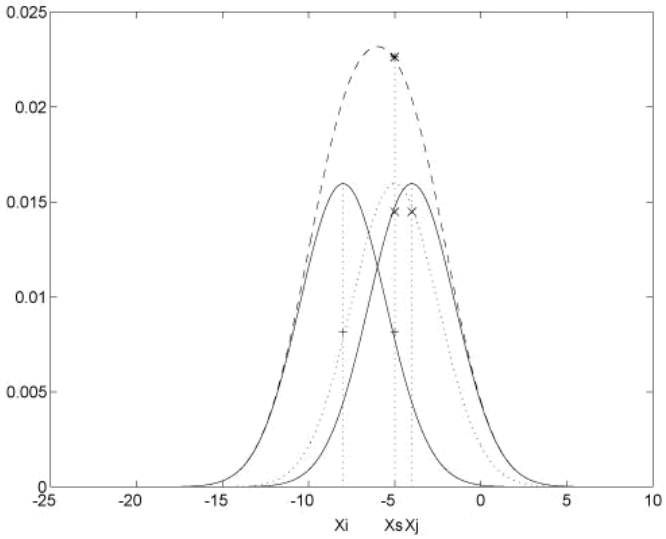

Fig. 1. The framework of incremental network information extraction

This framework completes the incremental corpus extraction for Web topic detection through using set of Web news and theme words and so on, this framework can measure Web news theme similarity. Through using selected Web news instance URL and source code related to theme, this framework can extract Web news instance information. Through using queue of InitialUrl and WaitingUrl, this framework can calculate importance of Web news links. Through using extracted Web news instance information, this framework can analyse mode of filtering and base filtering, and under background of theme, this framework can extract Web news instances incrementally, the result of keyword extracted can describe Web topic. In a word, this paper can effectively extract network information for massive Web news that report social events using this framework, and designs following algorithms in order to research a network information extraction method for Web topic detection.

4.1. The algorithm of theme similarity measurement for network information extraction

The design idea of theme similarity measurement algorithm is as follows, according to TopKeyWordSet, this algorithm can extract Web news instances related to theme from NewsSet and UrlSet, and calculate importance of Web news instances related to theme from InitialUrlQueue. The input content of this algorithm is set of Web news and theme words and so on, the output content of this algorithm is Web news instance information related to theme, the construction process is as follows.

Under background of social event occurrence, according to TopKeyWordSet, this process can form theme vector. As shown in Formula 1, si represents social event, tkwsij.weightvalue represents weight of theme word tkwsij.wordvalue and constitutes the component values of the theme vector. This vector is not constant, when massive information for Web news is extracted, according to its keyword set, this process can conduct iterative processing for theme words and weight value, and continue to learn extraction results of massive information for Web news. Among this process, the value of the number of theme words is determined in subsequent experiments.

$$\begin{gathered}\{s_i, (tkwsi_1.\text{wordvalue}, tkwsi_1.\text{weightvalue}), (tkwsi_2.\text{wordvalue}, \\ tkwsi_2.\text{weightvalue}), \ldots, (tkwsi_j.\text{wordvalue}, tkwsi_j.\text{weightvalue}), \\ \ldots, (tkwsi_n.\text{wordvalue}, tkwsi_n.\text{weightvalue})\}, \\ \text{The theme vector: } [tkwsi_1.\text{weightvalue}, tkwsi_2.\text{weightvalue}, \ldots, \\ tkwsi_j.\text{weightvalue}, \ldots, tkwsi_n.\text{weightvalue}]\end{gathered} \tag{1}$$

Through using open source library NekoHtml, this process can extract source code of Web news instance page, extract contents of ¡title¿ element label, and attribution value of content in ¡meta name=”keywords” content=”*”/¿ element tag, form element vector, and calculate similarity between Web news instance and theme as shown in Fig. 2. VectorTheme represents theme vector, VectorElement represents element vector, Relativity(D,Theme) represents the calculated result, which is compared with similar threshold value determined in subsequent experiments. If it is greater than or equal to this value, then it is similar to theme, conversely, it is not similar to theme.

As shown in Fig. 3, this process can calculate similarity between words. d1 represents the shortest distance between two words in WordNet, d2 represents depth of two words belong to same category in WordNet.

This process designs regular expressions, in order to eliminate interference of noise information for extracting Web news content. This process locates location of distribution for data items in Web news content, and extracts data items from Web news instance content related to theme. If there are URL sets pointing to next link target in extracted content, then it can be enqueued to InitialUrlQueue. This process calculates importance of similarity with theme; according to importance, instances can be ranked from high to low in InitialUrlQueue queue, and it can be enqueued to WaitingUrlQueue. When queue is not empty, this process calculates similarity between instances and theme. The importance of instance may be lower than its parent layer URL in InitialUrlQueue queue, but information contained in it may not exist in the parent layer URL instance. Therefore, this process conducts priority processing for link instances that have higher importance based on importing genetic factor, which is represented with σ, in order to ensure integrity of

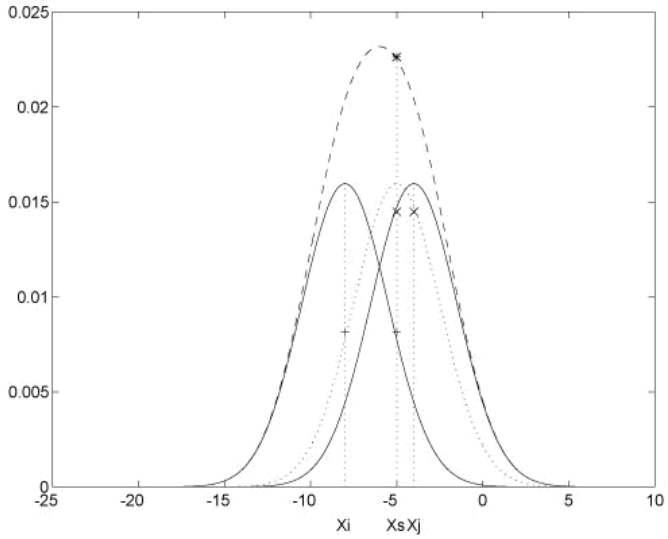

Fig. 2. The element vector and the computation process for Relativity(D,Theme)

information extraction, and consider how to do partial optimization as shown in Fig. 4. Among this process, the adjustment of factors is determined in subsequent experiments.

4.2. The algorithm of incremental instance extraction for Web topic detection

The design idea of incremental instance extraction algorithm is as follows, according to extracted Web news instance data items, this algorithm can analyze Pattern and BaseP-attern. According to TopKeyWordSet, this algorithm can conduct segmentation for Web news title and text content, and extract keywords, the result of keyword extracted can describe Web topic. Based on the imported self-adaption strategy, this algorithm can adjust size of theme vectors and similar thresholds. The input content of this algorithm are extracted Web news instance data items and TopKeyWordSet, the output content of this algorithm are filtering mode and Web news instances content extracted under background of theme incrementally, the construction process is as follows.

This process can extract URL information from Web news instances using ”//” separator as identification mark, and select its top-level nodes. Use ”/” separator as identification mark, this process can select its hierarchical nodes. The selected nodes are constructed into tree structure, and this process can analyze filtering mode Pattern among nodes for different levels. Grouping node tree structure one by one, according to filtering mode Pattern of previous level node for current node group, this process can analyze base filtering mode for regular hierarchical nodes as shown in Fig. 5.

This process can conduct segmentation for Web news title and text content, but in its result, there are some words that have nothing to do with topic detection. For example, the existence of words annotated as /p and /d and so on will contribute less to analysis of Web news texts, it will not only reduce efficiency and quality of Web news text analysis, and also reduce accuracy of topic detection. Based on occurrence of unexpected events, the

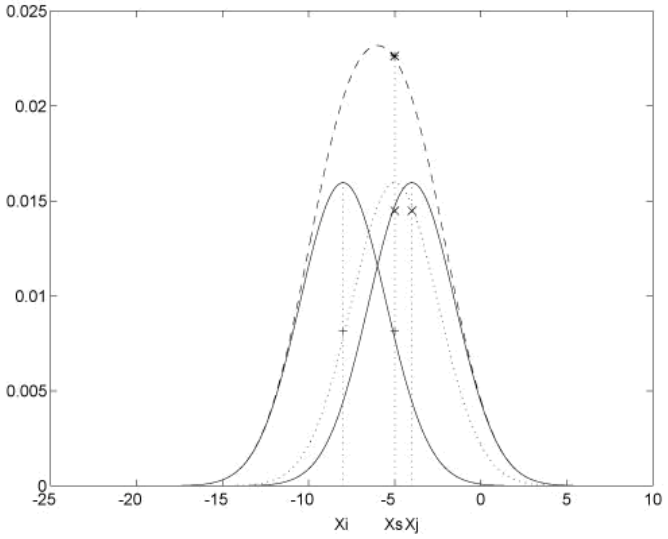

Fig. 3. The computation process for $\mathrm{Sim}(word_1, word_2)$

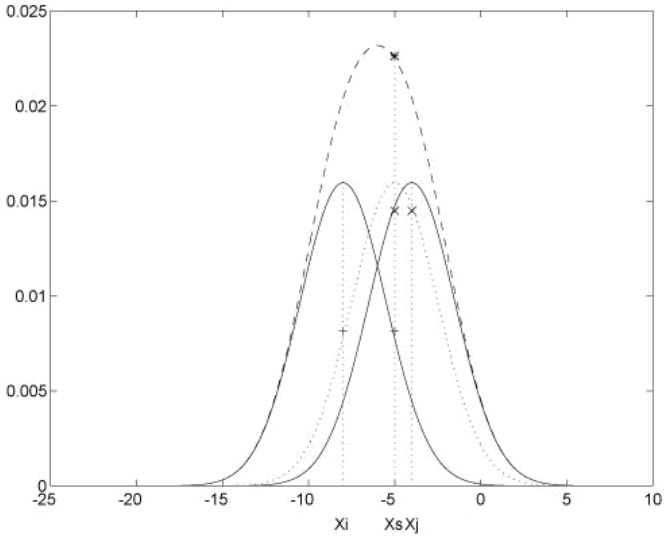

Fig. 4. The computation process for Relativity

Algorithm 1 Theme Similarity Measurement

Input: UrlSet, NewsSet, VectorTheme, Threshold, Parameters, InitialTime, T
Output: RTSet
WebNews ← (UrlSet, NewsSet)
for each $w[i]$ ($0 \leq i \leq w.\text{size}() - 1$) **do**
  sc ← Extract $w[i]$ target source code
  ec ← Extract sc element content
  VectorElement ← Generate element vector(ec, VectorTheme)
  Relativityparent ← Calculate $w[i]$ theme similarity(VectorElement, VectorTheme)
  if Relativityparent ≥ Threshold **then**
    rt ← Extract sc data items
    RTSet.addlist(rt)
    $urls$ ← Extract sc url
    EnInitialUrlQueue($urls$)
    if Time interval is T **then**
      $urlscobj$ ← Extract $urls$ target source code
      $urlecobj$ ← Extract $urlscobj$ element content
      VectorElementObj ← Generate element object vector($urlecobj$, VectorTheme)
      Relativity ← Calculate $urls$ theme similarity(VectorElementObj, VectorTheme)
      Call genetic factor strategy(Parameters)
      RankInitialUrlQueue(Relativity)
      DeInitialUrlQueue($urls$)
      EnWaitingUrlQueue($urls$)
      InitialTime ← Adjust InitialTime
    end if
  end if
end for
return RTSet

word segmentation process has some difficulties in detecting new words, so, it is impossible to accurately represent new words for emergencies in word segmentation result, these words are quite important in topic detection. The detection of these words will contribute more to Web news text analysis, it can not only improve efficiency and quality of Web news text analysis, and also improve accuracy of topic detection. Therefore, this process designs the corpus for filtering word, according to part of speech tagging and TopKeyWordSet in word segmentation result, and the corpus for filtering word, on the one hand, it can filter words that are not meaningful, on the other hand, it can detect new words with practical meaning, it will provide an effective segmentation result for Web news text analysis.

This process calculates weight of words in Web news instances, and extracts keywords as shown in Fig. 6. F(KeyWord,D) represents frequency of KeyWord occurrence in Web news instances, N represents total number of Web news instances involved in computing, n represents number of Web news instances containing KeyWord involved in computing for Web news group, Weight(KeyWord,T) represents weight in theme set for KeyWord, in order to consider importance of same word under different theme background.

On basis of traditional vector space model, this process imports self-adaption strategy, in order to consider feedback and guidance on theme in dynamically increasing Web

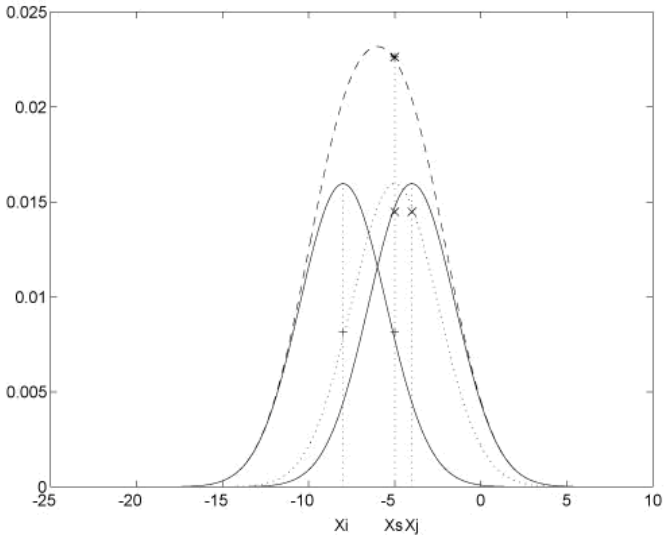

Fig. 5. URL network diagram structure is converted into tree structure

news instances. Within time interval, according to subsequent feedback information, this process automatically adjusts size of theme vectors and similar thresholds. The adjustment parameters are determined in subsequent experiments, if threshold increases, then it can improve accuracy of content extraction. However, when the calculated similarity of theme is generally low, if threshold is lowered, then scope of content extraction can be expanded. As shown in Fig. 7, Sumt1 represents the number of Web news instances extracted in t1 time, Sume represents the number of Web news instances extracted, it is expected to be within t time interval, Sumt2 represents the number of Web news instances extracted with t time interval, Sumt1/theme represents the number of Web news instances extracted that are similar to theme in t1 time.

Based on the method proposed in this paper, the effect of network information extraction for Web topic detection can be obtained, it can represent the extraction result for massive Web news in the context of social events. In the subsequent research, the analysis for semantic feature, feature evaluation and use behavior tracking on the extracted big data corpus can be conducted to further enhance the research value of incremental network information extraction for Web topic detection. To reflect the feasibility, verifiability and superiority for the method proposed in this section, the next section will complete the experimental analysis process.

5. The process of experimental analysis

This section will describe the experimental setup firstly, then analyze the evaluation metrics for the research method proposed in this paper, and then compare the performance of the experimental effect.

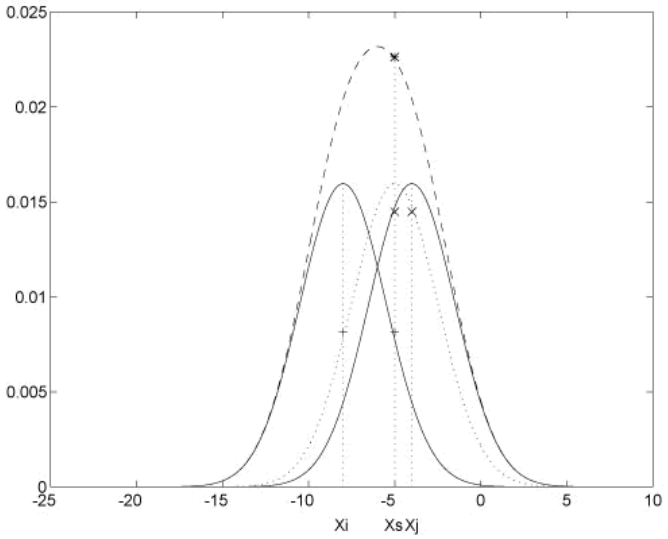

Fig. 6. The computation process for Weight(KeyWord,D)

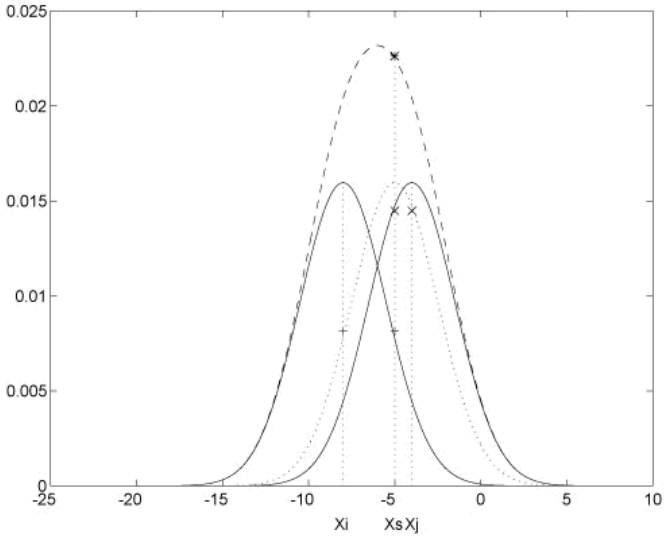

Fig. 7. The computation process for Threshold

Algorithm 2 Incremental Instance Extraction

Input: RTSet, Text corpus, Parameters, InitialTime, T
Output: Pattern, BasePattern, KeyWordSet
WebNews ← (RTSet)
for each $w[i]$ ($0 \leq i \leq w$.size() $- 1$) **do**
  Pattern ← Recursive analysis for $w[i].url$ filter mode
  BasePattern ← Recursive analysis for $w[i].childurl$ filter basemode
  ws ← WordSegment($w[i].dataitems$)
  $kwobj$ ← ExtractKeyWords(ws, Text corpus)
  for each $kw[i]$ ($0 \leq i \leq kwobj$.size() $- 1$) **do**
    KeyWordSet.addlist($kw[i]$)
  end for
  if Time interval is T **then**
    Call adaptive strategy(Parameters, KeyWordSet)
    InitialTime ← Adjust InitialTime
  end if
end for
Describe Web topic(KeyWordSet)

5.1. The experimental setup

Based on the design ideas and algorithms proposed in this paper, the hardware and software environments used in the experimental process are as follows. The processor is Intel 2.40GHz, the memory is 64GB, and the operating system is 64 bit Windows. The programming language is Java, mainly used for algorithm implementation. The network application research and development platform is MyEclipse, and the database management system is SQL Server, mainly used for storing and processing Web big data extracted [26,27,28]. The Web Project has been published on Big Data Analysis and Mining of My Teaching Classroom and My Practical Project for Tou Ge Practice Teaching Platform including program and data, due to dependency on the General Project of Science and Technology Plan of Beijing Municipal Education Commission, the Research Project on Graduate Education Science at Beijing Union University in 2025, and the Support Project of High-Level Teachers in Beijing Municipal Universities in the Period of 13th Five-Year Plan for this Web Project, therefore, it has been used in both graduate practical courses and undergraduate applied courses.

This paper uses the massive Web news generated by the German A320 aircraft crash event as the extraction target for network big data, these Web news were all published by authoritative websites. This event has gone through the process of beginning, development, and end, and the data is authentic, the experimental analysis process can verify the feasibility and effectiveness of the design ideas and algorithm design proposed in this paper.

5.2. The evaluation metrics

In the following experiments, the Precision evaluation index is used to measure not only the ratio of correctly extracted Web news instances to all extracted Web news instances that are similar to the theme, but also the ratio of correctly detected topics to all detected

topics, this ratio reflects the accuracy of extraction and detection. Based on the comprehensive consideration of Precision and Recall evaluation indicators, the F-Measure evaluation index is used to measure the overall extraction effect, this is a comprehensive performance of accuracy and comprehensiveness. Among them, the Recall evaluation index is used to measure the ratio of correctly extracted Web news instances to all Web news instances that should be extracted, this ratio reflects the comprehensiveness of extraction. In the evaluation process, the accurately extracted Web news instances can be annotated by automatically calculating the similarity between the keywords of the extracted Web news instance and the topic words.

In the following experiments, this paper also uses four other real data sets including the events of Shanghai Bund trample, Taiwan revival airliner falling river, Nepal 8.1 earthquake and Orient Star cruise overturn as the extraction target for network big data, in order to verify whether the proposed method in this paper is more universal and reliable for the evaluation index.

5.3. The performance comparison

Firstly, this section analyzes the accuracy of extracted Web news instances that are similar to the theme under different Web news information extraction methods as shown in Fig. 8 and Table 3, the accuracy represents whether the Web news instances extracted through two methods that are similar to the theme belong to the strongly or weakly associated annotation category with the theme. The red solid line represents the change in accuracy of Web news instances extracted under the theme matching method that are similar to the theme, from its trend, it can be seen that the extraction process is more open without detailed knowledge of the theme reported in Web news. With the continuous increase in the number of Web news, the accuracy has remained relatively stable, failing to break through 0.74. The blue solid line represents the change in accuracy of Web news instances extracted under the algorithm designed in this paper that are similar to the theme, from its trend, it can be seen that the information extraction filtering mode derived from continuously analyzing the extracted Web news instances has played a role in improving accuracy. Although its accuracy is similar to that of theme matching method when the number of Web news is small, as the number of Web news continues to increase, its accuracy also keeps climbing, reaching up to 0.83. So, this experiment demonstrates that the quality of Web news instances extracted by the algorithm in this paper is higher than that of theme matching method.

Next, this section analyzes the F-Measure values of extracted Web news instances that are similar to the theme under different threshold conditions to obtain the optimal threshold value as shown in Fig. 9 and Table 4, the F-Measure value represents the comprehensive quality of Web news instances extracted that are similar to the theme through the algorithm proposed in this paper, while continuously adjusting the threshold value. The red solid line represents the variation of F-Measure values under the theme background when the threshold takes different values, from its trend, it can be seen that when the threshold value is low, the comprehensive quality of Web news instances extracted that are similar to the theme is not high with the F-Measure value of approximately 0.6. As the threshold value increases, the comprehensive quality of Web news instances extracted that are similar to the theme also increases, and the F-Measure value also increases accordingly. When the threshold value is adjusted to about 0.65, the comprehensive quality of Web

news instances extracted that are similar to the theme reaches its highest level. When the threshold value is further increased, some Web news instances weakly associated with the extraction theme are not extracted, and the F-Measure value also decreases accordingly. The blue solid line represents the variation of F-Measure values under the method proposed in this paper when different threshold values are taken, from its trend, it can be seen that when the threshold value is low, the comprehensive quality of Web news instances extracted that are similar to the theme is high with the F-Measure value of approximately 0.75. As the threshold value increases, the comprehensive quality of Web news instances extracted that are similar to the theme remains stable, and the F-Measure value remains around 0.8. When the threshold value is adjusted to about 0.65, the comprehensive quality of Web news instances extracted that are similar to the theme reaches its highest level. When the threshold value is further increased, some Web news instances that are strongly or weakly associated with the extraction theme are not extracted, and the F-Measure value also decreases accordingly with a greater decrease than in the background of the theme. Before adjusting the threshold value to 0.78, the F-Measure value under the method proposed in this paper is higher than that under the theme background. In the context of the theme, when the threshold value is 0.65, the corresponding F-Measure value reaches its maximum, which is close to 0.75. Under the method proposed in this paper, when the threshold value is 0.65, the corresponding F-Measure value also reaches its maximum, which is close to 0.83. So, this experiment indicates that the optimal threshold value can be 0.65.

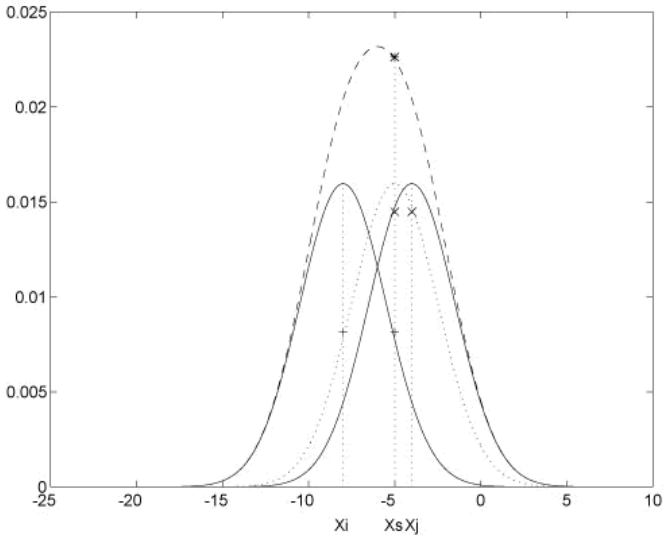

Fig. 8. The quality of Web news instances extracted that are similar to the theme under different methods

Next, this section analyzes the impact of Web news quantity extracted and threshold on the topic detection quality as shown in Fig. 10, the accuracy represents the quality of

Table 3. The comparison of accuracy rate for Web news instances extraction quality that is similar to the theme under different methods

The number of Web news (Unit: K) / The research method	20	40	60	80	100	120	140	160	180	200
Theme matching method	0.652	0.67	0.736	0.726	0.659	0.67	0.738	0.684	0.71	0.696
This paper's algorithm	0.65	0.664	0.694	0.725	0.754	0.764	0.794	0.804	0.822	0.833

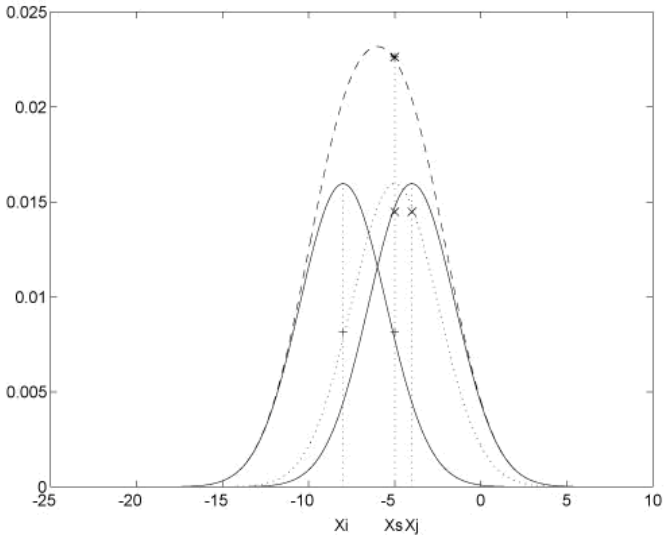

Fig. 9. The trend of F-Measure value variation with threshold value

Table 4. The comparison of F-Measure value for Web news instances extraction quality that is similar to the theme under different threshold conditions

The threshold value / The research method	0.1–0.15	0.2–0.25	0.3–0.35	0.4–0.45	0.5–0.55	0.6–0.65	0.65–0.7	0.75–0.8	0.85–0.9
The theme background	0.6–0.61 ↑	0.62–0.63 ↑	0.64–0.641 ↑	0.651–0.661 ↑	0.671–0.681 ↑	0.691–0.75 ↑	0.75–0.708 ↓	0.635–0.628 ↓	0.61–0.569 ↓
The incremental method	0.75–0.801 ↑	0.801–0.817 →↑	0.753–0.772 ↓↑	0.763–0.774 ↓↑	0.804–0.797 ↑↓	0.789–0.83 ↓↑	0.83–0.767 ↓	0.683–0.611 ↓	0.537–0.46 ↓

topic detection by adjusting the X-axis threshold and continuously increasing the number of Web news instances extracted on the Y-axis. When the threshold value is constant, it can be seen from the graph that as the number of Web news instances extracted continues to increase, the accuracy shows a trend of first increasing and then decreasing. The reason is that when the number of Web news instances extracted is small, it is not yet possible to fully understand the content reported by Web news. When there are a large number of Web news instances extracted, some Web news instances extracted that do not support topics are also analyzed, resulting in a decrease in accuracy. When the number of Web news instances extracted is constant, it can be seen from the graph that as the threshold value increases, the accuracy shows an increasing trend. The reason is that when the threshold value is low, most Web news instances extracted that do not support topics are extracted. When the threshold value is high, only a small number of Web news instances extracted that do not support topics are extracted. This experiment shows that when the number of Web news instances extracted is 100K and the threshold is 0.9, the quality of topic detection can reach the highest value, which is close to 0.82.

Next, this section analyzes the accuracy of topic detection under different numbers of theme keywords to obtain the optimal value for theme keywords quantity as shown in Fig. 11 and Table 5, the accuracy represents the quality of topic detection through the algorithm proposed in this paper with setting different numbers of theme keywords. The red solid line represents the change in accuracy of the theme background when the number of theme keywords is set differently, from its trend, it can be seen that when the number of theme keywords is set 3, the quality of topic detection tends to stabilize at approximately 0.682 to 0.685. The reason is that when the number of theme keywords is too small or too large, some Web news instances extracted that do not support topics are extracted. The blue solid line represents the change in accuracy of the incremental method when the number of theme keywords is set differently, from its trend, it can be seen that when the number of theme keywords is set 3, the quality of topic detection tends to stabilize at approximately 0.701 to 0.704 for the same reason as the red solid line trend. Overall, the accuracy under the incremental method is higher than that under the theme background, this experiment shows that the optimal value for the number of theme keywords can be set 4, so that under the algorithm proposed in this paper, the quality of topic detection is locally stable to the maximum value.

Next, this section analyzes the impact of the adjustment parameters in the algorithm on the topic detection quality to obtain the optimal adjustment range for these parameters as shown in Fig. 12, the accuracy represents the quality of topic detection when adjusting parameters for theme similarity measurement and incremental instance extraction algorithms. The red dashed line represents the variation in accuracy when the Alpha parameter takes different values for Formula 6, from its trend, it can be seen that when the Alpha value is adjusted between 1.15 and 1.3, the quality of topic detection is high and stable with an accuracy rate of approximately 0.76. The blue dashed line represents the change in accuracy when the Mu parameter takes different values for Formula 6, from its trend, it can be seen that when the Mu value is adjusted between 0.75 and 0.9, the quality of topic detection is high and stable with an accuracy rate of approximately 0.70. The green dashed line represents the change in accuracy when the Beta parameter takes different values for Formula 6, from its trend, it can be seen that when the Beta value is adjusted between 1.2 and 1.45, the quality of topic detection is high and stable with an accuracy

rate of approximately 0.77. The yellow dashed line represents the variation in accuracy when the Gamma parameter takes different values for Formula 6, from its trend, it can be seen that when the Gamma value is adjusted between 0.8 and 0.95, the quality of topic detection is high and stable with an accuracy rate of approximately 0.76. The pink dashed line represents the variation in accuracy when Sigma parameters take different values for Formula 4, from its trend, it can be seen that when the Sigma value is adjusted between 0.8 and 0.95, the quality of topic detection is high and stable with an accuracy rate of approximately 0.81. Overall, the adjustment of various parameters can locally stabilize the quality of topic detection to its maximum value, and determine the optimal adjustment range of each parameter.

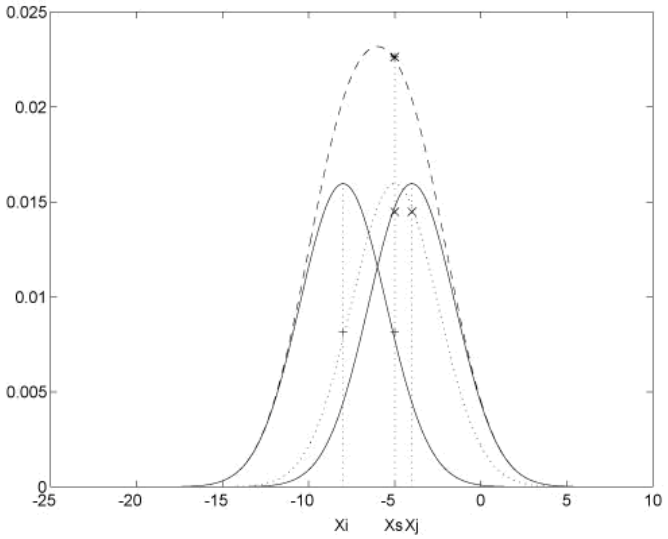

Fig. 10. The trend of accuracy changing with the number of Web news extracted and threshold value

Table 5. The comparison of accuracy rate for the topic detection quality under different numbers of theme keywords

The numbers of theme keywords / The research method	1	2	3	4	5	6	7	8
The theme background	0.625	0.684	0.682	0.685	0.682	0.685	0.662	0.662
The incremental method	0.708	0.703	0.701	0.704	0.701	0.704	0.718	0.719

Finally, this section analyzes the quality of topic detection on different data sets based on the extraction result for Web news as shown in Fig. 13 and Table 6, from its trend, it can be seen that under the method proposed in this paper, there is almost no difference in the quality of topic detection for the five events at the start, development, and end stages, this indicates that the quality of topic detection using the method proposed in this paper

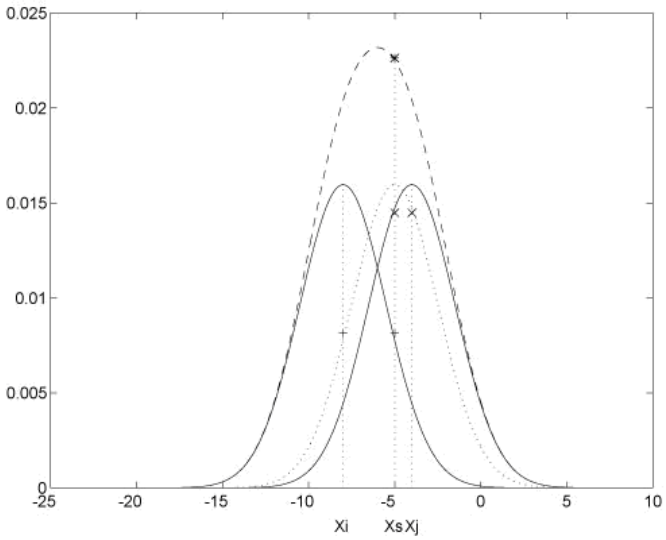

Fig. 11. The trend of accuracy changing with the number of theme keywords

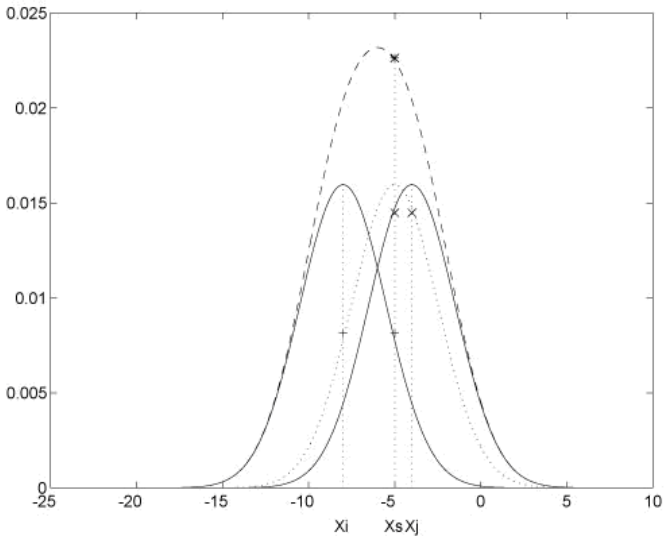

Fig. 12. The trend of accuracy changing with the adjustment for various parameters in the algorithm

is relatively stable under different events. However, as the number of Web news extracted continues to increase, it still has a certain impact on the quality of topic detection at different stages of the event. This is because, with the increase in the number of Web news extracted, the extraction result under the theme is constantly expanding, and the support for topic detection is also improving.

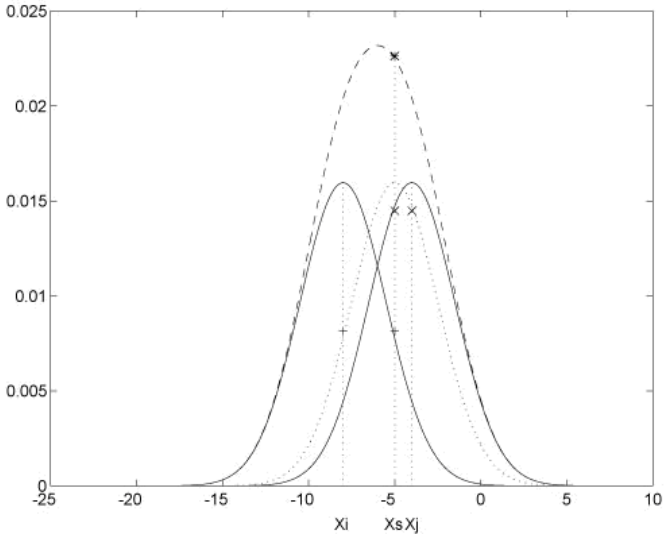

Fig. 13. The quality of topic detection on different data sets based on the extraction result

Table 6. The comparison of accuracy rate for the topic detection quality on different data sets based on the extraction result

The data sets / The evolution stage	Shanghai Bund trample	Taiwan revival airliner falling river	German A320 airliner crash	Nepal 8.1 earthquake	Orient Star cruise overturn
Start stage (About 50K)	0.701	0.698	0.706	0.692	0.704
Development stage (About 140K)	0.727	0.718	0.727	0.715	0.728
End stage (About 200K)	0.753	0.748	0.753	0.743	0.755

6. Conclusion

The research on the incremental network information extraction technology for Web topic detection, taking the network big data of Web news as the research object, has been completed, the results of this design and implementation are more valuable to scholars in related research fields. In the process of technology research, this paper proposes the theme element extraction module, the theme similarity calculation module, the instance data items extraction module, the link importance calculation module, the keywords extraction module, and the incremental instance extraction module, and proposes the algorithm

of the theme similarity measurement for incremental network information extraction and the incremental instance extraction for Web topic detection, in order to address the shortcomings in the current research status. The experimental analysis process shows that the method proposed in this paper is feasible, verifiable, and superior. It has played an important role in reconfiguring numerical intelligence warehouses for detecting Web topic, and inferring the Web hierarchical big data propagation path.

In subsequent research, the optimization of incremental network information extraction algorithm for Web topic detection can be continued, and the optimal range of parameters in the algorithm can be refined through experiments to handle fuzzy or conflicting semantic information, enhance the comprehensiveness, accuracy, and robustness of incremental network information extraction for Web topic detection. The research results can be applied to real-time Web news monitoring and multilingual Web topic detection processes, and further improve the processing efficiency for big data applications.

Acknowledgement. This paper is supported by General Project of Science and Technology Plan of Beijing Municipal Education Commission under Grant Nos. KM202011417011, Research Project on Graduate Education Science at Beijing Union University in 2025 under Grant Nos. YK202502, Support Project of High-Level Teachers in Beijing Municipal Universities in the Period of 13th Five-Year Plan under Grant Nos. CIT&TCD201704072.

References

1. Li, P., Zhang, L.: Application of big data technology in enterprise information security management. Scientific Reports 15(1), 1–5 (2025)
2. Arunkumar, M., Rajkumar, K., Jeyaseelan, W., Natraj, N.: Data mining, machine learning, and statistical modeling for predictive analytics with behavioral big data. Tehnicki Vjesnik - Technical Gazette 32(1), 72–74 (2025)
3. de Miguel, A., Sarasa-Cabezuelo, A.: A global approach to artificial intelligence. IEEE Access 13, 76946–76950 (2025)
4. Kaushik, M., Sharma, R., Koiva, P., Fister, I.J., Draheim, D.: An exhaustive multi-aspect analysis of swarm intelligence algorithms in numerical association rule mining. IEEE Access 12, 138985–138989 (2024)
5. Tang, J., Yan, Y., Bao, J., Huang, B.: Big data-driven control of nonlinear processes through dynamic latent variables using an autoencoder. IEEE Transactions on Cybernetics 55(5), 2411–2415 (2025)
6. Song, S., Pan, L., Liu, S.: A q-learning based auto-scaling approach for provisioning big data analysis services in cloud environments. Future Generation Computer Systems 154, 140–144 (2024)
7. Wang, S.: Research on the digital marketing strategies in the e-commerce logistics service mode under the influence of big data. Computer-Aided Design and Applications 21(S4), 39–43 (2024)
8. Wang, H., Zhang, S.: Research on the application of improved bert-dpcnn model in chinese news text classification. Concurrency and Computation: Practice and Experience 37(3), 1–3 (2025)
9. Tredinnick, L.: The intricate web: Network and rhizome metaphors in hypertext and the web and the epistemic challenge of fake news. Journal of Documentation 79(6), 1485–1489 (2023)
10. Li, T., Yu, J., Zhang, H.: Web of things based social media fake news classification with feature extraction using pre-trained convoluted recurrent network with deep fuzzy learning. Theoretical Computer Science 931, 65–69 (2022)

11. Li, X., Gao, N., Wang, Y.: Identifying disruptive technology using saox semantic analysis and web news data mining: A perspective of technology convergence. IEEE Transactions on Engineering Management 72, 2116–2120 (2025)
12. Mallick, P., Mishra, S., Chae, G.: Digital media news categorization using bernoulli document model for web content convergence. Personal and Ubiquitous Computing 27(3), 1087–1091 (2023)
13. Dritsas, E., Trigka, M.: Database systems in the big data era: Architectures, performance, and open challenges. IEEE Access 13, 95068–95072 (2025)
14. Angskun, T., Sritha, K., Srithong, A., Khopolklang, N., Kamollimsakul, S., Phithak, T., Angskun, J.: Using big data to assess an affective domain for distance education. Future Generation Computer Systems 160, 131–134 (2024)
15. Xi, Q., Jiang, P.: Design of news sentiment classification and recommendation system based on multi-model fusion and text similarity. International Journal of Cognitive Computing in Engineering 6, 44–47 (2025)
16. Wu, C., Wu, F., Huang, Y., Xie, X.: Personalized news recommendation: Methods and challenges. ACM Transactions on Information Systems 41(1), 1–4 (2023)
17. Xu, H., Peng, Q., Liu, H., Sun, Y., Wang, W.: Group-based personalized news recommendation with long and short term fine-grained matching. ACM Transactions on Information Systems 42(1), 1–6 (2023)
18. Guo, P., Song, Y., Wang, B., Liu, J., Zhang, Q.: Plbr: A semi-supervised document key information extraction via pseudo-labeling bias rectification. IEEE Transactions on Knowledge and Data Engineering 36(12), 9025–9036 (2024)
19. Chang, C., Tang, F., Yang, P., Zhang, J., Huang, J., Li, J., Li, Z.: Multi-view knowledge representation learning for personalized news recommendation. Scientific Reports 15(1), 1–13 (2025)
20. Li, Y., Yan, H., Zhang, Y., Wang, X.: Joint entity and relation extraction combined with multimodule feature information enhancement. Complex & Intelligent Systems 10(5), 6633–6645 (2024)
21. Serreli, L., Marche, C., Nitti, M.: Reducing data volume in news topic classification: Deep learning framework and dataset. IEEE Open Journal of the Computer Society 6, 152–163 (2025)
22. Zheng, Y., Tuan, L.: A novel, cognitively inspired, unified graph-based multi-task framework for information extraction. Cognitive Computation 15(6), 2004–2013 (2023)
23. Zhang, S., Tan, F., Peng, H.: Sample size determination for multidimensional parameters and the a-optimal subsampling in a big data linear regression model. Journal of Statistical Computation and Simulation 95(3), 628–632 (2025)
24. Matthews, S.: Review of statistical learning for big, dependent data. Journal of Official Statistics 40(4), 849–851 (2024)
25. Liu, H., Lu, F., Shi, B., Hu, Y., Li, M.: Big data and supply chain resilience: Role of decision-making technology. Management Decision 61(9), 2792–2796 (2023)
26. Ou, T., Chen, C., Tsai, W.: Establishing a dynamic recommendation system for e-commerce by integrating online reviews, product feature expansion, and deep learning. Applied Artificial Intelligence 39(1), 1–5 (2025)
27. Rui, G., Li, M.: Utilizing internet big data and machine learning for product demand forecasting and analysis of its economic benefits. Tehnicki Vjesnik - Technical Gazette 31(4), 1385–1388 (2024)
28. Zhu, Z., Sun, Y.: Personalized information push system for education management based on big data mode and collaborative filtering algorithm. Soft Computing 27(14), 10057–10060 (2023)

Mo Chen received Ph.D from School of Information, Renmin University of China in Computer Application Technology speciality, he is an associate professor of E-commerce

Department at Beijing Union University Business College. He engages in Data Structure, Database Principles and Applications, Data Acquisition and Preprocessing, Data Mining and Machine Learning, Business Big Data Analysis and Decision-Making, Big Data Technology and Application, Blockchain Application Technology and other courses teaching and researching work. His research interests are Big Data Analysis and Mining and so on, he publishes papers in the core journals, presides over research projects of the science and teaching.

Received: October 30, 2025; Accepted: March 20, 2026.