ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20241202002164
Original scientific paper

Research on Key Technology of Web Topic Detection Based on Tuple Semantic Description Analysis for Big Data

Mo CHEN

Abstract: In the context of the numerical intelligence age, how to further promote research on Web topic detection with heterogeneous big data as an important research object has attracted widespread attention from scholars. This paper takes network heterogeneous big data as the main research object and proposes a Web topic detection idea based on tuple semantic description analysis. Firstly, the main implementation methods are explained, including time item analysis, named entity item analysis, event item analysis, and semantic formal analysis. Secondly, the Web topic detection algorithm based on tuple semantic description analysis is elaborated. Through experiments, firstly, the impact of the quantity of Web news on the quality of time item description and five tuple semantic description is analysed, the influence of the adjustment parameters in the algorithm on the quality of the described Web news named entity items is also analysed to obtain the optimal adjustment range for these parameters, the impact of the long and short time selection range in the algorithm on the calculation results of the popularity weight of Web news named entity items is also analysed to obtain its optimal adjustment range, the impact of the adjustment parameter in the algorithm on the quality of the described Web news event items is also analysed to obtain the optimal adjustment range of the parameter. Secondly, the quality of topic detection under different semantic descriptions in Web news, the time consumed in the topic detection process under different methods, and the quality of topic detection under different datasets are analyzed. The experimental analysis process shows that the Web topic detection idea proposed in this paper is feasible, verifiable, and superior, and can play an important role in reconfiguring the Web topic corpus, inferring the Web hierarchical big data propagation path, and providing the numerical intelligence warehouses based on network information detection.

Keywords: big data; tuple semantic description; web topic detection

1 INTRODUCTION

Against the backdrop of the development of Web big data mining technology [1-3], scholars are still exploring and innovating in the face of the constantly emerging heterogeneous big data and diverse demands for numerical intelligence applications [4-6]. 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 [7, 8]. With the continuous development of Internet of Things, big data, cloud computing and other technologies [9-11], more and more devices and terminals are connected to the Internet, producing a large amount of data [12-14]. Therefore, scholars urgently need to think about how to detect more accurate knowledge from complex big data. In this exploration process, Web topic detection based on big data tuple semantic description analysis 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 [15-17]. 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 [18-21]. Based on the above characteristics, the existing statistical and deep learning methods are facing many challenges in Web topic detection, mainly due to the complexity, dynamics and heterogeneity for Web content, how to study the Web topic detection method based on tuple semantic description analysis, construct a Web topic corpus, infer Web hierarchical big data propagation path, and provide dynamic big data sources for network knowledge mining applications have become an urgent solution to the research problem.

To establish a strong foundation for the research of this paper, the next section will review existing methods related to Web topic detection, 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 Web topic detection, and elaborate on the main implementation methods and algorithms for tuple semantic description analysis.

This paper will make three key contributions: (1) it introduces a novel Web topic detection method using tuple semantic description analysis, (2) it optimizes named entity and event analysis for improved topic modelling, and (3) it provides extensive experimental validation and demonstrates the effectiveness of the proposed approach.

2 RELATED WORKS

Before detecting Web topic, 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. So, this paper intends to use Web news source material as the research object. The author of this paper has discussed the effective process of extracting Web news using incremental element extraction method through published paper [22], and this result of extraction is a foundation for continuing to research the detection technology for Web topic on tuple semantic description analysis for big data in this paper.

Based on having got Web news corpus, in order to accurately detect Web topic, it is important to carry out five tuple semantic description analyses, so that the computer can understand what Web news is about to report. In the 90s of last century, the researchers have begun to carry out semantic analysis for Web information. In this process, the technology of the semantic similarity computation is adopted. In recent years, the researchers are developing and making progress in the theory and technology of semantic description analysis, and some meaningful achievements have been made, shown in Tab. 1.

A solution of semantic analysis is proposed [23], in this solution, the aspect term is extracted, and the aspect

sentiment is predicted. Most solutions conduct the analysis task by handling the subtasks in a pipeline manner, whereby problems emerge in performance and real application. In this study, an end-to-end model is proposed, which fuses the syntactic structure information and the lexical semantic information, to address the limitation that existing solutions do not fully utilize textual information. This model integrates the part of speech, sememes, and context respectively by extracting syntactic structure information and lexical semantic information. Then, on the basis of an attention mechanism, this model further realizes the fusion of the syntactic structure information and the lexical semantic information to obtain analysis results for higher quality, in which way the text information is fully used. The subsequent experiments demonstrate that this model has certain advantages in using different text information, and the proposed solution outperforms all state-of-the-art related achievements. The testing also demonstrates the scalability of the solution, therefore, the entire research achievement has made a major contribution in the semantic analysis field.

Table 1 The related research status

The research direction	The research method	The research limitation	The proposed methodology
Web Topic Detection [23]	Propose an end-to-end model Extract the aspect term Predict the aspect sentiment	Ignore the logical association of information items between subtasks	The algorithm of time series construction
Web Topic Detection Semantic Analysis in Big Data [24]	Construct Chinese sentiment analysis model Integrate multi-granularity semantic features	Lack the logical description for multi-granularity semantic features	The algorithm of semantic description analysis
Semantic Analysis in Big Data Named Entity and Event Extraction [25]	Calculate semantic similarity Contrast the content similarity Follow the impact-factor	Ignore the setting of factor item weight	The algorithm of time series construction and semantic description analysis
Semantic Analysis in Big Data [26]	Design SciBERT model Propose a three-way framework	Lack the logical description based on the semantic relationships	The algorithm of semantic description analysis
Semantic Analysis in Big Data Named Entity and Event Extraction [27]	Construct a semantic recommendation model Introduce the formal concept analysis	Lack the logical description for multifactor items	The algorithm of time series construction and semantic description analysis

The solution for semantic analysis is presented [24], the Chinese sentiment analysis model integrating multi-granularity semantic features is constructed. The comparative experiments showed that the F1 values of this model reaches 88.28% and 84.80% on the man-made dataset and the NLPECC dataset, respectively. Most solutions have played a positive role in promoting the progress of semantic analysis research, but the characteristics of research objectives and downstream task requirements had not been thoroughly explored yet. This study introduces the radical and part-of-speech features based on the character and word features, with the application of bidirectional long short-term memory, attention mechanism and recurrent convolutional neural network. Meanwhile, an ablation experiment was conducted to verify the effectiveness of attention mechanism, part of speech, radical, character and word factors during the analysis process. So, the performance of the proposed model exceeds that of existing models to some extent. In view of the particularity of Chinese texts and the requirement of sentiment analysis, this paper borrows ideas from multiple interdisciplinary frontier theories and methods, such as information science, linguistics and artificial intelligence, which make it comprehensive and innovative. This paper deeply integrates multi-granularity semantic features such as character, word, radical and part of speech, which further complements the theoretical framework and method system of Chinese sentiment analysis.

A novel method of semantic analysis is proposed [25], in which both the content of the article and the authors profile are considered together to find the appropriate journal. In this study, one of the most important criteria to be considered is the content similarity of the journals and manuscript. For this purpose, the subject of the manuscript should be in accordance with the scope of the journal. Also, the manuscript content should be closed to the journals' trend for higher chance of acceptance. The second criterion is to take into account the impact-factor, acceptance-rate, review-time and publishing houses of the journal, which are suitable for the author's past publication profile. The experimental results conducted on real data sets have shown that the proposed method is applicable and performs high accuracy values based on hybrid semantic similarity and trend analysis, and the effectiveness of the proposed method is prior to previous methods with the development of big data technology. Therefore, the entire research achievement has made a major contribution in the field of semantic analysis and has strong application prospects in the semantic analysis of network data information.

A solution of semantic analysis is proposed [26], in this solution, the semantic main path network analysis approach based on citation context analysis is focused on research. Main path analysis is a method for extracting the scientific backbone from the citation network of a research domain, which has gradually attracted a great deal of attention from big data processors and scientists, but the existing approaches ignore the semantic relationships between the citing and cited objects resulting in several adverse issues, such as in terms of coherence of main paths and coverage of significant studies. This study advocates the semantic main path network analysis approach to alleviate these issues based on citation function analysis; meanwhile, a wide variety of SciBERT-based deep learning models are designed for identifying citation functions. The semantic citation networks are built by either including important citations, for example, extension, motivation, usage and similarity, or excluding incidental citations like background and future work. The semantic main path network is built by merging the top-K main paths extracted from various time slices of semantic citation network; in addition, a three-way framework is proposed for the quantitative evaluation of main path analysis results. Both qualitative and quantitative

analyses on three research areas of computational linguistics demonstrate that, compared to semantics-agnostic counterparts, different types of semantic main path networks provide complementary views of scientific knowledge flows. Combining them together, a more precise and comprehensive picture of domain evolution can be obtained, and more coherent development pathways can be uncovered between scientific ideas.

A solution of semantic analysis is proposed [27]. In this solution, a semantic recommendation model via fusing knowledge graph and formal concept analysis is constructed. The core idea of semantic recommendation is to incorporate semantic knowledge into the recommendation process. Most solutions have played a positive role in promoting the progress of semantic recommendation research. But the semantic recommendation algorithm, based on knowledge graph, ignores the deep implicit semantics of the evaluation data. The semantic recommendation algorithm based on the deep matrix decomposition model is limited to the implicit semantics of the evaluation data, the semantic recommendation algorithm based on the collaborative filtering algorithm performs only the selection of the nearest neighbors of the user or the item unilaterally, and ignores the influence of other aspects, which naturally leads to a decrease in the recommendation accuracy. To solve the above problems, this study introduces the formal concept analysis based on collaborative filtering. Using the property that the formal concept in the formal concept analysis can cluster objects of users and attributes of items simultaneously, this study proposes a semantic recommendation algorithm based on the knowledge graph and formal concept analysis to solve the problem of ignoring user or item factors. Finally, the proposed semantic recommendation algorithm is validated on two public datasets in this study. By using traditional algorithms and current semantic recommendation algorithms as benchmarks, extensive experiments show that the proposed semantic recommendation algorithm consistently outperforms state-of-the-art methods, therefore, the entire research process has made significant contribution in the field of semantic analysis.

Based on the above research status for Web information semantic analysis method, it can be summed up that most studies have taken the following processing steps including word segmentation, key characteristics extraction, similarity computation model construction, and semantic similarity computation and so on. However, the above process cannot fully consider how to accurately describe the topic for current internet news coverage published, what content can completely describe the topic it supports, how to standardize the various and complex expression mode of Chinese time, and how to analyse the tuple semantic information describing topic from time series. If above details can be studied in depth, the topic detection complexity can be reduced, the accuracy of topic detection can be improved, the research result can play an important role in 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, the next section will complete the problem definition firstly, and highlight the problem research boundaries.

3 PROBLEM DEFINITION IN WEB TOPIC DETECTION

From a global perspective, every Web news report can be viewed as 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. In this connection, the effective extraction and pre-processing process of Web news has been discussed in the paper published by the author of this paper [22], the Web news instances and their relationships have been deployed in tree structure, and noise information has also been filtered; it has provided a hierarchical and high-quality Web news source material for the research in this paper.

In text information related to Web news reports, it has presented unstructured characteristics, and it also contains more complex semantic characteristics composed of multiple information items. For example, there are core events reported by Web news and occurrence time of core events this time is different from Web news release time. There are occurrence locations of core events, initiating objects of core events, events associated with core events. The existence of these main information items is a requirement getting core content reported from Web news headline and large amounts of text for users. It is also key research point for analysing Web news contents. It interprets what core events were reported in Web news, when and where this core event occurs, what is the object of initiating this core event, what events are associated with this core event. However, these semantic characteristics behind Web news headline and large amounts of texts can refine core content reported by Web news, so it constitutes five tuples of Web news instances and can be used as a complete semantic description index for Web topic detection.

Given the problem definition for Web topic detection in Section 3, the next section will introduce the tuple semantic description analysis approach for Web topic detection, which aims to improve computational efficiency and accuracy.

4 PROPOSED METHODOLOGY: TUPLE SEMANTIC DESCRIPTION ANALYSIS 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 [28-30]. Based on above problem definition, this article proposes a five tuple semantic analysis method for Web topic detection, as shown in Fig. 1.

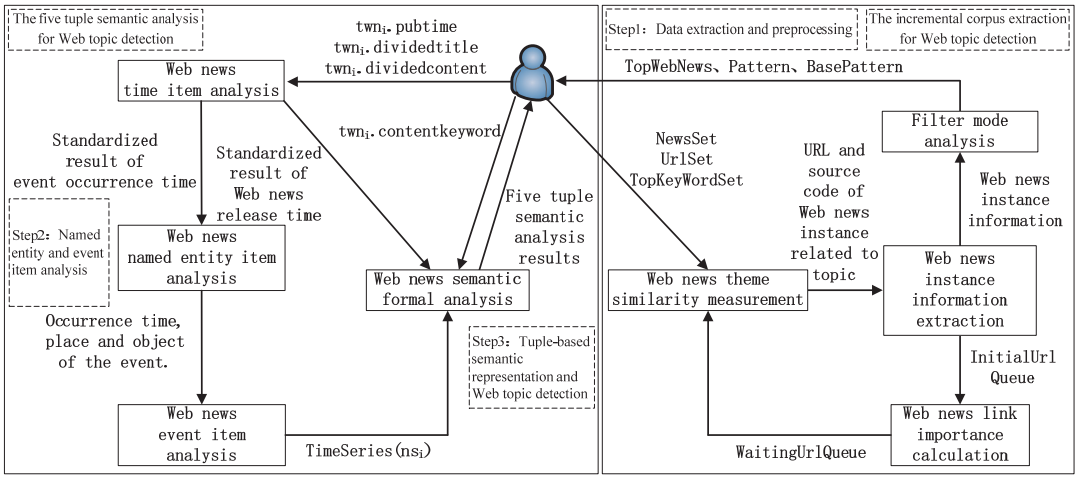

Figure 1 The framework of Web topic detection

The framework completes the incremental corpus extraction and the five tuple semantic analysis for Web topic detection, among them, this process of the incremental corpus extraction has been discussed in the paper published by the author of this paper [22], which provides the foundation of big data corpus for continuing to complete the five tuple semantic analysis in this paper. Through using data items of Web news release time and content segmentation results and so on, this framework can analyse Web news time items. Through using standardized results of occurrence time of events that have been analysed, this framework can analyse Web news naming entity items. Through using event occurrence time and naming entities that have been analysed, this framework can analyse Web news event item. Through using Web news content keywords and release time standardization results and Web news time point generation sequence, this framework can analyse five tuple semantic description results for Web news.

In this framework, the five defined data structures will be mainly applied, as shown in Tab. 2. News Set contains massive and authoritative URL in Web news network, the hls_{ij} represents the j hyperlink of ns_i in HyperLinkSet, UrlSet contains massive and authoritative instances in Web news network. If a Web news instance is published in $Page_i$ webpage, us_i can be presented with < url_i, $title_i$, $pubtime_i$, $pubsource_i$, $content_i$ >. url_i presents webpage address of $Page_i$, $title_i$ presents title of Web news, $pubtime_i$ presents release time of Web news, $pubsource_i$ presents release source of Web news, $content_i$ presents text of Web news. The key semantic characteristic information can be analysed from TopWebNews, the twn_i·url saves address for Web news, twn_i·title saves title for Web news, twn_i·pubtime saves release time for Web news, twn_i·pubsource saves release source for Web news, twn_i·content saves text for Web news, twn_i·dividedtitle saves segmentation result for Web news headline, twn_i·divided content saves segmentation result for Web news text, twn_i·content keyword saves keywords for Web news content, twn_i·relativity value saves similarity between Web news content and theme, twn_i·parenturl saves address of parent node for Web news instances, twn_i·pattern saves description for Web news instance filtering mode, twn_i·systemtime saves system time extracted for instances. FiveTuple describes fivetuple semantic information contained in Web news instances. The problem to be solved is as follows for Web topic detection method based on five tuple semantic description analysis, the ft_i = < Time, Place, Object, CoreEvent, RelationEvent >. ft_i·Time represents occurrence time of core event reported by Web news, ft_i·Place represents occurrence location of core event reported by Web news, ft.Object represents initiating object of core event reported by Web news, ft_i·CoreEvent represents core event information reported by Web news, ft_i·RelationEvent represents event information associated with core event reported by Web news, RelationEvent·preamble represents leading event information triggering core event, RelationEvent·derivative represents derivative event information generated by core event.

In a word, this paper can effectively analyse key information that represents its five tuple semantic characteristics from massive Web news instance data items using the framework proposed by this paper, in order to research a network topic detection method oriented semantic characteristics, and design the following algorithms.

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 based on Web news information extraction	$\{ns_1, ns_2, ns_3, \ldots, ns_{i-1}, ns_i, ns_{i+1}, \ldots, ns_n\}$
HyperLinkSet	The hyperlinks of massive Web news instances contained in NewsSet	$\{hls_{i1}, hls_{i2}, hls_{i3}, \ldots, hls_{i(j-1)}, hls_{ij}, hls_{i(j+1)}, \ldots, hls_{im}\}$
UrlSet	The big data source based on Web news information extraction	$\{us_1, us_2, us_3, \ldots, us_{i-1}, us_i, us_{i+1}, \ldots, us_n\}$
TopWebNews	The result of extracting massive Web news instances	$\{twn_1, twn_2, twn_3, \ldots, twn_{i-1}, twn_i, twn_{i+1}, \ldots, twn_k\}$
FiveTuple	The result of analysing semantic characteristic information	$\{ft_1, ft_2, ft_3, \ldots, ft_{i-1}, ft_i, ft_{i+1}, \ldots, ft_k\}$

4.1 The Algorithm of Time Series Construction for Web Topic Detection

The design idea of time series construction algorithm is as follows: According to results of Web news instance extraction, this algorithm can analyse the standardized time node from its content, and construct time series based on these nodes. The input content of this algorithm is Web news release time, content word segmentation results and other data items, the output content is standardized result of Web news release time and event occurrence time, and time series, the construction process is as follows.

This process can standardize Web news release time to YYYY-MM-DD HH:MM:SS format, it can be used as a reference for standardization of other time information. This process can standardize media coverage time in Web news. If this time exists, it will be standardized to YYYY-MM-DD HH:MM:SS format replacing Web news release time: it can be used as a reference for standardization of other time information. This process can standardize event occurrence time through extracting adjacent time words marked as /*t* or /*tg*, location words adjacent to time words marked as /*f*, quantifiers adjacent to time words marked as /*m* from word segmentation information of Web news title and text, the time phrases can be combined. According to the classification time shown in Tab. 3, different kinds of time information can be classified and conducted in time phrase.

Table 3 Time information classification table

Classification name	Classification description	Classification instance
C_1	Standard date	{2022 year}
C_2	Standard time	{8 hour}
C_3	Point to time	{last year, next month}
C_4	Time unit	{weekend, Friday}
C_5	Fuzzy time	{The third quarter, First ten days, afternoon, evening}
C_6	Specific time	{Spring Festival, National Day}

As shown in Eq. (1), it denotes n different atoms to be processed in time phrase from t_1 to t_n respectively, the value of n is greater than or equal to one.

$$T = \{t_1, t_2, ..., t_n\} \tag{1}$$

As shown in Eq. (2), C_j indicates that the time atom belongs to the category of processing, the value of j is between one and six, and the implementation process is shown in Fig. 2.

$$TC = \{(t_1, C_j), (t_2, C_j), ..., (t_n, C_j)\} \tag{2}$$

After processing the time points in the Web news instance, the time series can be formed as shown in Eq. (3). The $t_{i(j+1)}(ns_i \cdot \text{pubtime})$ represents the standardization result of Web news publishing time, $t_{i(j-1)}(ns_i \cdot \text{mediatime})$ represents the standardization result of media reporting time in Web news, $t_{ij}(ns_i \cdot \text{event}_j)$ represents other time point standardization results that appear in Web news.

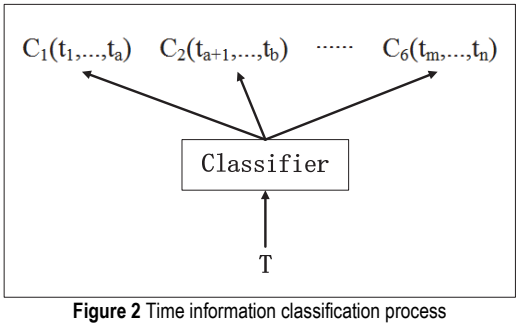

Figure 2 Time information classification process

$$\begin{aligned}&\text{TimeSeries}(ns_i) = \\ &(t_{i1}(ns_i \cdot \text{event}_1), t_{i2}(ns_i \cdot \text{event}_2), ..., \\ &t_{i(j-1)}(ns_i \cdot \text{mediatime}), t_{ij}(ns_i \cdot \text{event}_j), \\ &t_{i(j+1)}(ns_i \cdot \text{pubtime}), ..., t_{in}(ns_i \cdot \text{event}_n))\end{aligned} \tag{3}$$

Algorithm 1 TimeSeries_Construction
Input: NewsSet, Reg;
Output: EventTimeSet;
1: WebNews ← NewsSet;
2: MediaTime ← ø;
3: While each $w[i](0 \le i \le w\cdot\text{size}() - 1)$ Do
4: StandardTime for $w[i]\cdot$pubTime;
5: MediaTime ← Analyze mediatime from w[i].dividedcontent;
6: EventTimeSet ← Analyze eventtime using Reg;
7: Generate TimeSeries for $w[i]\cdot$pubTime $w[i]\cdot$mediatime and EventTimeSet;

4.2 The Algorithm of Semantic Description Analysis for Web Topic Detection

The design idea of semantic description analysis algorithm is as follows: according to the position of the event occurrence time in the sequence, this paper can analyze the place where the event occurred and the object of initiation, analyze the event information, and construct the sequence of events. According to this sequence, the core event nodes can be analyzed, the event nodes associated with the core events can be analyzed, and the analysis results can be described in a semantic formal way. The input content of this algorithm are content keywords, standardization results of publishing time, and the generation sequence of time points for Web news; the output content is the result of semantic description analysis for Web news. The construction process is described as follows.

This process can extract the occurrence place and initiating object of the event, in the existing location of the event occurrence time and within the search scope of all named entity, the words with named entity annotation features can be extracted. According to the named entity database, the heat of the extracted named entities can be sorted. Considering that the extracted named entities may come from the title or text of Web news, it is necessary to calculate the content weight of named entity item heat. Considering that the extracted named entities may come from different stages in the development of social events, the time weight of named entity item heat can be calculated. Combining these two weights, the balance weight of named entity item heat can be calculated as shown in Eq.

(4). $\text{Weight}_{\text{title}}$ represents the heat weight of named entity item in Web news title, $\text{Weight}_{\text{content}}$ represents the heat weight of named entity item in Web news text, $\text{Weight}_{\text{shorttime}}$ represents the heat weight of named entity item in a short time for the development of social events, $\text{Weight}_{\text{longtime}}$ represents the heat weight of named entity item in a long time for the development of social events, $\text{Weight}_{\text{entity}}$ represents the balance weight of the heat for named entity item. In the follow-up experiments, this paper will analyze the optimal value range for length of time and parameters. According to the sorting results, the primary and secondary named entity items can be analyzed, and the named entity database can be updated synchronously. As shown in Eq. (5), after extracting the occurrence place and initiating object of events for Web news, the semantics of Web news instance can be further described; $< t_j, p_j, o_j >$ (ns_i) represents the occurrence place p_j and the initiating object o_j of the event for Web news ns_i at t_j time point.

$$\begin{aligned}\text{Weight}_{\text{entity}} &= \gamma\left(\alpha\text{Weight}_{\text{title}} + \beta\text{Weight}_{\text{content}}\right) + \\ &+\delta\frac{\text{Weight}_{\text{shorttime}}}{\text{Weight}_{\text{longtime}}}\end{aligned} \tag{4}$$

$$\begin{aligned}&\text{TimeSeries}\left(ns_i\right) = \\ &\left(\left\langle t_1, p_1, o_1\right\rangle\left(ns_i\right), \left\langle t_2, p_2, o_2\right\rangle\left(ns_i\right), ..., \right. \\ &\left\langle t_{j-1}, p_{j-1}, o_{j-1}\right\rangle\left(ns_i \cdot \text{mediatime}\right), \left\langle t_j, p_j, o_j\right\rangle\left(ns_i\right), \\ &\left\langle t_{j+1}, p_{j+1}, ns_i \cdot \text{pubsource}\right\rangle\left(ns_i \cdot \text{pubtime}\right), ..., \\ &\left.\left\langle t_n, p_n, o_n\right\rangle\left(ns_i\right)\right)\end{aligned} \tag{5}$$

This process can extract the event information, in the event occurrence time and place, and the existing location of the object, the words with event annotation features can be extracted for search scope of all event items. According to the event corpus, the heat of the extracted event items can be sorted. Considering that the extracted event items will co-exist with the named entities within near range, the effective weight of the event items can be calculated. Considering that the extracted event items have also independent occurrence characteristics of high frequency under the effective condition, the heat weight of the event items can be calculated as shown in Eq. (6). TF_{event} represents the frequency of event items appearing in Web news, $\text{Max}(TF_{\text{event}})$ represents the highest frequency of event items appearing in Web news, TF_{leftword} represents the frequency of the named entity next to the left of the event item appearing in Web news, $TF_{\text{rightword}}$ represents the frequency of the named entity next to the right of the event item appearing in Web news, $\text{Weight}_{\text{event}}$ represents the heat weight for the event item. In subsequent experiments, the optimal range of parameter values will be analyzed in the formula. According to the sorting results, the primary and secondary event items can be analyzed, and the event corpus can be updated synchronously. As shown in Eq. (7), after extracting the event information of Web news, the semantics of Web news instance can be further described, $< t_j, p_j, o_j, e_j >$ (ns_i) represents occurance place p_j and initiating object o_j of event e_j for Web news ns_i at t_j time point.

$$\begin{aligned}\text{Weight}_{\text{event}} &= \alpha \times \frac{TF_{\text{event}}}{\text{Max}\left(TF_{\text{event}}\right)} + \\ &+\beta \times \frac{TF_{\text{event}}}{TF_{\text{leftword}} + TF_{\text{rightword}} - TF_{\text{event}}}\end{aligned} \tag{6}$$

$$\begin{aligned}&\text{TimeSeries}\left(ns_i\right) = \\ &\left(\left\langle t_1, p_1, o_1, e_1\right\rangle\left(ns_i\right), \left\langle t_2, p_2, o_2, e_2\right\rangle\left(ns_i\right), ..., \right. \\ &\left\langle t_{j-1}, p_{j-1}, o_{j-1}, \phi\right\rangle\left(ns_i \cdot \text{mediatime}\right), \\ &\left\langle t_j, p_j, o_j, e_j\right\rangle\left(ns_i\right), \left\langle t_{j+1}, p_{j+1}, ns_i \cdot \text{pubsource}, \phi\right\rangle \\ &\left(ns_i \cdot \text{pubtime}\right), ..., \left\langle t_n, p_n, o_n, e_n\right\rangle\left(ns_i\right)\left.\right)\end{aligned} \tag{7}$$

This process can determine the core events of Web news reporting, and the semantic description of core events in the sequence can be formed into a four tuple according to the heat weight of each event item and the keywords of Web news. The event items before the core event occurrence time point can be constituted as a leading event set, the event items after the core event occurrence time point can be constituted as a derivative event set. The set of leading and derived events, as well as four tuples, are recombined to form a five tuple semantics that can describe Web news, and generate Web news event sequence. As shown in Eq. (8), $< t_{ce}, p_{ce}, o_{ce}, e_{ce} >$ (ns_i) can describe core events, $< t_{pj}, p_{pj}, o_{pj}, e_{pj} >$ (ns_i) can describe leading events, $< t_{dk}, p_{dk}, o_{dk}, e_{dk} >$ (ns_i) can describe derivative events, $< t_{mt}, p_{mt}, o_{mt}, e_{ce} >$ $(ns_i \cdot \text{mediatime})$ can describe the core events reported by the media, $< t_{pt}, p_{pt}, ns_i.\text{pubsource}, e_{ce} >$ $(ns_i \cdot \text{pubtime})$ can describes the core events released by Web news. The result of this process can be shown in the webpage, as shown in Fig. 3.

$$\begin{aligned}&\text{TimeSeries}\left(ns_i\right) = \left(< t_{ce}, p_{ce}, o_{ce}, e_{ce}\left(ns_i\right), \right. \\ &\left\{< t_{p1}, p_{p1}, o_{p1}, e_{p1} >\left(ns_i\right), ..., \right. \\ &< t_{pj}, p_{pj}, o_{pj}, e_{pj} >\left(ns_i\right), ..., \\ &\left.< t_{pn}, p_{pn}, o_{pn}, e_{pn} >\left(ns_i\right)\right\}, \\ &\left\{< t_{d1}, p_{d1}, o_{d1}, e_{d1} >\left(ns_i\right), ..., \right. \\ &< t_{dk}, p_{dk}, o_{dk}, e_{dk} >\left(ns_i\right), ..., \\ &\left.< t_{dm}, p_{dm}, o_{dm}, e_{dm} >\left(ns_i\right)\right\}, \\ &\left\{< t_{mt}, p_{mt}, o_{mt}, e_{ce} >\left(ns_i \cdot \text{mediatime}\right), \right. \\ &\left.\left.< t_{pt}, p_{pt}, ns_i \cdot \text{pubsource}, e_{ce} >\left(ns_i \cdot \text{pubtime}\right)\right\}\right)\end{aligned} \tag{8}$$

Algorithm 2 Tuple_SemanticDescription
Input: NewsSet, EventTimeSet, Parameters, Reg;
Output: FiveTuple;
1: WebNews ← NewsSet;
2: Entity ← ø, Event ← ø, CoreEvent ← ø;
3: While each $w[i](0 \le i \le w\cdot\text{size}() - 1)$ Do
4: Entity ← Analyze entity using time position and Reg;
5: Event ← Analyze event using Reg and position of time and entity;
6: CoreEvent ← Analyze coreevent using event and $w[i]\cdot$contentkeyword;
7: Construct FiveTuple;

Web News Multi-Feature Introduction

The German A320 aircraft crash | Submit | Reset | Filter

Web News Instance								
Web News Title	Web News Five Tuple					Web News Operation		
	Time	Place	Object	CoreEvent	RelationEvent			
德国举行德翼空难国家纪念式500遇难者家属出席	2015-04-17 15:30:00	德国科隆大教堂	(德国)(遇难者)(家属)	(举行)(空难)(纪念)	(2015-03-24)(法国南部)(德国之翼公司空客A320飞机)(坠毁)(2015-04-15)(汉莎航空公司)(汉莎航空公司)(推迟、庆祝、重生周年)(2015-04-17 15:30:00)(德国科隆大教堂)(德国、遇难者、家属)(举行、空难纪念)(2015-04-18 08:16:00)(北京)(中国广播网)(发布)(2015-04-29)(调查部门)(调查部门)(浮现、坠机细节)	Browse	Delete	Update
汉莎无义务向当局通报德翼副驾驶曾患抑郁症	2015-04-07	德国	汉莎航空公司	(无须)(通报)(患抑郁症)	(2009-01-01)(德国)(卢比茨)(中断、飞行资格)(2012-01-01)(德国)(卢比茨)(取得、飞行资格)(2013-12-31)(德国)(航空当局)(实施、通报规定)(2015-03-27)(德国)(发言人)(知道、有精神问题)(2015-04-03)(法国)(民航事故技术调查分析局)(解读、黑匣子)(2015-04-05)(德国)(联邦飞航局)(不知、患抑郁症)(2015-04-07)(德国)(汉莎航空公司)(无须、通报、患抑郁症)(2015-04-07 08:24:00)(北京)(中国新闻网)(发布)	Browse	Delete	Update
德翼坠机罹难者遗体搜寻工作结束 遗体辨认持续	2015-04-05	法国	当局	(搜寻遗体)(结束)(分析DNA)(继续)	(2009-01-01)(德国)(德国汉莎航空)(卢比茨曾患抑郁症)(2014-11-01)(德国)(联邦航空局)(接受、改正缺失)(2015-03-24)(法国、阿尔卑斯山)(德翼航空4U9525航班)(坠毁、遇难)(2015-03-29)(法国)(搜救人员)(收集残骸、寻找黑匣子)(2015-03-30)(法国)(蓬图瓦兹、国家宪兵犯罪研究所法医)(收集、坠毁遇难者DNA)(2015-04-05)(法国)(当局)(搜寻遗体、结束、分析DNA、继续)(2015-04-05 08:44:00)(北京)(中国新闻网)(发布)(2015-04-19)(法国)(当局)(完成、辨认遗体)	Browse	Delete	Update
Next Page Last Page								

Figure 3 The effect of web news five tuple semantic description analysis

Based on the method proposed in this paper, the effect of five tuple semantic description analysis can be obtained for Web news. It can represent Web topic detected and hidden in the semantic characteristic of massive Web news in the context of social events, and then it can infer Web hierarchical big data propagation path. In the subsequent research, the feature evaluation and use of behavior tracking on the analyzed big data corpus can be conducted to further enhance the research value of Web topic detection. To reflect the feasibility, verifiability and superiority of 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.

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 2.40 GHz, the memory is 64 GB, 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 preprocessing Web big data [31, 32].

This paper uses the massive Web news generated by the German A320 aircraft crash event as the experimental network big data corpus. 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 analyzed Web news instances to all analyzed Web news instances, but also the five tuple semantics of correctly described Web news instances to the five tuple semantics of all described Web news instances. It is also used to measure the ratio of correctly detected topics to all detected topics; this ratio reflects the accuracy of analysis and detection.

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, 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 impact of the number of Web news on the described time items and the five tuples semantic quality; it can reflect the effect of algorithm implementation for time series construction and tuple semantic description as shown in Fig. 4.

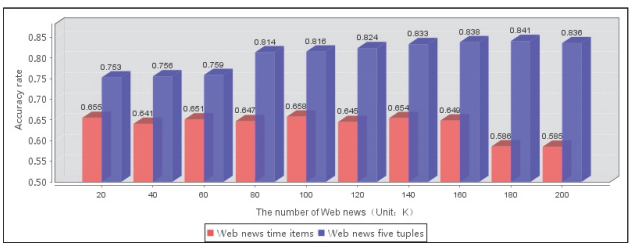

Figure 4 The accuracy change trend of the time items and five tuples description for Web news with their number

The accuracy indicates whether the described Web news time items and five tuples semantics belong to the accurate or approximately accurate annotation category. The red bar indicates the accuracy change situation of the described Web news time items. It can be seen from its trend that in the case of a small amount of Web news, the accuracy rate changes more smoothly in fluctuating, around 65%. However, with the continuous increase in the quantity for Web news, the accuracy has a significant downward trend. The reason is that for social events that occur in a certain range, the relative or implicit time used in the subsequent released Web news is significantly increased, which increases the standardization difficulty of Web news time items. The blue bar represents the accuracy changed situation of the described Web news five tuples semantics, from its trend. It can be seen that as the number of Web news increases, the semantics of the named entity and event items described become richer, which makes the accuracy rate have a stable trend of improvement up to 84.1%. In general, the description of Web news named entity items and event items is based on the description of time items. Therefore, the blue column is significantly higher than the red column, which is in line with the expected experimental effect. This experiment shows that the five tuples semantic description index analysis for Web news has higher completeness and accuracy.

Next, this section analyzes the influence of the adjustment parameters in the algorithm on the quality of the described Web news named entity items, in order to obtain the optimal adjustment range of these parameters for Eq. (4), as shown in Fig. 5.

The accuracy represents the quality of Web news named entity items described when the parameters in the semantic description and analysis algorithm are adjusted. The red dotted line indicates the changed situation of accuracy when the Alpha parameter in the algorithm takes different values. From its trend, it can be seen that when the Alpha value is adjusted to between 1.2 and 1.35, the quality of the described Web news named entity items is higher and stable with the accuracy of about 75%. The blue dotted line indicates the change of situation of accuracy when the Beta parameter in the algorithm takes different values. From its trend, it can be seen that when the Beta value is adjusted to between 0.75 and 0.9, the quality of the described Web news named entity items is higher and stable, with the accuracy of about 65%. The green dotted line indicates the changed situation of accuracy when the Gamma parameter in the algorithm takes different values. From its trend, it can be seen that when the Gamma value is adjusted to between 0.65 and 0.7, and the Sigma value is adjusted to between 0.3 and 0.35, the quality of the described Web news named entity items is higher and stable with the accuracy of about 80%. In general, the adjustment of each parameter can make the quality of the described Web news named entity items locally stable to the maximum value, which is in line with the expected experimental effect, and the optimal adjustment range for each parameter is determined.

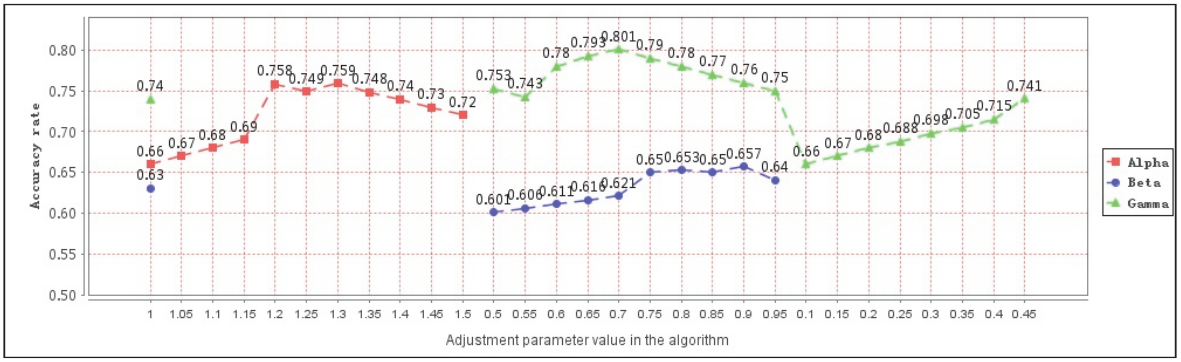

Figure 5 The change trend of accuracy with the adjustment of various parameters in the algorithm

Next, this section analyzes the influence of the long and short time selection range in the algorithm on the calculation result of the heat weight for Web news named entity items, in order to obtain its optimal adjustment range for Eq. (4), as shown in Fig. 6.

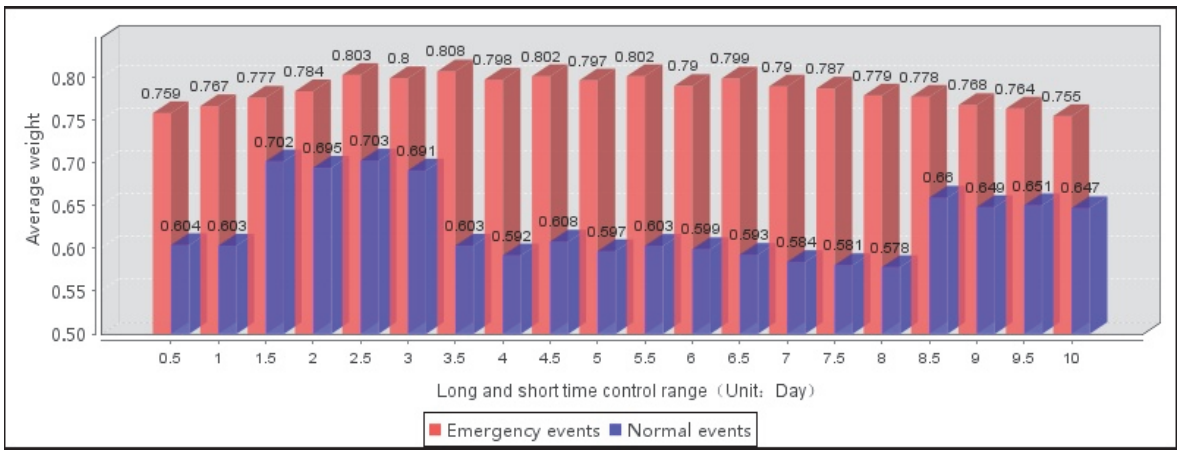

Figure 6 The change trend of average weight with the adjustment of long-term and short-term control range

The red column indicates the change situation of average weight according to the selected range of long and short time in the case of emergencies for society. It can be seen from its trend that when 2 to 3 days are selected in a short time, the average weight has increased significantly and is stabilized at about 0.8. When choosing 9 to 10 days for a long time, the average weight decreases significantly and stabilizes at about 0.76. The blue column indicates the change situation of the average weight according to the range selected in the long and short time in the case of normal events for society, from its trend, when 1.5 to 3 days are selected in a short time, the average weight is higher and stable at about 0.7. When 7 to 8 days are selected for a long time, the average weight is lower and stable at about 0.58. This experiment shows that when determining different long-term and short-term selection ranges for social events, it will affect the time weight calculation result for Web news named entity items heat.

Next, this section analyzes the impact of the adjustment parameter in the algorithm on the quality of the described Web news event items to obtain the optimal adjustment range of the parameter for Eq. (6), as shown in Fig. 7.

When adjusting the parameter in the semantic description and analysis algorithm, accuracy represents the quality of the described Web news event items. The red dashed line represents the change in accuracy when the Beta parameter in the algorithm takes different values. From its trend, it can be seen that when the Beta value is adjusted between 0.55 and 0.65, and the Alpha value is adjusted between 0.35 and 0.45, the quality of the described Web news event items is higher, stabilizing at around 70%. Overall, adjusting parameter can stabilize the quality of the described Web news event items locally to the maximum value, which is in line with the expected experimental effect and determines the optimal adjustment range of parameter.

Next, this section analyzes the quality of topic detection under a different semantic description for Web news, as shown in Fig. 8 and Tab. 4.

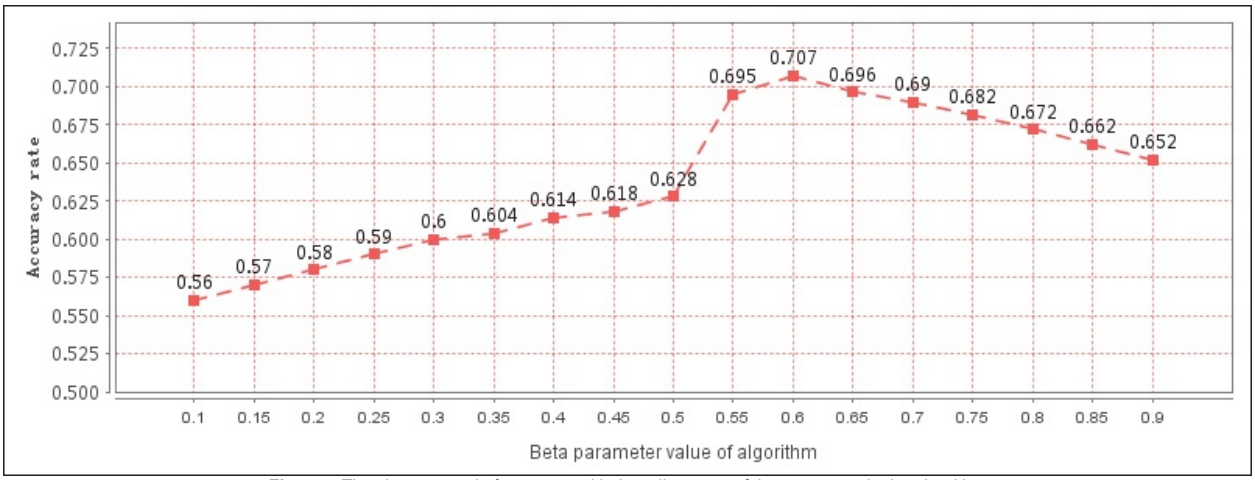

Figure 7 The change trend of accuracy with the adjustment of the parameter in the algorithm

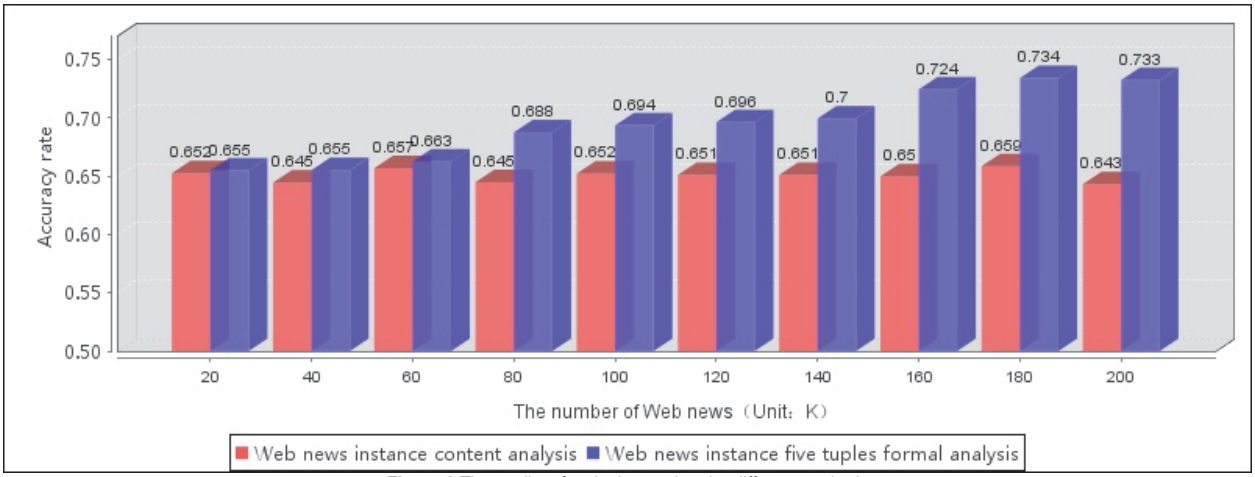

Figure 8 The quality of topic detected under different methods

Table 4 The comparison of accuracy rate for the topic detection quality under different methods

The number of Web news (Unit: K) / The research method	20	40	60	80	100	120	140	160	180	200
The method of Web news instance content analysis	0.652	0.645	0.657	0.645	0.652	0.651	0.651	0.65	0.659	0.643
The method of Web news instance five tuples formal analysis	0.655	0.655	0.663	0.688	0.694	0.696	0.7	0.724	0.734	0.733

The accuracy represents the quality of topic detected through two methods. The red bar represents the accuracy changes of topic detected under semantic analysis of Web news instance content. From its trend, it can be seen that as the number of Web news instances increases, comparable Web news instance content becomes richer, and the quality of topic detected improves to a certain extent, but its accuracy can only fluctuate around 65%. The blue bar represents the accuracy changes of the topics detected under the five tuple semantic formal analysis of Web news instance. From its trend, it can be seen that although the number of Web news is small, the accuracy is not much different from the semantic analysis of Web news instance content; however, as the number of Web news increases, the comparable five tuple semantic content of Web news instance becomes more accurate, and the accuracy rate is also increasing, reaching up to 73.4%. This experiment shows that under the five tuple semantic formal analysis of Web news instance, the quality of topic detected is higher than that of content semantic analysis in Web news instance.

Next, this section analyzes the time consumed in topic detection process under different methods, as shown in Fig. 9. This experiment compares the execution time consumed to solve the problem proposed in this paper under the method of combining theme matching and semantic analysis in Web news instance content, as well as under the method proposed in this paper. For the German A320 plane crash incident, the *x*-axis represents the time of massive amounts of Web news released by the authoritative Web news network. The red solid line represents the execution time consumed under the combination method of Web news instance content theme matching and semantic analysis, while the blue solid line represents the execution time consumed under the method proposed in this paper. From the trend of the two-color solid line, it can be seen that as this event progressed, the number of Web news increased sharply, released in March, April, and some parts of May 2015 respectively. Therefore, whether under the combination method of theme matching and semantic analysis in Web news instance content, or under the method proposed in this paper, the execution time required to solve the problem proposed in this paper is significantly higher than the relative stationary period of the solid line. However, under the method proposed in this paper, the execution time consumed is significantly less than the combination method of Web news instance content theme matching and semantic analysis. This experiment demonstrates that using the method proposed in this paper, it can effectively solve the uncertainty problem presented in this paper. Finally, this section analyzes the quality of topic detection on different datasets, as shown in Fig. 10 and Tab. 5.

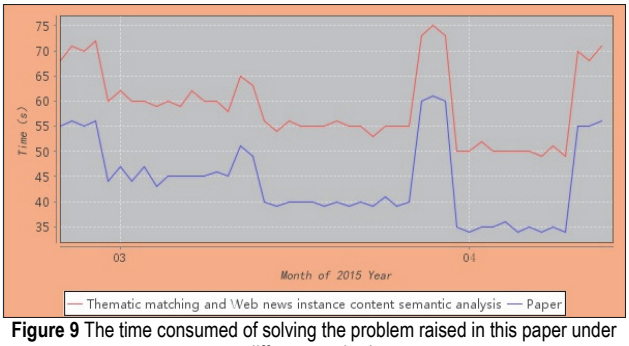

Figure 9 The time consumed of solving the problem raised in this paper under different methods

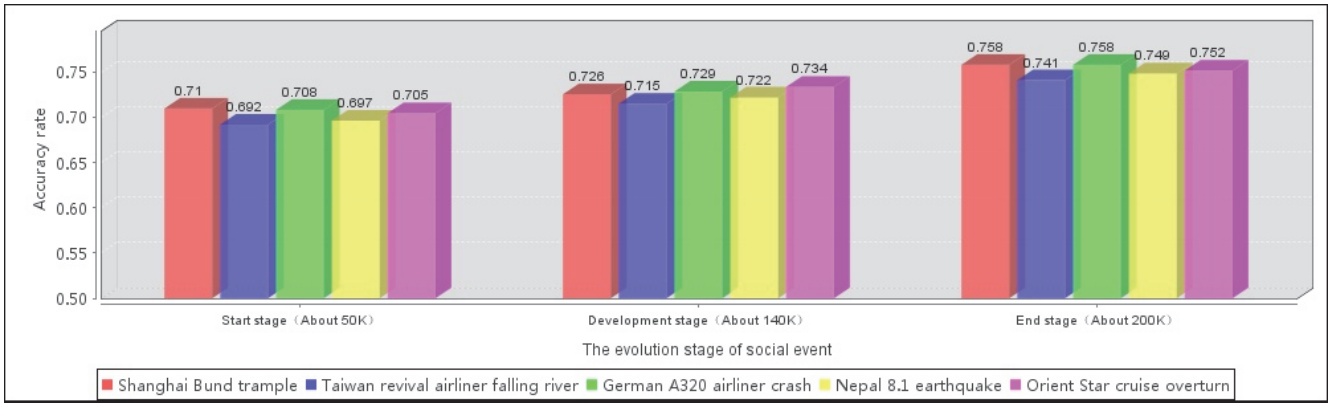

Figure 10 The quality of topic detection on different data sets

Table 5 The comparison of accuracy rate for the topic detection quality on different data sets

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 50 K）	0.71	0.692	0.708	0.697	0.705
Development stage（About 140 K）	0.726	0.715	0.729	0.722	0.734
End stage（About 200 K）	0.758	0.741	0.758	0.749	0.752

In addition to analyzing the topic detection quality of the German A320 aircraft crash data set, topic detection quality analysis is also conducted on the standard data sets of the other four events. 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 beginning, development, and end stages. This indicates that the quality of topic detection using the method proposed in this paper is relatively stable under different events; however, as the number of Web news 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, the semantic description of the five tuple under the theme is constantly expanding, and the support for topic detection is also improving.

6 CONCLUSION

The research on the Web topic detection method based on the five tuple semantic description analysis technology, taking the big data of Web news network 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 method research, this paper proposed the Web news time item analysis module, named entity item analysis module, event item analysis module, and semantic formal analysis module, and proposed time series construction and semantic description analysis algorithm 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 the Web topic corpus, improving the understanding of network big data, inferring the propagation path of Web hierarchical big data, and providing an intelligent big data warehouse for network information detection application.

In subsequent research, the optimization of the Web topic detection algorithm 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 Web topic detection. The research results can be applied to real-time Web news monitoring and multilingual Web topic detection processes, and can further improve the computational efficiency of big data applications.

Acknowledgements

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.

7 REFERENCES

[1] [1] Li, P. & Zhang, L. M. (2025). Application of Big Data Technology in Enterprise Information Security Management. *Scientific Reports*, *15*(1), 1-5. https://doi.org/10.1038/s41598-025-85403-6

[2] Khairy, D., Alharbi, N., Amasha, M. A., Areed, M. F., Alkhalaf, S., & Abougalala, R. A.(2024). Prediction of Student Exam Performance Using Data Mining Classification Algorithms. *Education and Information Technologies*, *29*, 21621-21626. https://doi.org/10.1007/s10639-024-12619-w

[3] Kim, M. & Yoo, H. (2024). Identification of Key Service Features for Evaluating the Quality of Metaverse Services: A Text Mining Approach. *IEEE Access*, *12*, 6719-6723. https://doi.org/10.1109/ACCESS.2024.3352008

[4] Arunkumar, M., Rajkumar, K., Jeyaseelan, W. R. S., & Natraj, N. A. (2025). Data Mining, Machine Learning, and Statistical Modeling for Predictive Analytics with Behavioral Big Data. *TehnickiVjesnik-Technical Gazette*, *32*(1), 72-74. https://doi.org/10.17559/TV-20231102001073

[5] Song, S. H., Pan, L., & Liu, S. J. (2024). A Q-Learning Based Auto-Scaling Approach for Provisioning Big Data Analysis Services in Cloud Environments. *Future Generation Computer Systems*, *154*, 140-144. https://doi.org/10.1016/j.future.2024.01.003

[6] Soldatenko, S. A. (2024). Artificial Intelligence and Its Application in Numerical Weather Prediction. *Russian Meteorology and Hydrology*, *49*(4), 283-287. https://doi.org/10.3103/S1068373924040010

[7] Wang, X. X., Zhang, H. P., Li, Z. T., Yang, X., Qian, Z. Y.,Bian, D., & Zhao, Y. X.(2025). Revealing the City Influence and Its Pattern Using Web Search Data: A New Perspective Through Attention Flow. *ISPRS International Journal of Geo-Information*, *14*(1), 1-3. https://doi.org/10.3390/ijgi14010024

[8] Xu, Y. Q., Kang, C. G., Jiao, W., & Jia, Y. H. (2024). Discovering Structure and Influencing Factors of Chinese City Directed Network (CCDN) from Web Search Engine Data. *Applied Geography*, *49*(4), 283-287. https://doi.org/10.1016/j.apgeog.2025.103564

[9] Liu, B. T. & Tang, M. (2025). A Very Compact and a Threshold Implementation of uBlock for Internet of Things. *Tsinghua Science and Technology*, *30*(5), 2270-2273. https://doi.org/10.26599/TST.2024.9010257

[10] Tang, J. W., Yan, Y. T., Bao, J., & Huang, B. (2025). Big Data-Driven Control of Nonlinear Processes Through Dynamic Latent Variables Using an Autoencoder. *IEEE Transactions on Cybernetics*, *55*(5), 2411-2415. https://doi.org/10.1109/TCYB.2025.3544257

[11] Andreoli, R., Mini, R., Skarin, P., Gustafsson, H., Harmatos, J., Abeni, L., & Cucinotta, T. (2025). A Multi-Domain Survey on Time-Criticality in Cloud Computing. *IEEE Transactions on Services Computing*, *18*(2), 1152-1155. https://doi.org/10.1109/TSC.2025.3539197

[12] Wang, S. B. (2024). 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. https://doi.org/10.14733/cadaps.2024.S4.39-55

[13] Lazzeretti, L., Domenech, R. B., Hervas-Oliver, J. L., & Innocenti, N. (2023). Artificial Intelligence, Big Data, Algorithms and Industry 4.0 in Firms and Clusters. *European Planning Studies*, *31*(7), 1297-1302. https://doi.org/10.1080/09654313.2023.2220490

[14] Jensen, M. H., Nielsen, P. A., & Persson, J. S. (2023). From Big Data Technologies to Big Data Benefits. *Computer*, *56*(6), 52-56. https://doi.org/10.1109/MC.2022.3206032

[15] Wang, H. D. & Zhang, S. Y. (2025). Research on the Application of Improved BERT-DPCNN Model in Chinese News Text Classification. *Concurrency and Computation: Practice and Experience*, *37*(3), 1-3. https://doi.org/10.1002/cpe.8338

[16] Mallick, P. K., Mishra, S., & Chae, G. S. (2023). Digital Media News Categorization Using Bernoulli Document Model for Web Content Convergence. *Personal and Ubiquitous Computing*, *27*(3), 1087-1091. https://doi.org/10.1007/s00779-020-01461-9

[17] Jadhav, D. & Singh, J. (2025). Web Information Extraction and Fake News Detection in Twitter Using Optimized Hybrid Bi-Gated Deep Learning Network. *Multimedia Tools and Applications*, *84*(11), 9471-9475. https://doi.org/10.1007/s11042-024-19225-5

[18] Sharma, S., Aujla, G. S., & Bali, R. S. (2024). Big Data Techniques Utilization in Intelligent Transportation System Environment. *Communications in Computer and Information Science*, *1930*, 368-373. https://doi.org/10.1007/978-3-031-48781-1_29

[19] Aseman-Manzar, M. M., Karimian-Aliabadi, S., Entezari-Maleki, R., Egger, B., & Movaghar, A. (2023). Cost-Aware Resource Recommendation for DAG-Based

Big Data Workflows: An Apache Spark Case Study. *IEEE Transactions on Services Computing*, *16*(3), 1726-1729. https://doi.org/10.1109/TSC.2022.3203010
[20] Shi, H. G. & Yang, Q. (2023). Linguistic Multidimensional Perspective Data Simulation Based on Speech Recognition Technology and Big Data. *Soft Computing*, *27*(14), 9967-9970. https://doi.org/10.1007/s00500-023-08191-z
[21] Malysiak-Mrozek, B., Wieszok, J., Pedrycz, W., Ding, W. P., & Mrozek, D. (2022). High-Efficient Fuzzy Querying With HiveQL for Big Data Warehousing. I*EEE Transactions on Fuzzy Systems*, *30*(6), 1823-1826. https://doi.org/10.1109/TFUZZ.2021.3069332
[22] Chen, M. & Yang, X. P.(2016). Research on Model of Network Information Extraction Based on Improved Topic-Focused Web Crawler Key Technology. *Tehničkivjesnik/Technical Gazette*, *23*(4), 1025-1035. https://doi.org/10.17559/TV-20150314134638
[23] Bie, Y., Yang, Y., & Zhang, Y. L. (2023). Fusing Syntactic Structure Information and Lexical Semantic Information for End-to-End Aspect-Based Sentiment Analysis. *TSINGHUA Science and Technology*, *28*(2), 230-243. https://doi.org/10.26599/TST.2021.9010095
[24] Liu, Z. B. & Zhao, W. J. (2023). Chinese Sentiment Analysis Model by Integrating Multi-granularity Semantic Features. *Data Technologies and Applications*, *57*(4), 605-622. https://doi.org/10.1108/DTA-10-2022-0385
[25] Yasar, M. Y. & Kaya, M. (2023). Author-Profile-Based Journal Recommendation for a Candidate Article: Using Hybrid Semantic Similarity and Trend Analysis. *IEEE Access*, *11*, 45826-45837. https://doi.org/10.1109/ACCESS.2023.3271732
[26] Jiang, X. R. & Liu, J. J. (2023). Extracting the Evolutionary Backbone of Scientific Domains: The Semantic Main Path Network Analysis Approach Based on Citation Context Analysis. *Journal of the Association for Information Science and Technology*, *74*(5), 546-569. https://doi.org/10.1002/asi.24748
[27] Wei, L. N. & Zhu, X. T. (2023). Semantic Recommendation Model via Fusing Knowledge Graph and Formal Concept Analysis. *IEEE Access*, *11*, 62337-62347. https://doi.org/10.1109/ACCESS.2023.3287778
[28] Angskun, T., Sritha, K. N., Srithong, A., Khopolklang, N., Kamollimsakul, S., Phithak, T., & Angskun, J. (2024). Using Big Data to Assess an Affective Domain for Distance Education. *Future Generation Computer Systems*, *160*, 131-139. https://doi.org/10.1016/j.future.2024.05.057
[29] Choi, U. & Lee, K. (2023). Dense or Sparse: Elastic SPMM Implementation for Optimal Big-Data Processing. *IEEE Transactions on Big Data*, *9*(2), 637-641. https://doi.org/10.1109/TBDATA.2022.3199197
[30] Silva, D. H., Maziero, E. G., Saadi, M., Rosa, R. L., Silva, J. C., Rodriguez, D. Z., & Igorevich, K. K. (2022). Big Data Analytics for Critical Information Classification in Online Social Networks Using Classifier Chains. *Peer-to-Peer Networking and Applications*, *15*(1), 626-632. https://doi.org/10.1007/s12083-021-01269-1
[31] Rui, G. W. & Li, M. G. (2024). Utilizing Internet Big Data and Machine Learning for Product Demand Forecasting and Analysis of Its Economic Benefits. *Tehnicki Vjesnik-Technical Gazette*, *31*(4), 1385-1394. https://doi.org/10.17559/TV-20240318001408
[32] Zhu, Z. F. & Sun, Y. L. (2023). Personalized Information Push System for Education Management Based on Big Data Mode and Collaborative Filtering Algorithm. *Soft Computing*, *27*(14), 10057-10064. https://doi.org/10.1007/s00500-023-08213-w

Contact information:

Mo CHEN, Ph.D, Associate Professor
Big Data Management and Application Major,
School of Business, Beijing Union University,
No. 3, Yanjing Dongli, Chaoyang District, Beijing
E-mail: mo.chen@buu.edu.cn; 962387563@qq.com