Web News Media Retrieval Analysis Integrating with Knowledge Recognition of Semantic Grouping Vector Space Model

Wenting Xiong
School of Journalism and communication, Hunan Mass Media Vocational and Technical College,
Changsha, Hunan, 410100, China
E-mail: 2016150180@jou.edu.cn

Keywords: vector space model, semantic grouping, searching for information, web news media retrieval

Received: October 26, 2023

Traditional Web news media retrieval technology can only meet the specific requirements of customers. Because of its universal characteristics, it cannot meet the needs of different environments, different purposes, and different times simultaneously. Researchers have proposed a search method for online news media, which is used for computing the semantic grouping vector space model. The customer's interest model is analyzed through the characteristics of the user's different classification areas. In this paper, we propose a vector space model that performs semantic grouping based on feature words. The model divides four groups that are relatively independent in the meaning of feature words in a news report: time, place, person, and event, and then forms four vector spaces and calculates the weight value and similarity of each vector space. Theoretical analysis and experimental results show that the improved model is suitable for searching Web news information and improves the calibration rate, query speed, and calibration rate.

Povzetek: V študiji je predstavljen izboljšan model vektorskega prostora za iskanje spletnih novic, ki vključuje semantično grupiranje po ključnih besedah v nekaj kategorij.

1 Introduction

Early search engines generally used content-based search methods, which were developed on the basis of the theory and technology of traditional information retrieval. The primary consideration is the relationship between web pages and search terms and the frequency and location of users querying documents. This method improves the search quality and accuracy to a certain extent, but these methods are based on keyword queries. Synonyms and polysemous words in natural language cannot be retrieved, so the search rate is not high and has certain limitations[1-2]. In recent years, the rapid progress of information technology has ushered in the digital age, and the digital and electronic files of past paper files are gradually being replaced. How to quickly and accurately query the user's demand information in the Web news media database? The current significant problem is how to deal with the issue of archive information supply and demand. Web news media retrieval is a helpful solution. By using a personalized search system to refer to the same information among users, new calculation methods for the semantic grouping vector space model can be discovered.

The most important thing is the user type, but the participation of users is required because of the calculation method and collaboration Personalized search methods of semantic grouping vector space model have their characteristics.

With the continuous development of the multimedia entertainment field, Web news media has gradually integrated into people's lives. However, how large-capacity Web news media data can better adapt to the environment and different user characteristics and introduce Web news media suitable for users is a challenge facing people. This paper presents the knowledge recognition of the semantic grouping vector space model in the Web news media retrieval process. It accurately conducts the Web news media retrieval by using the deep belief network algorithm in the audio segmentation stage of the Web news media. Finally, the results of the experimental analysis show that the method proposed in this paper can quickly retrieve the corresponding type of Web news media according to the user's preferences.

References	Key findings	Methodology	Limitation
10	The study focused on the Hermes Framework for Personalized News Services, with an emphasis on implementing Semantic Web standards and utilizing topic data.	The study included approaches such as Ontology-based Knowledge Representation, Natural Language Processing (NLP) for Semantic Text Analysis, and Semantic Query Languages for Information Specification.	The challenges included complexity, difficulty in acquiring knowledge, reliance on Semantic Web standards, and scalability concerns.
11	The study examined the effective extraction of events from a news corpus, emphasizing the strength of Latent Dirichlet Allocation (LDA) in collecting semantic connections. It also evaluated the efficacy of Continuous Bag of Words (CBOW) and Skip-gram models in obtaining semantic similarity.	The process of preparing the Online news corpus involved the utilization of diverse research approaches.	The study addressed limitations within the news corpus, generalization difficulties, and reliance on algorithmic parameters.
12	The research focused on gathering data using web crawling and extracting features using NLP techniques.	The study comprised primary stages: Online News Corpus Preparation, NLP for Feature Extraction, and the Selection of Machine Learning Models.	The study examined many factors pertaining to generalization, the ever-changing nature of fake news, and ethical implications.
13	The study discussed the challenges of collecting high-level information, proposed a compact event representation, and introduced a hypothesis for the facilitation of social and geopolitical analysis in the context of the rise of the Social Web.	The study emphasized the use of data mining tools to examine event representations, integrating case studies and user evaluation procedures.	Factors including data velocity and volume, the study's dependence on data quality, and the extent of user examination were among the constraints.
14	The study focused on the deployment of the Hermes Framework for Personalized News in the Hermes News Portal.	Three main approaches were examined in the study: the implementation of the Hermes News Portal, semantic query languages, and semantic text analysis.	The complex nature of complexity and the learning curve, as well as its dependence on the quality of ontology.
15	The study concentrated on the challenges associated with traditional search engines and the integration of NLP.	The study involved using Web Crawling and NLP techniques to gather and preprocess data, followed by applying Sentiment Analysis algorithms.	The study factored in reliance on data sources, sentiment analysis limitations, and resolving plurality.
16	The paper focused on the Semantic Model that employed Term Frequency-Inverse Document Frequency (TF-IDF) and evaluated its performance using evaluation metrics.	The research focuses on three key aspects: data preprocessing, the utilization of machine learning and deep learning algorithms, and the creation of a robust prediction model.	The study presented challenges in generalizing findings, allocating resources, and accounting for temporal variables.
17	The study focused on analyzing the level of financial area analysis and the importance of material on news websites.	The study utilized various methods such as data collection, the implementation of a supervised machine	The Study complexities of the Lithuanian language presented research restrictions due to its algorithmic sensitivity.

		learning model, and hyper-parameter optimization using Grid Search.	
18	The study focused on the various challenges faced by conventional search engines, the integration of sentiment analysis technologies, and the subsequent deployment of intelligent search capabilities.	The research involved various methodologies, such as data collection, web crawling, Natural Language Processing (NLP) for text preprocessing, and the implementation of sentiment analysis algorithms.	The research involved depending on the data source and carefully considering temporal aspects.
19	The study's primary findings included an analysis of annual publication trends and subject distribution, highly cited literature and research hotspots, and an evaluation of module functionality and integration.	The research involved analyzing annual trends and subject distribution, as well as creating a deep-learning model.	The study's findings are limited due to assumptions made in vote prediction, the complexity of the model, and resource constraints.

To overcome these challenges, we proposed a vector space model that focuses on semantic grouping based on feature words. The paper aims to organize news information into different categories based on the meaning of specific words.

2 Semantic grouping vector space model

The knowledge recognition process based on the semantic grouping vector space model is mainly for the problems that occur during database retrieval. If this problem is completely solved, the semantic grouping vector space model knowledge algorithm needs to be used to optimize the parameters of the model. This paper combines the strategy of weight sharing. Weight sharing refers to making the semantic grouping vector space model different in connection mode and parameter sharing mode to the conventional vector model. The semantic grouping vector space model can be locally connected and data information can be shared. Weight sharing mainly refers to the collection of parameter data based on multiple nodes in the hierarchical process. The feasibility analysis of shared data parameters is primarily related to various goals in the calculation process.

Different from the traditional method, the semantic grouping vector space model mainly uses the initial feature value of the input signal of the collected data and learns the input signal according to the hierarchical retrieval [3]. Usually, it includes the average time amplitude difference of the time domain features. This article uses the energy in a short time, etc., for the initial input characteristics of the model.

The Web news media search signal can be represented by x(n), and the short-term energy balance can be expressed as follows:

$$E_n = \sum_{m=-\infty}^{\infty} \left[x(n) w(n-m) \right]^2 = \sum_{m=-\infty}^{\infty} x(m)^2 h(n-m) = x(m)^2 * h(n) \quad (1)$$

However, the high dimensionality of the time-domain features under initialization will cause a lot of interference and noise. Therefore, the input search signal needs to be reduced in dimensionality.

The primary cause analysis method is used to make statistics on the multiple variables of the investigation, and the internal structure among various variables can be analyzed by studying multiple main components. After the data of Web news media is processed by dimensionality reduction, the input data information can be retrieved. In deep learning, the semantic grouping vector space model can perform data processing on the output data. Among them, the Web news media retrieval structure at this stage is shown in Figure 1.

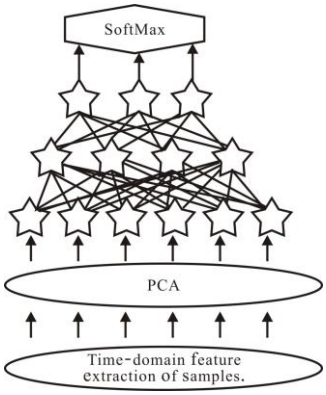

Figure 1: Semantic grouping vector space structural diagram.

2.1 Deep belief network algorithm in the audio segmentation stage

The Deep Belief Network (DBN) algorithm, utilized in audio segmentation, employs a hierarchical, generative model consisting of multiple layers of stochastic, latent variables. Initially trained layer by layer, the DBN captures intricate patterns in audio data through unsupervised learning. The top layer of the network functions as a discriminative model, facilitating the identification of relevant features for segmentation. Through iterative fine-tuning of weights during training, the DBN learns hierarchical representations, enabling the extraction of meaningful audio features. This hierarchical approach enhances the algorithm's efficiency in discriminating between various audio segments, thereby improving its ability to segment and classify different components within the audio signal accurately.

2.2 Knowledge recognition algorithm of semantic grouping vector space model

A news report contains four elements: time, place, person, and event. Therefore, for Web news information, at first, feature words are distinguished based on these four elements, respectively define the associated semantic groups to form 4 vectors, and then determine which vector space each feature word belongs to, and establish an inverse index corresponding to each vector space, calculate the weight value and similarity of feature words for each meaning group. Finally, the weighted sum of similarity is obtained, and the search results that are greater than a specific valve value are sorted using link analysis[4-5].

2.3 Weight and similarity of feature word

In Web news information, the position of the feature word on the document is different, and the importance of the document ability expressed is also other. The feature word level describes this characteristic. If the feature word TE appears n times in the document, as in formula 2, the feature word level score T_k is the k-th appearance level of the feature word T in the document.

$$score(T)=\sum_{k=1}^{n}\frac{1}{2^{\ln T_k}} \tag{2}$$

The weight of the feature word is

$$w(T)=score(T)\times idf_k \tag{3}$$

In the case of calculating the weighted similarity between the query Q and a specific news feature word group D, due to the large amount of calculation and time overhead of the conventional VSM similarity, the ratio of the weight value of the QD cross part to the sum of the QD weight value is used for calculation (4).

$$sim(Q,D)=\frac{\sum_{k=1}^{|Q\cap D|}\left(w(T_{qk})+w(T_{dk})\right)}{\sum_{k=1}^{|Q|}\left(w(T_{qk})+\sum_{k=1}^{|D|}w(T_{dk})\right)} \tag{4}$$

2.4 Semantic grouping vector space calculation method

For the vector space model, the conventional method of semantic grouping vector space calculation is to calculate the cosine similarity between vectors. The semantic grouping vector space of user u and Web news media d can be defined as:

$$Sim(u,d)=\frac{u\cdot d}{\|u\|\cdot\|d\|} \tag{5}$$

Regarding the probability model, the cosine similarity of vectors cannot be calculated by self-connection[6-7]. The following propositions are proposed to express the diversity of user interests.

Proposition 1. Assuming that user you have conditions independent of multimedia digital archive d in the predetermined classification model $C=\{c_1,c_2,\cdots,c_n\}$, the probability that multimedia digital archive d recommends to user u is:

$$p(u|d)=p(u)\sum_{j=1}^{n}\frac{p(c_j|u)\,p(c_j|d)}{p(c_j)} \tag{6}$$

Proof: It can be known from the total probability formula,

$$p(u,d)=\sum_{j=1}^{n}p\left(u,d\left|c_j\right.\right)p\left(c_j\right) \quad (7)$$

Assuming that the user u exists independently in the multimedia digital file d under condition C, so $p\left(u\left|d,c_j\right.\right)=p\left(u\left|c_j\right.\right)$, and then $p\left(u\left|d,c_j\right.\right)=p\left(u\left|c_j\right.\right)p\left(d\left|c_j\right.\right)$ is obtained. Therefore, formula (7) can be transformed into

$$p(u,d)=\sum_{j=1}^{n}p\left(u\left|c_j\right.\right)p\left(d\left|c_j\right.\right)p\left(c_j\right) \quad (8)$$

Get

$$p\left(u\left|d\right.\right)=\sum_{j=1}^{n}\frac{p\left(u\left|c_j\right.\right)p\left(d\left|c_j\right.\right)p\left(c_j\right)}{p(d)} \quad (9)$$

The purpose is to transform the semantic grouping vector space problem of the probability model into a situation of seeking conditional probability, presenting the diversity of user interests.

The adopted system has the memory of recording the user's search history and clicks and continues to search for the data information source of the user's operation behavior model. The system automatically completes this coherent operation, and the user experience is not disturbed. First, the user's historical search information in the browser is saved to learn the user's interest, and then the user's interest in the search information through the user's operation on the search results. Add time stamps to the data of interest. Therefore, update the points of interest that users need to be more interested. In the user interest model, the design process is shown in Figure 2.

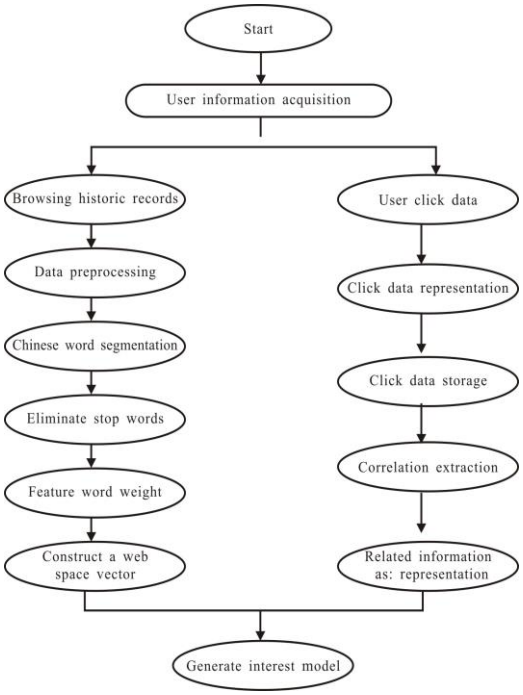

Figure 2: Design process of user modeling.

After completing the Chinese word segmentation with the IK Analyzer Chinese, the vector space model is constructed, and the weight w_i can be obtained through the frequency of occurrence of keywords k_i in the document through the calculation formula of TF-IDF.

$$w_i=tf_i\times\mathrm{idf}_i \quad (10)$$

Among them, tf_i represents the frequency of the keyword k_i appearing in all generated texts and idf_i represents the frequency of k_i in the reverse order of all generated texts. The calculation method is as follows:

$$\mathrm{idf}_i=\log\frac{N}{n} \quad (11)$$

Where N is the number of generated texts, and n is the number of all texts containing the keyword k_i.

The method of calculating the rights of keywords k_i can be adjusted as follows.

$$w_i' = w_i \times e^{-t} \tag{12}$$

$$d = \left(\left(k_1, w_1'\right), \left(k_2, w_2'\right), \cdots, \left(k_i, w_i'\right)\right) \tag{13}$$

When comparing the data of the model and the document $X\left(x_1, x_2, \cdots, x_n\right)$ that the user is interested in $W\left(w_1, w_2, \cdots, w_n\right)$, the size of θ is evaluated by calculating the angle θ of the vector and $X\left(x_1, x_2, \cdots, x_n\right)$ composition, which is inversely proportional to the degree of user concern. The smaller the θ, the higher the correlation between this file and the user's interests and preferences. The calculation formula is as follows.

$$sim\left(X, W\right) = \cos\theta = \frac{\sum_{i=1}^{n} X_i W_i}{\sqrt{\left(\sum_{i=1}^{n} X_i^2\right)\left(\sum_{i=1}^{n} W_i^2\right)}} \tag{14}$$

2.5 Web news media retrieval analysis

In multimedia digital archives, user Web news media retrieval is completed by three stages: collecting and analyzing user data, constructing user interest model, and updating user interest model. Suppose Web News Media wants to obtain user information. In that case, it must first get the user's obvious information, such as the user's registered account number, age, education, occupation, unit, keywords of interest, etc[8-9]. The user can modify and reply to this significant information, to gradually improve the information. However, some users are unwilling to provide accurate registration information due to personal privacy or time issues. To solve this problem, you can set up implicit information to extract the user's knowledge. The test will remove the bookmark of the keyword searched for, download and save the files, etc. According to the bookmarks maintained by the user, the downloaded and saved document information, the user's long-term concern, research field, and other issues can be determined, which can become important sources of information for establishing a model.

Because the user's interest is not fixed, it is necessary to establish an update mechanism when building a model to remove the forgotten topics in time to add new content, calculate the weight of the user's interest, and rank them according to the proportion of the weight. People's forgetting value is the trend of forgetting from the beginning and gradually becoming late. In the interest model system, the weight of the keyword of interest is multiplied by the update time, the weight of the phrase is sorted, and the forgotten interest topics are deleted. Complete the tracking of the effective behavior of the user and harvest a new keyword to recalculate the proportion. If the weight exceeds the threshold, it is added to the user interest model to complete the model update. It can be proved from Proposition 1 that according to the results of the ranking query based on the recommended ratio, the semantic grouping vector space model calculation can be used to query the media digital archive users. Based on p(u) of inequality (9), since p(u) does not interfere with the results of the recommendation probability, according to this method, the detailed explanation of the retrieval calculation of the multimedia digital archive users is performed.

Algorithm 1. Web news media retrieval analysis algorithm based on knowledge recognition of semantic grouping vector space model.

Input: domain classification model, user interest model, retrieval keywords, search engine, output: multimedia digital archive users' Web news media search results.

(1) According to the search keywords, the search engine is used to generate a preliminary search result set X.

(2) Set the number of iterations i=0.

(3) For the i-th Web news media in the set X, formula (1) is used to calculate the probability distribution in the field's classification model.

(4) Equation (9) is used to calculate the probability that the multimedia digital file i is recommended to the current user and added to the list Y.

(5) If the multimedia digital file i is the last multimedia digital file in the set X, go to (6); otherwise, set i=i+1 and return to (3).

(6) Sort and output the multimedia digital files according to the probability in the list Y in descending order.

Because the algorithm is actually based on another search engine, for each multimedia digital archive of search results, the probability distribution in the domain classification model must be calculated. This has a great impact on the performance of the algorithm. If the search engine calculates the probability distribution in the domain classification model of each Web news media in advance, the performance of the algorithm will be

significantly improved to meet the needs of real-time processing.

3 Web news media retrieval analysis process

3.1 Web news media retrieval

The four parts of the browser plug-in, personal manager, user model learner, and information personalized searcher, constitute the experimental system; as shown in Figure 3, it is the browser plug-in that provides users with convenient tools. After the user logs in and register information, the browser plug-in can be used to complete the Web news media retrieval of multimedia digital archives—no need to log in to the server. In addition, the browser plug-in mainly collects the user's personal information and transmits it to the server. The personal manager is used to manage the user's personal information, hobbies, and bookmarks through the personal manager. The purpose of tracking user behavior is to learn user interests. The information Web, a news media retriever, can complete the user's query and recommendation in the multimedia numbers calculated by the semantic grouping vector space model.

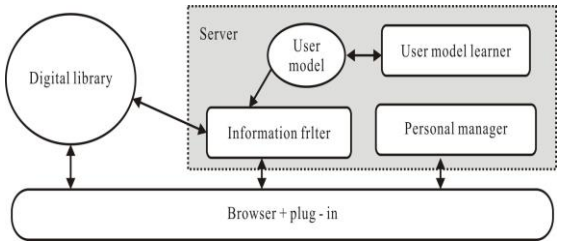

Figure 3: System architecture.

Our system can track the behavior of guests; it is distributed on the edge of the client and server and will not affect the customer's reading and system performance.

3.2 News page judgment and related information extraction information extraction

Through retrieval and learning, the link weights and node offsets of each layer are obtained, and the network initialization is completed. The reverse conduction algorithm (BP)is adopted, and the deep trust network model monitored from top to bottom is fine-tuned to overcome the shortcomings of local optimization and long search time. Although the performance of the deep belief network model shows strong characteristic learning ability, from the above principles, Internet search requires a large amount of sample data to generate more parameter values. On the other hand, based on the problem of Web news media recommendation search, there is a lack of a large amount of sample data, and it is found that the generation of a large number of parameters takes a long time, which is not good for practical applications. During the calculation process, the feature vector is used to represent the web page, and if the keyword weight We are determined by the TF*IDF method, and the term item is determined to be a named entity, the weight value shall be appropriately enhanced. The specific definition is as follows.

$$W_e = \begin{cases} idf_e \times \alpha, e \text{ is named entity} \\ idf_e, e \text{ is others} \end{cases} \quad (15)$$

Among them, α is the weighting factor, which is 5 in this experiment.

Finally, if m word items with large weight values are selected to generate web page feature vectors and applications. The number of shared term items in the two webpage feature vectors is used as the basis for judging similarity. If the number of shared terms is greater than the threshold, the two web pages are similar.

After determining the reprinting or similar relationship, relevant information is extracted and recorded. The leading information recorded is the reprinted website, the source site of the reprinted website, the number of responses to the reprinted website, and the time of the news release. The reprinted website and the source site here are only records of the reprinting relationship, not the finalized accurate source site and reprinted website. The last source site will be determined in the next step.

3.3 Judgment of news reprinting relationship and calculation of authority of news source sites

$$\text{News reprint rate (denoted as Trams)} = \frac{\text{reprint times}}{\text{source website clicks}} \quad (16)$$

However, since the reprinting relationship of Web news has two types, direct reprinting, and indirect reprinting, the source site cannot be determined initially, and the two attribute values of all nodes are initialized to 1 in the entire network layer. Then, in pt→qt, the website pt describes that the news of the website qt is reproduced. The content quality attribute value and the reprint attribute value are calculated using the following repetitive formula. The attribute value of all web pages is normalized to 1 when each iteration is completed.

$$A_0(pt) = \sum_{qt \to pt} A_1(qt) \quad (17)$$

$$A_1(pt) = \sum_{pt \to qt} A_0(qt) \tag{18}$$

$$A_0(pt) = \frac{A_0(pt)}{\left[\sum_{\forall pt} \left(A_0(pt)\right)^2\right]^{\frac{1}{2}}} \tag{19}$$

$$A_1(pt) = \frac{A_1(pt)}{\left[\sum_{\forall pt} \left(A_1(pt)\right)^2\right]^{\frac{1}{2}}} \tag{20}$$

Iteratively update the attributes of each node $A_0(pt)\ A_1(pt)$ according to the above formula.

The extracted reprinting information is used first to extract the relationship between news reprinting sites, calculate the authority value of each reprinting site, and use the website with the most reprinting times as the source site, including the relationship between direct reprinting and indirect reprinting, and that authority value is treated as the value of the reprint rate of the news.

3.4 Calculation of new response rate

The response rate (denoted as Rep) directly reflects people's reaction to Web news. usually

$$Response\ rate = \frac{amount\ of\ responses}{number\ of\ clicks} \tag{21}$$

Observation results show that most news pages only provide the number of answerers rather than the number of clicks/viewers. The number of clicks/views on the page is stored on the page server-side and cannot be obtained through simple capture and information extraction. Based on a large number of observations, a response rate ratio is summed up based on the relative number of news responses, and this ratio is used as the news response rate. Here, the number of responses is the total of the number of responses from the source site and the number of responses from the reprint website[8-9]. Figure 4 shows the distribution of the number of news responses.

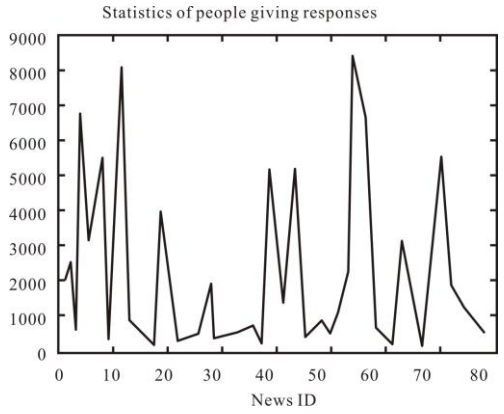

Figure 4: Statistics of the person-time number of responses.

It can be seen from Figure 4 that the number of responses to most news is within 1,000 people. There is very few news with more than 3,000 people. According to the statistical rules in the figure above, the relative recovery rate values are shown in Table 1. As an example, the number of responses (0~500) indicates the range of the number of people who responded to this event, and the relative response rate indicates that the number of people who responded to this event is between (0~500). It is considered that the number of people who responded to this event accounted for 5% of the number of viewers. If there are more than 5,000 respondents, those who have read this report will basically give answers, and the relative answer rate is 100%.

Table 1: List of relative response rates.

Number of responses	Relative response-rate (%)
5000~	100
4500~5000	90
4000~4500	80
3500~4000	70
3000~3500	60
2500~3000	50
2000~2500	40
1500~2000	30
1000~1500	20
500~1000	10
0~500	5

3.5 The influence of time factors on news ranking

There are usually two trends in people's interest in news, as shown in Figure 5. The attention here is measured by the number of news viewers per unit of time. The first is the slow-growing type of interest in knowledge such as national policy news. The timeliness of news in these categories is not strong, and people's concern is slowly increasing with time. The other is the type that grows rapidly and declines. It is mainly for news on current events; this kind of news is very time-sensitive. People's attention to this kind of news has increased rapidly in a short period, and after some time, the attention has quickly dropped [10-12]. Therefore, the sorting of news must first be classified and judged, taking into account the influence of time. From this perspective, the importance of news is inversely proportional to the time of publication.

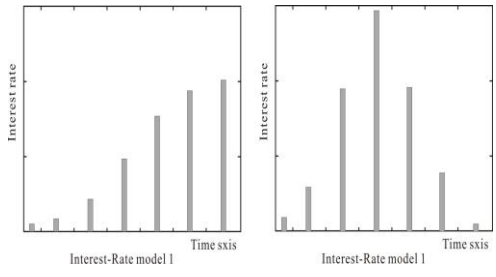

Figure 5: News attention.

In addition, the longer the release time, the higher the probability of reprinting and replying and the greater the number of responses and reprints. If the time factor is not taken into consideration, it is unfair to the newly published news. Therefore, when selecting parameters, the time factor will affect the importance of news. For reports with a long submission period, the number of responses and reprints will be reduced. Summarizing the above two points, combined with the definition of the news decay time parameter in literature [4], the definition of the time parameter is as follows.

$$D(t_s,t)=e^{-\alpha(t-t_s)} \tag{22}$$

Among them is the publication time of the news. The determination of α depends on the recession time of the news category to which the news belongs. Recession time refers to the time from the news release to the intermediate experience that no one cares about, and it is defined here. The relationship between news and recession time is $t_s\ t\geq t_s$

$$\alpha=\begin{cases}e^{-\beta\alpha}=0.5, Current\ affairs\ news\\ e^{-\gamma\alpha}=0.5, Non-current\ news\end{cases} \tag{23}$$

Here β is the decline time of current affairs news. γ is the decline time of non-current affairs news.

3.6 Judgment of news influence

Through the above steps, data on news reprint rate, news response rate, and influence factor of news source sites can be obtained(W_S), as well as the time parameter of the news release $D(t_s,t)$.

Reprinting and replying to news are considered to be the recognition of news, so the Web news recognition rate (denoted as Rec) is defined as news recognition rate = a. ×Reload rate+b×Recovery rate.

In order to ensure that the authorization rate is less than 1, the relationship between a and b is defined as a+b=1. Since there is no suitable corpus, the values of a and b cannot be obtained through the training method so these decisions can be obtained according to the 80/20 rule. There may be a lot of people watching the news, but few people answer it, and even fewer people understandably do repost. So, I think the reprint rate can better reflect the influence of news. Experiments show that this definition method is feasible[13-14].

Finally, combining the above information, define the influence of news (NF) as follows.

$$N_F=e^{-\alpha(t-t_s)}\times W_S\times\left(a\times Trans+b\times \mathrm{Rep}\right) \tag{24}$$

4 Experiment results and analysis

The performance evaluation of the information retrieval system is generally used as a benchmark, and the comprehensive evaluation rate F can also be used for evaluation (25)

$$F=\frac{2\times precision\times recall}{precision+recall} \tag{25}$$

4.1 Experimental data set

The top nine news information websites were tracked on the Chinese website rankings for a week. As many features based on the summary of Chinese webpages appeared in the algorithm research, Chinese webpages were still used as experimental subjects in the experiment. The news on the homepage of these nine websites is captured every hour. The list of captured experimental data is shown in Figure 6.

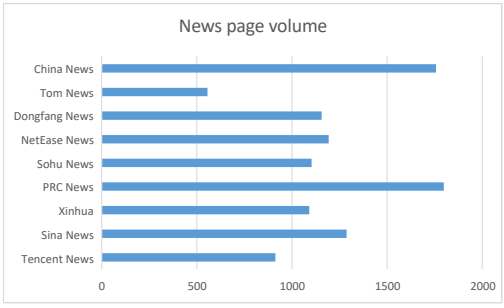

Figure 6: List of experimental web pages.

After the internal deduplication of the website, these pages are classified into six types according to their content. The news distribution of each category is shown in Figure 7. Here, the strength of news timeliness is obtained from the attention model to which news belongs.

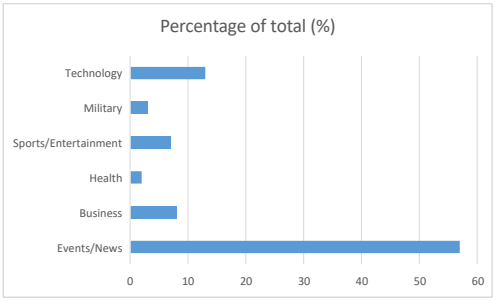

Figure 7: Classification list of experimental web pages.

4.2 Experimental result

The recommended task is executed using Python 3.5. Numpy, Pandas, Scikit-learn, Natural Language Toolkit (NLTK), and Matplotlib software are required to be installed alongside Python to carry out the procedure

(1) News influence ranking for the week from September 10 to 16th, 2007

Figure 8 shows the top 10 news and their influence values in the week from September 10 to 16th, 2007. Here, the recession time of current affairs news is defined as 72 hours instead of 120 hours. Figure 9 shows the distribution of news influence values within a week.

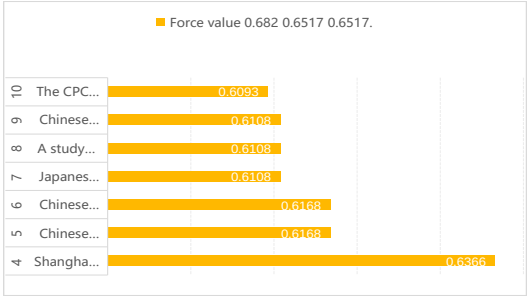

Figure 8: TOP10 news list from September 10 to 17th, 2007.

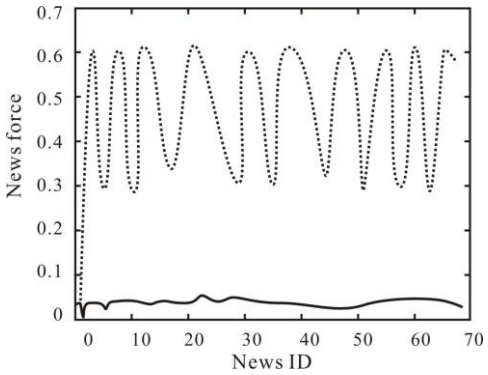

Figure 9: Distribution of news influence values collected from September 10, 2007, to September 16, 2007.

(2) Designated news influence ranking

The algorithm is also suitable for the sorting of designated news, giving some news on different topics, and using the topics of these news as keywords to search for relevant news pages using popular search engines. Select the top 50 from the search results for statistical calculations. The reason for ranking in the top 50 is that basically all reprinted pages and almost all similar pages, as well as the top 50 websites with news information in the Chinese website rankings, are included. It is sufficient to determine the source website of the news and the reprint rate of the news. After browsing many web pages, it can be found that all the comments of netizens are on the top websites. Netizens on other websites make almost zero responses, so selecting these pages can also get a more accurate news response rate value. After obtaining each news topic and reprinting the page, the relevant information is extracted to analyze the influence of each topic news according to the above algorithm, find the influence coefficient, sort according to the influence coefficient, and obtain the ranking result of the quantitative analysis. Next, we investigated the sorting results of these topics by multiple people. After synthesis, we got the sorting results of manual qualitative analysis. Finally, the consistency of the two results is compared. From the results of multiple comparisons, it is found that the sorting results calculated by this method are almost the same as the manual sorting results. The comparison results are shown in Figure 10 and Figure 11. Here is a comparison of the influence of news on non-related topics. Experiments show that this method is also applicable to related topics.

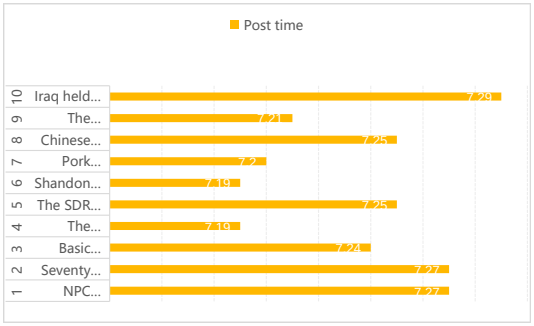

Figure 10: Results of manual ranking of designated news influence.

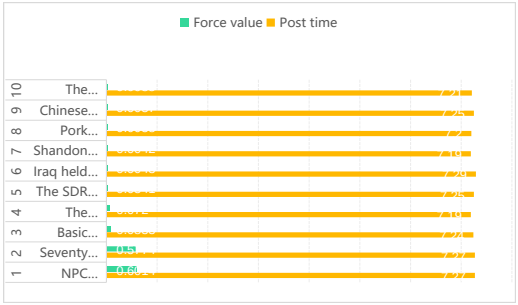

Figure 11: Specify the results of Web news media retrieval analysis ranking.

Query speed is a metric that measures the efficiency of a system in processing and responding to user queries. It quantifies the time it takes for a system to execute a search or retrieval operation and deliver the relevant results to the user. Table 2 and Figure 12 show the query speed result. While comparing the proposed method (SGVSM - 5 sec) with the other existing methods (Generalized Vector Space Model (GVSM)- 10 sec, TF-IDF -12 sec), it shows that our proposed method is superior for prediction accuracy in Web News Media Retrieval to other methods.

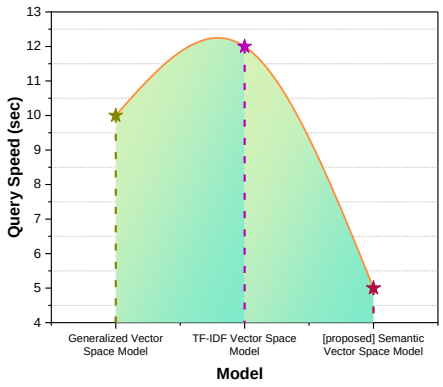

Figure 12: Query speed

Table 2: Query speed

Model	Query Speed (sec)
Generalized Vector Space Model [20]	10 seconds
TF-IDF Vector Space Model [21]	12 seconds
Semantic Vector Space Model[proposed]	5 seconds

Calibration rate is a metric used to assess the accuracy of predicted probabilities in a predictive model. It measures how well the predicted probabilities align with the actual outcomes or events. Table 3 and Figure 13 show the Calibration rate results. While comparing the proposed method (SGVSM - 82%) with the other existing methods (GVSM – 70%, TF- IDF -68%), it shows that our proposed method is superior for prediction accuracy in Web News Media Retrieval to other methods.

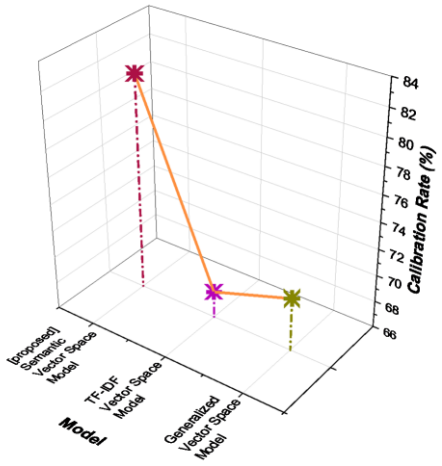

Figure 13: Calibration rate

Table 3: Calibration rate

Model	Calibration Rate (%)
Generalized Vector Space Model[20]	70
TF-IDF Vector Space Model[21]	68
Semantic Vector Space Model[proposed]	82

Accuracy is a metric that assesses the correctness of a proposed model's predictions by comparing them to observed values. The correctness of case predictions is measured as a percentage of complete occurrences. Figure 14 and Table 4 depict the comparative evaluation of accuracy in suggested and traditional methods. When compared to currently existing methods such as GVSM

and TF-IDF, which have accuracy values of 85% and 82%, respectively, the suggested SGVSM achieves an accuracy value of 90%. Our proposed method provided superior results for Web News Media Retrieval.

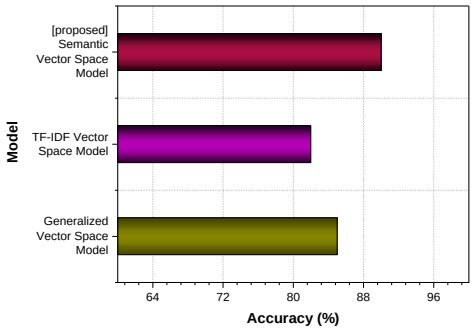

Figure 14: Accuracy

Table 4: Accuracy

Model	Accuracy (%)
Generalized Vector Space Model[20]	85
TF-IDF Vector Space Model[21]	82
Semantic Vector Space Model[proposed]	90

Precision is a fundamental parameter utilized in the field of statistics to evaluate performance. Figure 15 and Table 5 depict the comparative evaluation of precision in suggested and traditional methods. When compared to currently existing methods such as GVSM and TF-IDF, which have Precision values of 85% and 82%, respectively, the suggested SGVSM achieves an accuracy value of 90%. Our proposed method provided superior results for Web News Media Retrieval.

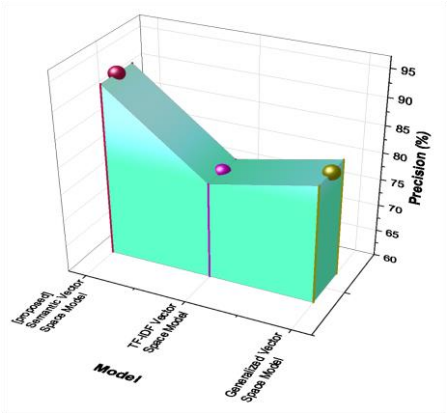

Figure 15: Precision

Table 2: Precision

Model	Precision (%)
Generalized Vector Space Model[20]	82
TF-IDF Vector Space Model[21]	78
Semantic Vector Space Model[proposed]	92

Recall is a performance metric used in data categorization that represents the proportion of actual positives that a model correctly retrieves. The recall result is shown in Table 6 and Figure 16. When compared to currently existing methods such as GVSM and TF-IDF, which have Recall values of 88% and 85%, respectively, the suggested SGVSM achieves an accuracy value of 92%. Our proposed method provided superior results for Web News Media Retrieval.

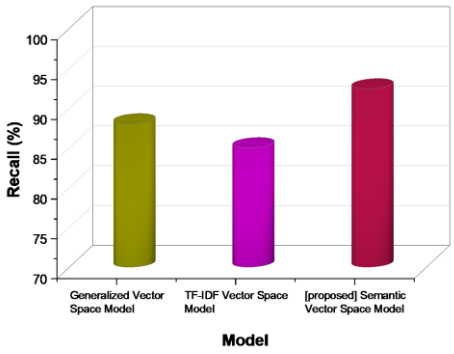

Figure 16: Recall

Table 6: Recall

Model	Recall (%)
Generalized Vector Space Model[20]	88
TF-IDF Vector Space Model[21]	85
Semantic Vector Space Model[proposed]	92.4

Computational time, also known as execution time or runtime, refers to the amount of time it takes for a computer program or algorithm to complete its execution. It is a crucial metric in evaluating the efficiency and performance of computational processes. The recall result is shown in Table 7and Figure 17. When compared to currently existing methods such as GVSM and TF-IDF, which have Recall values of 10.9 sec and 8.7 sec, respectively, the suggested SGVSM achieves an accuracy value of 6.4 sec. Our proposed method provided superior results for the Retrieval of news media from the web.

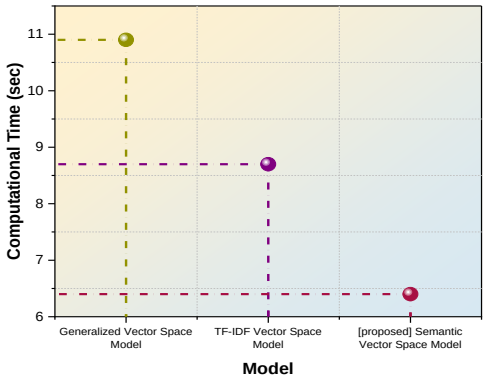

Figure 17: Computational time

Model	Computational time
Generalized Vector Space Model[20]	10.9
TF-IDF Vector Space Model[21]	8.7
Semantic Vector Space Model[proposed]	6.4

Table 7: Computational time

5 Implication

The combination of Web News Media Retrieval Analysis and Knowledge Recognition of Semantic Grouping Vector Space Model has significant practical applications for information retrieval and analysis in the field of online news media. This strategy seeks to improve the efficiency and relevancy of web news searches by integrating sophisticated approaches in information retrieval, semantic grouping, and knowledge recognition. The model enhances the extraction of significant insights by identifying semantic relationships within the material, allowing for more precise categorization and collection of articles based on their underlying knowledge structures. This integration enhances a news retrieval system by making it more sophisticated and contextually aware. Additionally, it has the potential to improve the user experience by delivering more coherent and informative results. Furthermore, this approach may be utilized in several domains, such as journalism, research, and data analysis, providing a valuable instrument for effectively navigating and understanding the extensive and ever-changing realm of web-based news media.

6 Discussion

The generalized Vector Space Model may struggle to capture semantic nuances and ambiguity in language. Words with multiple meanings or contexts may be represented by a single vector, leading to potential confusion. TF-IDF Vector Space Model faces challenges with synonymy and polysemy, where different words may have similar meanings or a single word may have multiple meanings. When implementing a proposed Semantic Vector Space Model, achieving superior performance in capturing semantic nuances and addressing the challenges of ambiguity becomes possible. This model endeavors to represent words in a multidimensional space, taking into account their semantic relationships, thereby offering a more nuanced and context-aware representation. By considering the inherent meanings and associations between words, the Semantic Vector Space Model aims to enhance accuracy and effectiveness.

7 Conclusions

Compared with previous vector space models, the vector space model based on the semantic grouping of feature words proposed in this paper is more accurate for Web news information and is suitable for the retrieval of Web news information systems. The calibration rate of document query, the comprehensive evaluation rate F, and query speed have been significantly improved. The current situation is personalized services. The general retrieval system in the past can no longer meet the retrieval requirements in different environments, purposes and different times. This paper has carried out a series of research and analysis on Web news media retrieval. Through experiments, we can see the interference factors for the calculation of the semantic grouping vector space model. Experiments have proved that the accuracy of analysis has been improved, and the interests and needs of users can be correctly expressed so that the accuracy of Web news media retrieval is further enhanced. Keeping up with real-time updates and delivering the latest news can be a challenge. Systems might struggle to provide up-to-the-minute information due to processing delays or the dynamic nature of news. In future research, systems may focus on improving semantic understanding to provide more accurate and contextually relevant results. This could involve advanced natural language processing techniques, including sentiment analysis, entity recognition, and topic modeling.

Acknowledgements

2022 Hunan Natural Science Foundation "Research on the Construction of Teaching Innovative Team of Higher Vocational Teachers Based on the Background of Improving Quality and Cultivating Excellence" （No.: 2022JJ60020）

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest

Funding Statement

This study did not receive any funding in any form.

References

[1] Solainayagi, P., & Ponnusamy, R. (2019). Trustworthy media news content retrieval from web using truth content discovery algorithm. Cognitive Systems Research, 56(AUG.), 26-35.

[2] Solainayagi, P., & Ponnusamy, R. (2019). Trustworthy media news content retrieval from web using truth content discovery algorithm. Cognitive Systems Research, 68(5), 566-588.

[3] Kai, S., Wang, S. , & Liu, H. . (2018). Understanding User Profiles on social media for Fake News Detection. FakeMM'18 Workshop, 31(6), 1-7.

[4] D Corney, Gonzalo, J., Martinez, M. , Poblete, B. , & Valochas, A. . (2018). Recent Trends in News Information Retrieval, 24, 1012‑1017.

[5] Davies R. J. (2016). Digital computing techniques in the manufacture and operation of engine management systems. Aeronautical Journal, 79(776), 349-353.

[6] Solainayagi, P. , & Ponnusamy, R. . (2019). Trustworthy media news content retrieval from web using truth content discovery algorithm. Cognitive Systems Research, 56(8), 26-35.

[7] Chen, H. , Huang, B. , Liu, W. Z. , Gao, Y. B. , & Jiang, X. Y. . (2019). Python-based web news crawler and retrieval. Software Guide, 47(1):1-38.

[8] [Dewandaru, A. , Supriana, I. , & Akbar, S. . (2017). Keyword and event extraction for thematic map retrieval from indonesian online news site. Journal of Physics Conference, 38(8), 10320-10342.

[9] George Dimitrakopoulos, Panagiotis Demestichas, & Vera Koutra. (2012). Intelligent management functionality for improving transportation efficiency by means of the car pooling concept. IEEE Transactions on Intelligent Transportation Systems, 13(2), 424-436.

[10] Fei-Yue Wang. (2011). Intelligent systems and social management. IEEE Intelligent Systems, 26(6), 2-3.

[11] Meesad, P. . (2021). Thai fake news detection based on information retrieval, natural language processing and machine learning. SN Computer Science, 2(6), 1-17.

[12] Jing Fan, Tianyang Dong, Xinxin Guan, & Ying Tang. (2013). A rapid simulation system for decision making in intelligent forest management. IEEE Intelligent Systems, 28(5), 2-9.

[13] Pe?A-Araya, V. , Quezada, M. , Poblete, B. , & Parra, D. . (2017). Gaining historical and international relations insights from social media: spatio-temporal real-world news analysis using twitter. Epj Data Science, 6(1), 25.

[14] Abbas Rajabifard, Russell G. Thompson, & Yiqun Chen. (2015). An intelligent disaster decision support system for increasing the sustainability of transport networks. Natural Resources Forum, 39(2), 83-96.

[15] Frasincar, F., Borsje, J. and Levering, L., 2009. A semantic web-based approach for building personalized news services. International Journal of E-Business Research (IJEBR), 5(3), pp.35-53.

[16] Nkongolo Wa Nkongolo, M., 2023. News Classification and Categorization with Smart Function Sentiment Analysis. International Journal of Intelligent Systems, 2023.

[17] Bangyal, W.H., Qasim, R., Rehman, N.U., Ahmad, Z., Dar, H., Rukhsar, L., Aman, Z. and Ahmad, J., 2021. Detection of fake news text classification on COVID-19 using deep learning approaches. Computational and mathematical methods in medicine, 2021, pp.1-14.

[18] Štrimaitis, R., Stefanovič, P., Ramanauskaitė, S. and Slotkienė, A., 2021. Financial context news sentiment analysis for the Lithuanian language. Applied Sciences, 11(10), p.4443.

[19] Sun, N. and Du, C., 2021. News text classification method and simulation based on the hybrid deep

[20] Tsatsaronis, G. and Panagiotopoulou, V., 2009, April. A generalized vector space model for text retrieval based on semantic relatedness. In *Proceedings of the Student Research Workshop at EACL 2009* (pp. 70-78).

[21] SUBA, C., Retrieval of Information Document Using TF-IDF Algorithms and Vector Space Model Representation.