Web Crawling Algorithm Fusing TF-IDF and Word2Vec Feature Extraction

Xinyue Feng

School of Electronic Information, Foshan Polytechnic, Foshan, 528137, China
E-mail: xinyue6570@fspt.edu.cn

Received 14 March 2025; Accepted 06 May 2025

Abstract

Current research focuses on how to efficiently extract and crawl network information because, with the growth of the Internet, network information is becoming more and more diverse. To address the problem of incorrect data extraction and topic judgment of web crawlers, this study proposes a novel approach based on a file inverse frequency algorithm and Word2Vec feature extraction. The new method improves the retrieval capability of web crawlers by using the file inverse frequency algorithm and uses Word2Vec to extract data features, which improves the data extraction capability of current crawlers. The results showed that the F1 values of the research use model were 25.8% and 26.2% higher than those of the digital filtering algorithm, respectively. The total number of localization resources for the research use strategy was 2800 and the network coverage was 81%, which was 12% higher than the optimal strategy. The research use strategy had a shorter retrieval time and the model could recognize the vocabulary of the keywords. Finally, the model used by the research also had a good model processing

Journal of Web Engineering, Vol. 24_5, 713–738.
doi: 10.13052/jwe1540-9589.2452

capability when compared to other models. In summary, the new model built by the research can improve the data retrieval ability and data extraction ability of the web crawler, which provides new research ideas for future web information extraction.

Keywords: Network information, retrieval, internet worm, TF-IDF, Word2Vec, data extraction.

1 Introduction

The network data volume is increasing at an exponential rate due to the fast development of Internet technologies. Internet worm (IW) technology is becoming more and more important as a vital tool for gathering and organizing web data in this era of explosive information [1]. IW can automatically collect information from the Internet to support various data-intensive applications such as search engines (SE), market analysis, social network analysis, etc. However, traditional IWs are finding it more and more challenging to handle vast amounts of heterogeneous network data as a result of the network environment's complexity [2]. The selection of feature extraction techniques becomes crucial to enhancing crawler performance during the information extraction process. In the field of text analysis, the term frequency-inverse document frequency (TF-IDF) technique has been around for a while and is particularly useful for extracting keywords and determining word importance [3]. It is also able to reflect the importance of words in documents by calculating the inverse ratio of their frequencies in documents to their distribution frequencies in the corpus. However, the contextual link and semantic information between words are ignored by the TF-IDF algorithm, which only takes word frequency and inverse document frequency into account. Word2Vec is a neural network-based word embedding technique that turns words into a set of numerical vectors while capturing the intricate semantic links between words [4]. In this way, Word2Vec not only preserves the semantic information of words but also provides more in-depth semantic parsing when dealing with natural language tasks. Moreover, in the study of web information Word2Vec and TF-IDF algorithms are widely used in web data processing. In order to cope with the increasing security threats in IoT, Aqeel et al. analyzed more than 170 research articles, proposed threat classification and analysis for IoT systems, and explored the development of recovery mechanisms. The study's findings demonstrated that integrating cutting-edge technologies like blockchain, AI, and machine learning provided

a practical means of ensuring security and privacy in Internet of Things systems [5]. Deng et al. conducted research and suggested a susceptibility infection recovery death (SIRD) model to simulate the propagation process of computer worms in order to obtain a deeper understanding of the propagation mechanism and features of worms. For parameter estimation, the new model included the least squares method, the Markov chain Monte Carlo method, and the ensemble Kalman filtering method. It was based on the ordinary differential equation model. According to the findings, the suggested SIRD model outperformed the conventional model in parameter estimation [6]. In order to forecast words in social media and perform time-based prediction, Lubis et al. suggested a way to perform word prediction for Twitter and interpret tweets as weighted documents based on the TF-IDF algorithm. By using time series prediction on the test data of 1734 tweets, the study's findings showed that the results had a high degree of accuracy. Furthermore, two categories of active and inactive users were created for the word prediction, and the MAPE computation was able to achieve 50% [7]. In order to maximize the efficiency of keyword extraction in information retrieval techniques, Cheng et al. examined the current TF-IDF methodology and suggested an enhanced TF-IDF algorithm. Using information entropy and relative entropy from information theory as calculation parameters, the program optimized keyword extraction. According to the study's findings, the algorithm could greatly increase keyword extraction's efficiency and accuracy [8]. Ao et al. put forth a novel technique to increase keyword extraction accuracy by combining the benefits of the TF-IDF and TextRank algorithms. The new method used a similar historical news library weighted TF-IDF and TextRank algorithm (TFSL-TR). First, the news was categorized using an LSTM-based classification model, then the TF-IDF value and TextRank value are calculated, and finally the two are summed up by weights and TopK words are taken as keywords. According to the study's findings, the TFSL-TR algorithm could significantly raise keyword extraction accuracy [9]. Because of this, the research suggests an algorithmic model based on the combination of TF-IDF and Word2Vec feature extraction. Traditional IW research is therefore insufficient to improve the crawler's retrieval ability and data extraction (DE) ability. By merging the features of topic networks, the model offers a more comprehensive DE approach and enhances IW's web data retrieval capability through the use of the TF-IDF algorithm. The model also incorporates the Word2Vec feature extraction method to feature the existing topic vocabulary in order to enhance the topic judgment and feature extraction ability of the IW. Meanwhile, the research innovatively combines TF-IDF and Word2Vec

to enhance the DE efficiency of the IW and the accuracy of topic information collection through the construction of topic IW.

2 Methods and Materials

2.1 Analysis of Internet Worm Search Engines and Topic Crawling Techniques

SE is a special way to retrieve some specific data and information from IoT data, process this data content, build an IW SE database through data analysis, and provide more friendly data information to users through SE [10]. The network SE working process is shown in Figure 1.

As illustrated in Figure 1, the SE first retrieves data information from the Internet via the IW, and subsequently generates the original web page interface using this data. The SE is capable of performing data processing, classification, data mining, and other operations on the web page. Subsequently, the acquired data are stored in order to produce the relevant SE. Ultimately, the data are deposited in the SE database. Users have the ability to simultaneously store the data engine and search the data via the web page. The IW technology must mine and classify the current data information to ensure the functionality of web data and search. For this reason, IW technology is the key to the operation of web SE [11]. Figure 2 is the process of IW technology operation.

In Figure 2, the IW will first obtain the required localization resources from the web structure, such as the current data to be obtained is apple, then the IW will first search from the fruit database. Second, in the process of searching by the crawler, it will extract the corresponding localization

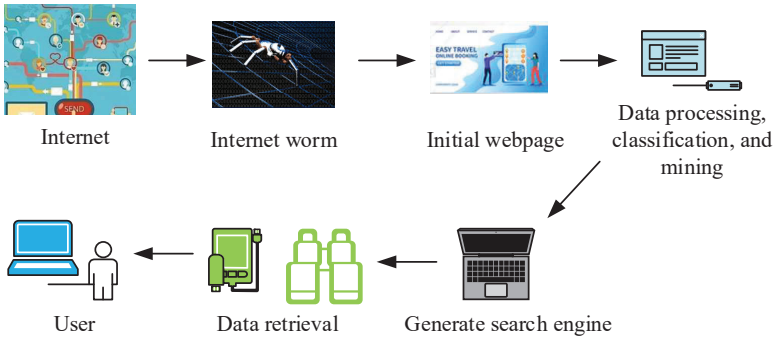


Figure 1 Working process of the network search engine.

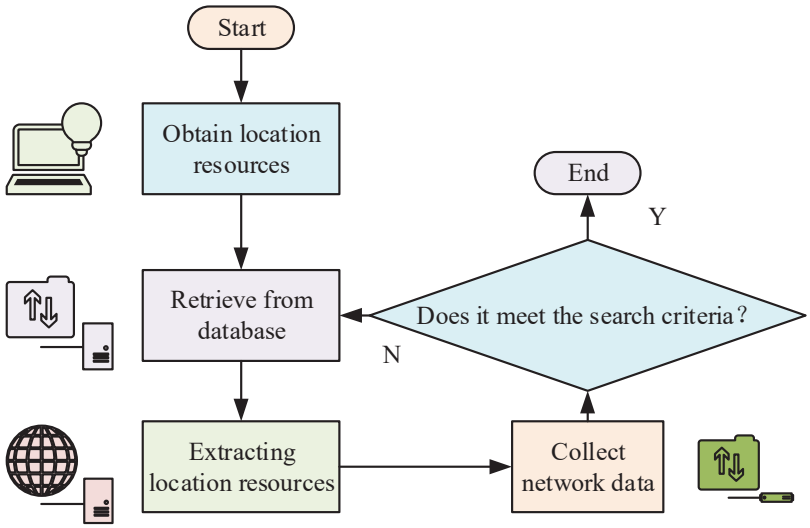


Figure 2 Operation process of web crawler technology.

resources, and arrange the resource data that need to be acquired in the order of localization resources before and after. The crawler collects the currently required network data from the located resources. The process is looped without interruption until the crawler searches and grabs all current location resources. Finally, it determines whether the search condition is satisfied or not. If it is satisfied, the data is output and if not, the data is reacquired and retrieved.

The IW structure mainly includes the modules of data scheduling, data acquisition, data parsing and data analysis; data scheduling is the deployment and scheduling of resource data in the web page so that the current data can be better extracted and analyzed [12]. For the data storage function, there is a need to ensure that the current database can contain a larger data space and better data analysis capabilities. Meanwhile, retaining document data needs to be irrelevant to ensure that the data program is not retrieved by other SEs [13]. At the same time, the use of a get database has a strong data storage function. Therefore, the current database selects a MySQL database as the data storage database [14]. Since the commonly used IW has the problem of poor purposefulness in the collection and engine search of web data, a topic IW technique is proposed to address this problem. The IW theme is required to initiate the search data, assess the relevance of the current data and the search data, and ensure the positioning resources of the current data

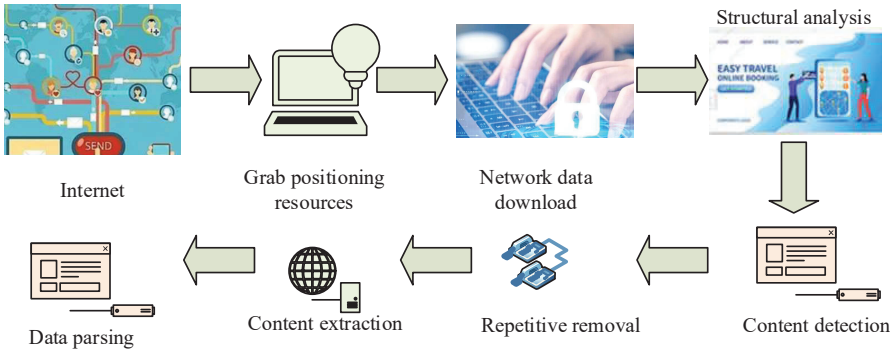


Figure 3 Main structure of the field theme network crawler.

by formulating the corresponding search process and strategy. This is done to present the corresponding theme in the crawler retrieval process [15]. Therefore, the judgment of the IW theme in the current study is crucial. Figure 3 shows the wield theme IW main structure.

In Figure 3, the subject IW will first respond to the request for data positioning resources through the Internet, followed by the need to crawl through the positioning resources in the network data download; positioning resources are needed to analyze the structure, content detection and prediction and then extraction and duplication removal and then crawler retrieval. The retrieved web data is then subjected to the next step of content extraction and the data parsing process. Finally, the relevance of the web pages is analyzed by filtering the data to get the final topic crawler network database.

2.2 Theme Crawler Network Building for TF-IDF and Word2 Vec

For the construction of the topic crawler network, it is necessary to statistically analyze the network data, and the TF-IDF algorithm is a data-weighted statistical algorithm that is able to reflect the degree of contribution of the topic data obtained from the IW, so as to better retrieve the information of the IW data. In addition, the algorithm can lower the theme weights of recurring words throughout the process of creating thematic crawler network data, allowing for a more efficient statistical analysis of the existing vocabulary. One of the most effective algorithmic processes is through the initial document data vocabulary and the current occurrence of data frequency statistics. Equation (1) depicts the expression [16].

$$f_{y,x} = yf(y,x) \tag{1}$$

In Equation (1), $f_{y,x}$ denotes the frequency of occurrence (FoO) of the current search term and y denotes the occurrence of the term; x denotes the document and $yf_{y,x}$ denotes the frequency of the current vocabulary appearing in the document. Equation (2) shows how the main degree of contribution is computed. The inverse document frequency, which measures the role of the current vocabulary feature in the network-wide data, reflects the degree of contribution of the current data vocabulary [17].

$$idf(t, D) = \log \frac{N}{|x \in D, y \in x|} = \log \frac{N}{n_y} \tag{2}$$

In Equation (2), t is a particular vocabulary and N is the total number of documents in the current search database. D denotes the database and d denotes a particular data file in the database. Among them, the size of the documents is the same as the size of the vocabulary data in the database, n_t is the number of documents in which the vocabulary appears in the database. Due to the diversity of the data network, different metaphorical vocabularies, harmonic vocabularies, etc. can interfere with the topic judgment of IW [18]. Therefore, judging the keywords can cause difficulties in understanding the data vocabulary and there is a strong subjective factor. Second, the theme key words obtained in this way cannot judge the size of its relevance when matching with the network text information. Therefore, it cannot help the next crawling strategy; adding the TF-IDF algorithm to the theme crawler judgment can better avoid this situation [19]. Figure 4 shows the theme crawler data judgment process.

In Figure 4, all topic information in the domain is capable of performing data preprocessing and data analysis when the keyword weight of the vocabulary is high. However, in order to join the algorithmic model to judge the topic vocabulary, the initial data information must be obtained first. At this point these topic data can be transmitted into the machine for learning through the documents of the relevant neighborhood. Additionally, because of the data features and relevant data, the crawler SE is able to continuously assess the topic relevance of the data web page information. This allows it to better mine the relevance and relevant characteristics of the topic vocabulary. Following data processing, the output topic data is fed into the TF-IDF algorithm to determine the weight size of the current network data, or the TF-IDF value of the current data network.

When the TF-IDF value is calculated, the current topic data features are trained on the vocabulary by the Word2Vec model to obtain the correlation and feature size of the data features [20]. However, due to the small number

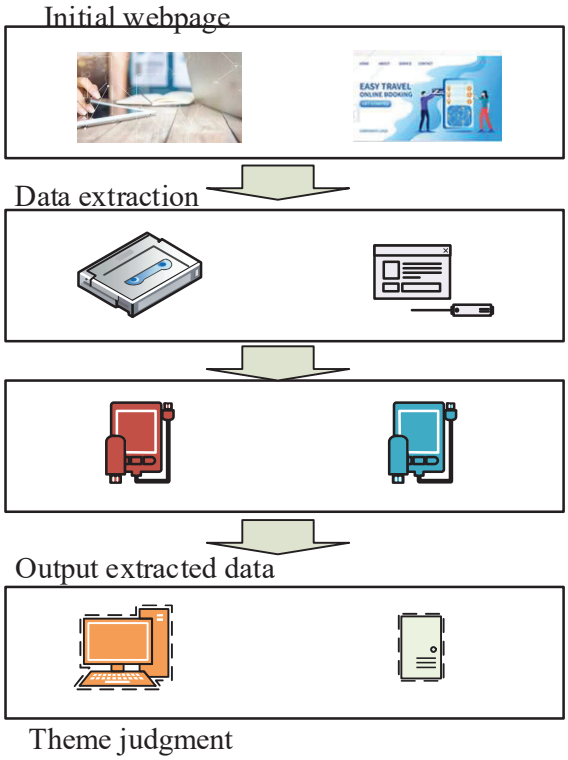


Figure 4 Theme crawler data judgment process.

of features obtained from the original data, the feature data obtained by correlation is not accurate enough, and irrelevant topic features tend to appear. For this reason, by collecting non-thematic data samples, the non-relevant web page topic samples are feature extracted, and then more effective predictive topic feature values are obtained by positive feature data filtering. Secondly, according to the calculated TF-IDF value to establish the theme feature gradient, each layer of the gradient is the initial keywords and corresponding weights. Each lower layer of the gradient of the feature value size is the previous layer of the feature prediction value, corresponding to the size of the weights. Finally, the theme network of the feature value inventory is used to get the final theme crawler similarity [21].

Due to the varying nature of the subject network, data labels, and subject vocabulary in relation to the influence of the feature vector, it is possible to construct the structure of the text of different types of expressions and to

calculate the different network weights, thereby improving the proportion of weights of the size of the current feature value [22]. In this study, the TF-IDF technique is used to compute the weights of the crawler, as shown in Equation (3). However, the use of separate data vocabulary is required to compute the contribution value of the subject data, so it is easy to find the weight deviation [23].

$$T_{wf}(k) = \frac{\sum_{i=0}^{n} m(i)}{\sum_{j=1}^{N} m(j)} \tag{3}$$

In Equation (3), $m(i)$ denotes the weight size at label i, k denotes the number of feature data, $\sum_{i=0}^{n} m(i)$ denotes the total weight size of the label at that number of features, $\sum_{j=1}^{N} m(j)$ denotes the total weight size of all labels, $T_{wf}(k)$ is the average weight size of all labels. The quantity of feature data that is displayed at this time, is indicated by Equation (4) [24].

$$tf_{ik} = \frac{n_{ik}}{\sum_{i=0}^{N} n_{mk}} \tag{4}$$

In Equation (4), n_{ik} denotes the FoO of the ith feature value in the kth network data, $\sum_{i=0}^{N} n_{mk}$ denotes the total frequency size of feature values in the current data network, f_{ik} denotes the total frequency of feature data appearing in the kth network data of the i eigenvalues, and t denotes the word frequency. The less feature data that occur, the smaller the data network feature weights that are obtained, and the higher the total frequency at this point in time, the lower the data network contribution value. It is essential to discuss the feature theme of the frequency and provide various positive and negative correlations in addition to the statistical analysis of the feature frequency of the current data document. Equation (5) shows the improved weights formula for positive and negative correlations [25].

$$tw_{ik} = \frac{\sum_{i=0}^{N} tf_{ik}}{N} * \frac{r \sum_{m=0}^{N} n_{mk}}{\sum_{m=0}^{N} n_{mk} - \sum_{i=0}^{N} n_{ik}} \tag{5}$$

In Equation (5), tw_{ik} denotes the global importance of the positively correlated features in the subject network data, and $\sum_{i=0}^{N} tf_{ik}$ is the sum of the times the feature documents appear. N is the documents, $\sum_{i=0}^{N} n_{ik}$ denotes the sum of the number of relevant features of the subject network appearing in the documents, and $\sum_{m=0}^{N} n_{mk}$ is the sum of the features in the current topic network. When the number of features occupies an increased

proportion in the whole document, the sharing of the subject network data at this time increases. r indicates the size of the positive convergence of the subject data network. This parameter can reduce the impact of the subject network due to the proportion of the correlation feature data being too high, caused by the weight overflow. When the degree of contribution of the frequency occurrence of relevant and irrelevant topics in the document varies greatly, the degree of influence of feature data on the document will be reduced, as shown in Equation (6) [26].

$$idf_i = \frac{a}{b} \log \frac{N}{1 + n_i} \tag{6}$$

In Equation (6), idf_i is the FoO of the non-relevance feature of the document at the feature value of i and N is the total size of the non-relevance document. n_i is the total number of documents in which non-relevant documents appear in the subject document at feature value i, a is the network data in which the feature value occurs in the subject network data, and b is the network data in which the feature occurs in the non-relevant network data. The subject share of a feature value is reduced when the value of idf_i is higher, and increased when idf_i is smaller. When the frequency of data network features appearing in the correlation network is positively correlated with the frequency of the topic network, and the non-correlation network is negatively correlated, the correlation change of the network needs to be obtained by adding the weights of the two network topics, as shown in Equation (7).

$$\textit{TF-IDF} = T_{wf}(k) * tw_{ik} * idf_i \tag{7}$$

The TF-IDF technique can evaluate and compute both positive and negative correlations based on Equation (7). The Word2Vec model is required to achieve the topic network's feature extraction. It is vital to evaluate the network model and key feature extraction in order to actualize the subject crawler's main DE. Therefore, the study adopts the Word2Vec model for DE, which is based on the procedure of first computing the probability of the word vectors of the topic network as indicated in Equation (8), in order to increase the extraction effectiveness of the current topic crawler [27].

$$h = \frac{1}{2c} \sum_{-c \le j \le c} VWotd_{(n+j)},\ j \neq 0 \tag{8}$$

In Equation (8), $Wotd_{(n+j)}$ denotes the textual information in the topic network and h denotes the probability. c denotes the window size of the topic

network, $VWotd$ denotes the word vector size, and j and n are real numbers. Secondly by processing the text information we can calculate the set of words as shown in Equation (9).

$$S = [word_1, word_2, \ldots word_n] \tag{9}$$

In Equation (9), S denotes the set of text vocabulary. The set of vectors of the vocabulary is then computed as shown in Equation (10) [28].

$$VS = [Vword_1, Vword_2, \ldots Vword_n] \tag{10}$$

In Equation (10), VS denotes the set of current word vectors. The third step is to get the full text word vectors by accumulating them, as shown in Equation (11) [29].

$$VText = \frac{1}{n} \sum_{i=1}^{n} VWord_i \tag{11}$$

In Equation (11), $VText$ denotes the lexical computation vector of the full text. Finally, the feature vectors of words are obtained by similarity calculation as shown in Equation (12).

$$SemWord_i = \cos(VWord, VText) \tag{12}$$

In Equation (12), $SemWord_i$ represents the feature size of words. After the feature vector process, we can obtain the word vector similarity of the text. The obtained text similarity and feature data are subject filtered to finally get the corresponding subject data, and finally the features are arranged by building a feature vector tree to get the final gradient feature data.

2.3 Theme Crawler Feature Strategy Analysis and Algorithm Implementation

As the crawler will get more localization resources from different networks when performing web parsing, this creates the situation where the relevance of lexical topics will be different in the process of implementing IW. Therefore, how to reduce the relevance difference of all the web pages number by different localization resources is the main problem to solve during the operation of IW. Thus, the study proposes a deep traversal crawling strategy, firstly the strategy learns by strengthening the crawler network to obtain more information about the web data, as shown in Figure 5 for the web page learning process.

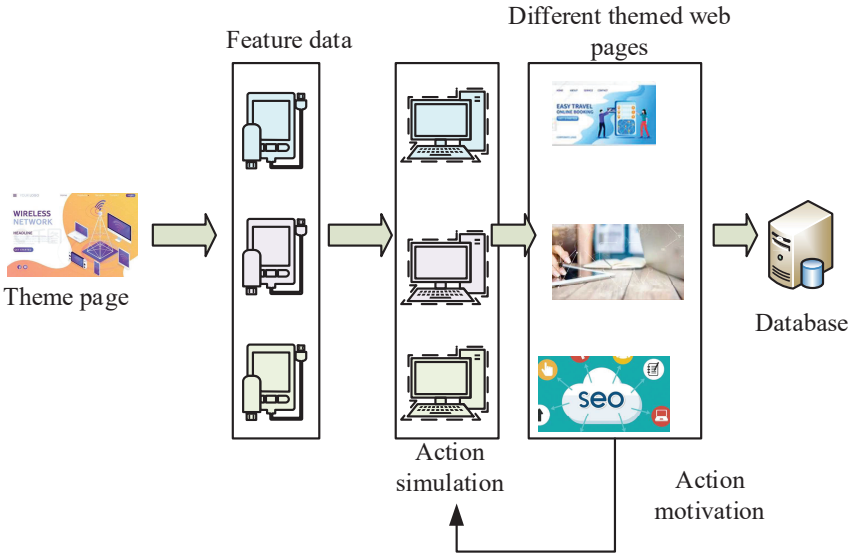


Figure 5 Web page learning process.

The key feature data of the present network will be obtained by the IW initially crawling the dynamic topic page of the web page, as shown in Figure 5. This data will then be simulated through the feature data, with the related topic information page of the current web page being analyzed through the simulation action. The feature action will then be rewarded through the topic information page in order to improve the simulation action. Finally, all the topic information data will be transferred into the feature database. Analyzing the actions of IWs can improve the network information processing ability of the crawler strategy [30]. The entire crawler network is traversed to obtain the crawler network strategy as shown in Figure 6.

In Figure 6, the crawler network strategy is to firstly determine the initial interface of the network for one-stage localization resources, followed by analyzing the localization resources of the current web page of the network DE to get M localization resource networks. Then it combines the one-stage localization resource data and M localization resource networks. Then the topic network is judged. Judge whether the current topic network exists relevance or not, if there exists data relevance then extract the localization resources of one stage again by the above action simulation. If no relevance exists then the crawler search strategy is directly ended [31]. Therefore, the model of the crawler algorithm is constructed as illustrated in Figure 7 in

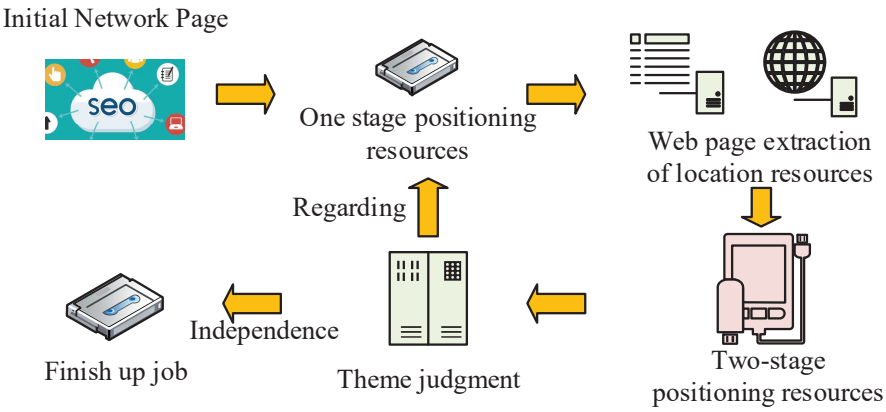


Figure 6 Crawler network strategy.

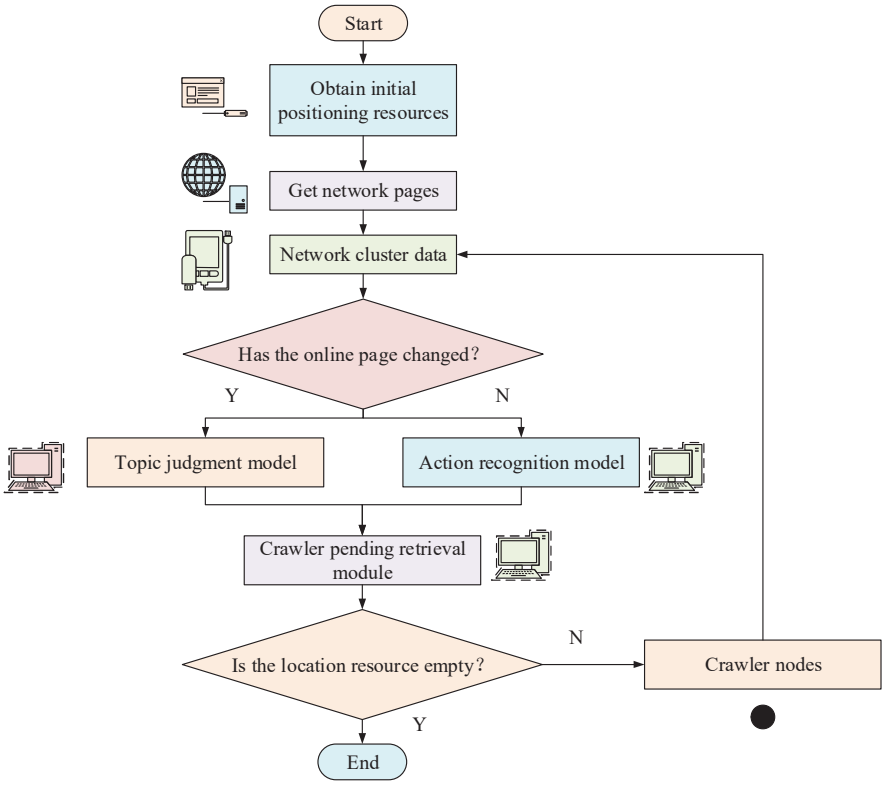


Figure 7 Building a crawler algorithm model.

order to validate the impact of the model of the present topic network and the impact of the crawler strategy.

In Figure 7, in the beginning phase of the algorithm it is necessary to obtain the initial localization resources of the network, and then obtain the network page through the resources. Second, the current network cluster data is obtained by parsing the network page, and the network cluster data is judged to determine whether the current network page is changed or not. If it is changed, then the topic judgment model is used to generate the positioning resources of the sub-stage. If it does not change, then get the behavioral resources of the same level through the action recognition of the network, merge the sub-stage and the localization resources of the same level and add them to the crawler to be retrieved module, and then judge whether the localization resources are empty or not. If it is empty, the crawler nodes are parsed separately and then clustered. If it is not empty, the current crawler retrieval results are output directly, thus realizing the whole crawler retrieval.

3 Results

3.1 Comparative Analysis of the Retrieval Effect of the Internet Worm

To test the current operation effect of the IW, the Baidu network is used as the training network of the crawler, and the first stage localization resources of the IW are set to 40. The crawler will extract the localization resources of the second stage when running, and get 500 localization resources that need to be retrieved by the crawler, and set the current threshold to 5, and the crawler will flip the web page 10 times each time. The node of the crawler is 3. The operating system is Windows 10, the utterance analysis tool is Pyhanlp, the participle tool is Jieba, and the hardware is Intel(R) Core(TM) i7-6700. The data model selects the current news, education, and other neighboring data information. Moreover, the total amount of data is 50,000, and 7000 data are selected as the data of the training set. The comparison model selects the TextRank algorithm, the digital filtering algorithm (DFA), and the traditional TF-IDF algorithm for algorithm comparison, and the research strategy compares the priority strategy, breadth strategy, depth strategy, and deep learning search strategy for comparison. Figure 8 shows the comparison of the change in accuracy under different algorithms using the current crawler.

The accuracy of the four algorithms in the dataset varies less in Figure 8(a) as the number of text extractions increases, while the research

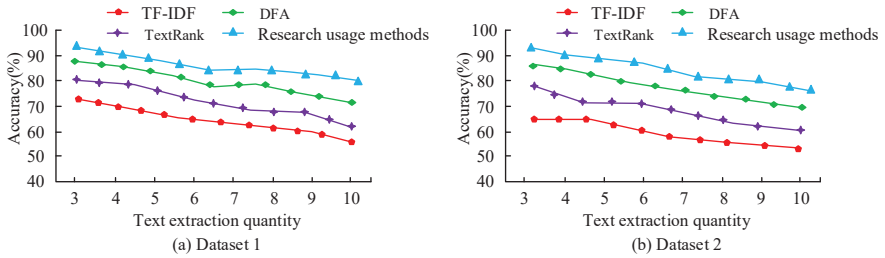


Figure 8 Comparison of data extraction accuracy of different algorithms.

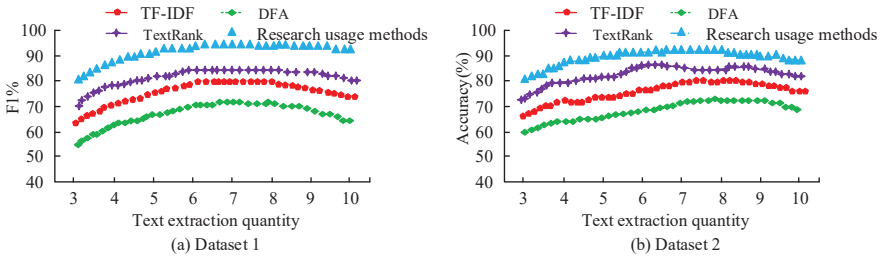


Figure 9 Comparison of F1 values for different algorithm models.

use algorithm model's average accuracy is roughly 86.5%. The lowest accuracy among the four algorithms is the traditional TF-IDF algorithm with an average value of about 64.3%, which is about 22.2% lower compared to the research use method of its algorithmic accuracy. In Figure 8(b), the accuracy of the four algorithms varies in the same way as in Figure 8(a), but the average value of the accuracy of the research use algorithm model averages 85.6%. The average value of the traditional TF-IDF algorithm is 58.4%, which is 27.2% lower compared to the research use method traditional TF-IDF algorithm. This demonstrates that, in comparison to other algorithmic models, the research usage algorithm has a greater DE capability. The comparison of the F1 values from several algorithmic models is displayed in Figure 9.

In Figure 9(a), the F1 values of the four algorithms increase with the number of keywords and then show a decreasing trend, which may be due to the algorithmic model reaching the optimum first and then experiencing a performance degradation. However, the F1 value of the model used in the study is at a higher value among the four algorithmic models, with the highest F1 value of 93.5%, which is 25.8% higher compared to the lowest DFA model of 67.7%. The trend of Figure 9(a) is mostly maintained by the shift in the F1 values of the four models in Figure 9(b). The research use algorithm

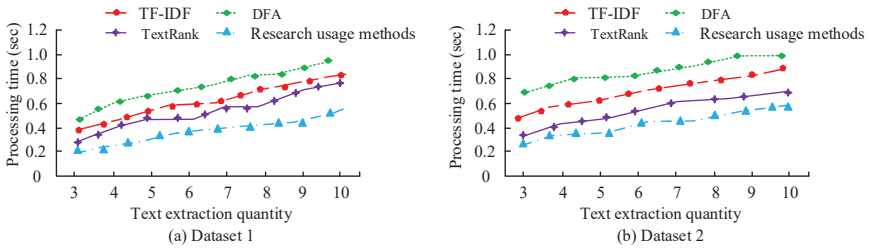


Figure 10 Comparison of text processing times.

Table 1 Comparison of the effect of running the crawler with different strategies

Thematic Strategy	Number of Resources Positioned	Site Coverage	Target Number	Ratio of Theme Pages	Running Time
Breadth first strategy	11700	80%	1582	13%	6.1 h
Deep prioritization strategy	20280	81%	1687	9%	11.4 h
Best priority search	4944	69%	1541	25%	5.2 h
Deep learning-based strategy	3700	74%	1455	39%	1.4 h
Research use strategy	2800	81%	1465	49%	1.0 h

model has the highest F1 value at 91.6%, which is approximately 26.2% higher than the DFA model's 65.4%. This demonstrates that the research use algorithm model performs better than the conventional use algorithm model. Comparison of the processing time of the text is obtained as shown in Figure 10.

In Figure 10(a), the text processing time of several algorithmic models increases with the increase of text data, but the model used in the study uses the shorter time used in the text processing time, in which the processing time is 0.54 s when the text data is 10, and the DFA model has the longest time for text processing of 0.87 s, and there is a difference between the two models of 0.33 s. In Figure 10(b), the change in the text processing time of the model is the same as in Figure 10(a), while the research use model has the shortest text processing time, in which the processing time when the number of texts is 10 has a difference of 0.50 s with the highest DFA model. This shows that the research use model, with a shorter processing time for text data processing, is more effective. The comparative analysis of the various ways for the crawler's operation is presented in Table 1.

In Table 1, the total number of localization resources of the research use strategy among several strategies is 2800, which indicates that the current use strategy does not require more localization resources to be able to complete

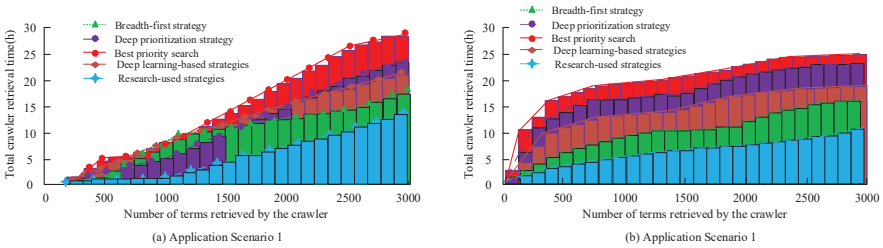


Figure 11 Comparison of total running time of different strategy crawlers.

the crawler retrieval process. At the same time the network coverage of the research use strategy is 81%, which is 12% higher compared to the optimal strategy with the lowest network coverage. The number of retrieval targets of the research using strategy method in the middle level is as large as 1455, which is 222 targets lower compared to the highest number of the depth-first strategy. However, the percentage of topic pages for the research use strategy is 49%, which is 40% higher compared to the lowest depth-first strategy. Finally, in the comparison of strategy retrieval time, the retrieval time of the research use strategy is 1.0 h compared to the highest breadth optimized strategy, which is 5.1 h lower. These results display that the research use strategy has a better operational effect compared to the other strategies in the overall operational effect of the strategies. The comparison of the crawler's overall running time with various techniques is displayed in Figure 11.

In Figure 11(a), the total time variation of crawler retrieval with different strategies increases with the increase in the number of retrievals in test scenario 1. The total running time of the crawler is shorter and also the time variation is less when the number of retrievals is around 1000, and the total time of retrieval increases more after the number of retrievals exceeds 1000. At the same time the total time of the strategy used in the study reaches 14.5 h at a number of 6000 and the total time of the optimal strategy varies more with a total time of 28.2 h. In Figure 11(b) the variation of retrieval time of the crawler with different strategies in test scenario 2 increases with the increase in the number of files and then the variation is less. At the same time the research use strategy time is smaller, the average retrieval time is 7.2 h, the optimal strategy retrieval time is longer the average retrieval time at 17.6 h, and the difference between the two strategies is 10.4 h. It can be noticed that compared with the optimal strategy the research use strategy the total time is lower, which suggests that the research use strategy can effectively reduce the IW running time in different strategies.

3.2 Actual Operation Effect of the Internet Worm

We compared the actual operation of different methods in the previous section, and use different algorithm models for crawling the keywords "Vaccine Passport", "Cross Border" and "Pneumonia" as an example. The combinations of research models are the TF-IDF algorithm, the model adding lexical features, the model adding positional features, the model adding semantic features, and the research use model, and the combinations are denoted as combinations 1, 2, 3, 4, and research use model, respectively. The retrieval effect is compared to get the keyword retrieval results of the different algorithm models, as shown in Table 2.

When it comes to accuracy comparison, Table 2 shows that the study's model outperforms the other algorithmic models by a large margin. Notably, the model achieves the highest accuracy of 93% when it comes to the detection of the keyword "pneumonia." At the same time in the detection of this keyword the retrieval time is significantly shorter, only 1.9 s and the number of keywords retrieved is significantly more at 343. Compared to other retrieval algorithms, the effect and performance of the retrieval algorithm used in the study is significantly higher than other algorithmic models. This shows that the algorithmic model used in the study is more effective for keyword extraction. The operational performance of different combinations is compared, as shown in Figure 12.

In Figure 12(a), the performance of all three models of the current research use model combination is relatively high, where the retrieval accuracy variation is 82% for the research use model, which is 19% higher compared to the lowest combination 1 accuracy. In Figure 12(b), the highest recall value of 72 for the research use model is 20% higher compared to the lowest combination 1. In Figure 12(c), the research uses the model with

Table 2 Keyword retrieval results for different algorithmic models

Manual Annotation	Search Keywords								
	Vaccine Passport			Cross Border			Pneumonia		
Performance	Accuracy (%)	Time (s)	Total Number of Keywords	Accuracy (%)	Time (s)	Total Number of Keywords	Accuracy (%)	Time (s)	Total Number of Keywords
TF-TDF	86	3.5	130	88	3.4	140	90	3.2	142
TextRank	84	4.2	210	86	41	252	83	4.5	236
DFA	83	3.6	213	84	3.7	234	86	3.8	224
Research use model	91	2.1	321	92	2.1	324	93	1.9	343

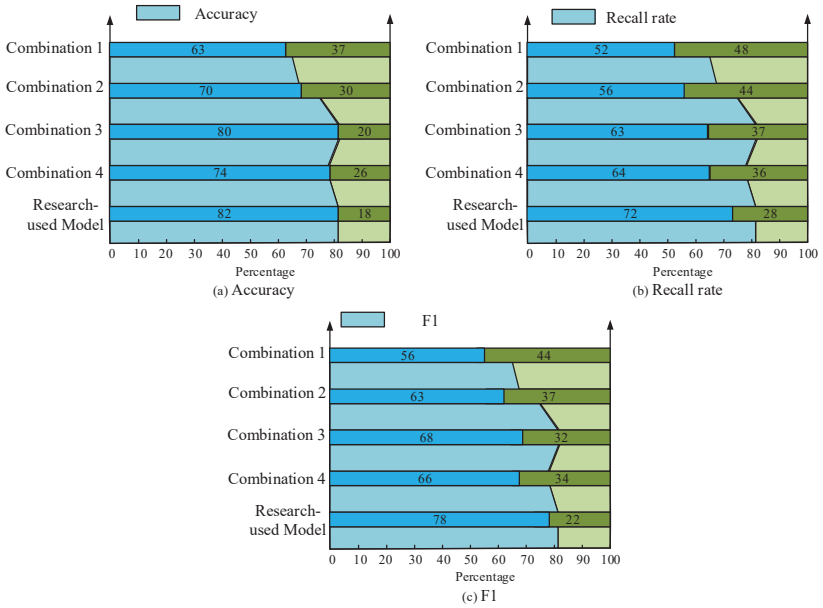


Figure 12 Different combinations of crawler retrieval effect.

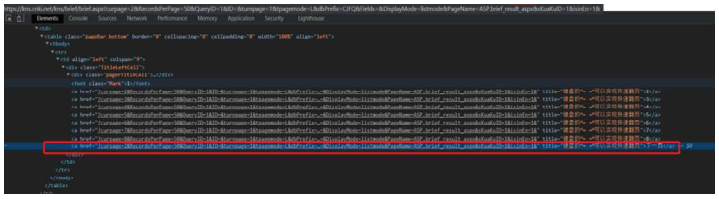
Table 3 Comparison of the effect of different algorithms for theme judgment

Algorithmic Model	Completeness Rate	Accuracy
Keyword matching	40%	58%
Multi pattern matching	55%	72%
K-nearest neighbor + TextRank	60%	80%
Naive Bayes	58%	60%
Topic recognition algorithm for deep learning	85%	75%
TF-IDF	76%	68%
Research use model	88%	82%

the highest F1 value of 78%, which is 22% higher compared to the lowest combination 1. This shows that the model used in the current research is more efficient and performs better than the other retrieval combinations of the crawler in the process of crawler retrieval. Table 3 presents a comparative analysis of the efficacy of several subject judgment algorithms.

In Table 3, among several common topic judgment models, the model used in the study has the best algorithmic results in terms of the change in the check rate and accuracy. Comparing several traditional algorithms, the highest check rate of the algorithmic model used in the study is 88%, which

(a) The process of crawling and retrieving articles



(b) Completion status of article keyword search

Figure 13 Display of the crawler retrieval effect.

is 48% higher than the 40% of the keyword extraction model, which may be the reason why the algorithmic model improves the keyword extraction of the data. Meanwhile, the research use model has the highest accuracy value of 82%; compared to the lowest keyword extraction model, its accuracy is improved by 24%, which may be caused by the addition of the Word2Vec model. This shows that the research use model has better model keyword retrieval. Figure 13 shows the actual situation of the current model crawler retrieving Web documents.

Figure 13(a) shows the current crawler retrieval of articles, i.e. the selection of articles with the keyword “Vaccine Passport” in the article content of the web page. During the crawler retrieval process, the articles are first imported as a whole, and then preprocessed after the data import is completed. Finally, the keywords are retrieved from the crawler. In Figure 13(b), the keyword extraction process entails the retrieval of keywords from the

article by the crawler. This retrieval process enables the extraction of data information containing keywords from the article and its subsequent storage in the database, thereby finalizing the retrieval of the article.

4 Discussion

The study primarily addressed the challenges that the existing IW faced in acquiring and gathering web page data and evaluating web page themes. It also suggested a novel approach that made use of the TF-IDF algorithm and Word2Vec feature extraction. The new method transformed and analyzed the structure of thematic web pages using the TF-IDF algorithm and Word2Vec feature extraction for the IW, which improved the theme judgment and vocabulary retrieval ability of the crawler network. The outcomes revealed that the accuracy of the research use model was higher than the traditional TF-IDF algorithm by 22.2% and 27.2%, respectively. The F1 value of the model used in the study was the highest among the four algorithmic models at 93.5%, compared to the DFA whose F1 values were 25.8% and 26.2% higher, respectively. The research use algorithmic model had better processing results in terms of text processing time. The total number of localization resources for the research using strategy was 2800 and network coverage was 81% which was 12% higher than the optimal strategy. The retrieval time of 1.0 h for the research use strategy was 5.1 h lower than the highest breadth optimal strategy. The research used the shortest running time of only 14.5 h in the running time of the strategy. The research used the algorithmic model to be able to recognize the vocabulary of the keywords. The model used in the study had a maximum retrieval accuracy of 82%, which was 19% higher compared to other models. The recall rate was also 20% higher than the other models. The research used model had the highest search rate of 88%, which was 48% higher compared to other models. It can be concluded that the research using algorithmic models and strategies are able to improve the performance and retrieval ability of the current crawler retrieval. Although the research has achieved a lot of results, there are still some shortcomings. The model used in the study needs further data analysis on crawler sharing, and also needs to conduct model training with higher quality documents to improve the model's topic judgment ability. Future research directions will also combine advanced deep learning models such as BERT and GPT to improve semantic understanding, integrate multimodal data to improve Web content understanding, optimize real-time data processing to adapt to dynamic network content, extend cross-domain and cross-language applications, and develop

user-centered retrieval strategies. It can also be combined with other algorithms and models such as TextRank, LDA, or transfer learning to further improve performance.

Declarations

Availability of Data and Material

The data will be made available on the request.

Competing Interests

The author claims no conflict of interest.

Authors' Contributions

The paper was written by Xinyue Feng.

Acknowledgments

The research was supported by: Guangdong Higher Vocational Education Innovation and Entrepreneurship Training Program Project, Smart Wearable Device for Pets Based on ARM and Big Data (No. 20220830); Guangdong Science and Technology Innovation Strategy Special Fund Project (No. pdjh2022b1000); Project of Young Innovative Talents in General Universities of Guangdong Province (No. 2022KQNCX243); Guangdong Province Ordinary Higher Education Engineering Technology Research (Development) Center, Industrial Internet Enables Data Protection Technology Engineering Research Center (No. 2024GCZX028); Key Fields Special Projects of Colleges & Universities in Guangdong Province, Research on the Intelligent Aquaculture Industry Chain Service Platform Based on the High-Quality Development Project for Hundreds of Counties, Thousands of Towns, and Tens of Thousands of Villages (No. 2024ZDZX4164).

References

[1] K. Manjari, R. Sumanth, S. Rousha, and J. Devi. "Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm," Proc. Int. Conf. Trends Electron. Inform. (ICOEI), vol. 15, no. 2, pp. 648–652, April, 2020, DOI: 10.1109/ICOEI48184.2020.9142938.

[2] A. Jalilifard, V. F. Caridá, A. F. Mansano, R. S. Cristo, and F. P. da Fonseca. "Semantic sensitive TF-IDF to determine word relevance in documents," Adv. Comput. Netw. Commun., vol. 2021, no. 2, pp. 327–337, June, 2021, DOI: 10.1007/978-981-33-6987-0_27.

[3] F. Lan. "Research on text similarity measurement hybrid algorithm with term semantic information and TF-IDF method," Adv. Multimed., vol. 23, no. 5, pp. 2022–2023, April, 2022, DOI: 10.1155/2022/7923262.

[4] M. Suma and P. Madhumathy. "Brakerski-Gentry-Vaikuntanathan fully homomorphic encryption cryptography for privacy preserved data access in cloud assisted Internet of Things services using glow-worm swarm optimization," Trans. Emerg. Telecommun. Tech., vol. 33, no. 12, pp. 4641–4642, December, 2022, DOI: 10.1002/ett.4641.

[5] M. Aqeel, F. Ali, M. W. Iqbal, T. A. Rana, M. Arif, and M. R. Auwul. "A review of security and privacy concerns in the internet of things (IoT)," J. Sensors, vol. 2022, no. 10, pp. 29–30, September, 2022, DOI: 10.1155/2022/5724168.

[6] Y. Deng, Y. Pei, and C. Li. "Parameter estimation of a susceptible–infected–recovered–dead computer worm model," Simul., vol. 98, no. 3, pp. 209–220, March, 2022, DOI: 10.1177/00375497211009576.

[7] A. R. Lubis, M. K. Nasution, O. S. Sitompul, and E. M. Zamzami. "The effect of the TF-IDF algorithm in times series in forecasting word on social media," Indones. J. Electr. Eng. Comput. Sci., vol. 22, no. 2, pp. 976–984, February, 2021, DOI: 10.11591/ijeecs.v22.i2.pp976-984.

[8] L. Cheng, Y. Yang, K. Zhao, and Z. Gao. "Research and improvement of TF-IDF algorithm based on information theory," Proc. Int. Conf. Comput. Eng. Netw. (CENet), vol. 13, no. 6, pp. 608–616, April, 2020, DOI: 10.1007/978-3-030-14680-1_67.

[9] X. Ao, X. Yu, D. Liu, and H. Tian. "News keywords extraction algorithm based on TextRank and classified TF-IDF," Proc. Int. Wireless Commun. Mob. Comput. (IWCMC), vol. 15, no. 6, pp. 1364–1369, June, 2020, DOI: 10.1109/IWCMC48107.2020.9148491.

[10] W. Zhuohao, W. Dong, and L. Qing. "Keyword Extraction from Scientific Research Projects Based on SRP-TF-IDF," Chin. J. Electron., vol. 30, no. 4, pp. 652–657, April, 2021, DOI: 10.1049/cje.2021.05.007.

[11] R. Rawat, V. Mahor, S. Chirgaiya, R. N. Shaw, and A. Ghosh. "Analysis of darknet traffic for criminal activities detection using TF-IDF and light gradient boosted machine learning algorithm," Innov. Electr. Electron.

Eng., vol. 2021, no. 5, pp. 671–681, May, 2021, DOI: 10.1007/978-981-16-0749-3_53.

[12] S. Rahman, K. H. Talukder, and S. K. Mithila. "An empirical study to detect cyberbullying with TF-IDF and machine learning algorithms," Proc. Int. Conf. Electron. Commun. Inf. Tech. (ICECIT), vol. 2021, no. 14, pp. 1–4, September, 2021, DOI: 10.1109/ICECIT54077.2021.9641251.

[13] H. Yu, Y. Ji, and Q. Li. "Student sentiment classification model based on GRU neural network and TF-IDF algorithm," J. Intell. Fuzzy Syst., vol. 40, no. 2, pp. 2301–2311, February, 2021, DOI: 10.3233/JIFS-189227.

[14] Z. Jiang, B. Gao, Y. He, Y. Han, P. Doyle, and Q. Zhu. "Text classification using novel term weighting scheme-based improved TF-IDF for internet media reports," Math. Probl. Eng., vol. 2021, no. 4, pp. 1–30, Mar, 2021, DOI: 10.1155/2021/6619088.

[15] T. Korkmaz, A. Çetinkaya, H. Aydın, and M. A. Barışkan. "Analysis of whether news on the Internet is real or fake by using deep learning methods and the TF-IDF algorithm," Int. Adv. Res. Eng. J., vol. 5, no. 1, pp. 31–41, April, 2021, DOI: 10.35860/iarej.779019.

[16] M. Mohammed and N. Omar. "Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec," PLoS ONE, vol. 15, no. 3, pp. 19–20, March, 2020, DOI: 10.1371/journal.pone.0230442.

[17] V. D. Antonio, S. Efendi, and H. Mawengkang. "Sentiment analysis for covid-19 in Indonesia on Twitter with TF-IDF featured extraction and stochastic gradient descent," Int. J. Nonlinear Anal. Appl., vol. 13, no. 1, pp. 1367–1373, January, 2022, DOI: 10.22075/IJNAA.2021.5735.

[18] G. Yunanda, D. Nurjanah, and S. Meliana. "Recommendation system from microsoft news data using TF-IDF and cosine similarity methods," Build. Inform. Tech. Sci. (BITS), vol. 4, no. 1, pp. 277–284, Jun, 2022, DOI: 10.47065/bits.v4i1.1670.

[19] S. Amin, M. I. Uddin, S. Hassan, A. Khan, N. Nasser, A. Alharbi, and H. Alyami. "Recurrent neural networks with TF-IDF embedding technique for detection and classification in tweets of dengue disease," IEEE Access, vol. 8, no. 5, pp. 131522–131533, May, 2020, DOI: 10.1109/ACCESS.2020.3009058.

[20] Y. Li, and H. Ning. "Multi-feature keyword extraction method based on TF-IDF and Chinese grammar analysis," Proc. Int. Conf. Mach. Learn.

Intell. Syst. Eng. (MLISE), vol. 2021, no. 9, pp. 362–365, November, 2021, DOI: 10.1109/MLISE54096.2021.00075.

[21] J. Li. "A comparative study of keyword extraction algorithms for English texts," J. Intell. Syst., vol. 30, no. 1, pp. 808–815, Jul, 2021, DOI: 10.1515/jisys-2021-0040.

[22] J. Qin, Z. Zhou, Y. Tan, X., and Z. He. "A big data text coverless information hiding based on topic distribution and TF-IDF," Int. J. Digit. Crime Forensics, vol. 13, no. 4, pp. 40–56, Jul, 2021, DOI: 10.4018/IJDCF.20210701.oa4.

[23] I. Ghozali, M. F. Asy'ari, S. Triarjo, H. M. Ramadhani, H. Studiawan, and A. M. Shiddiqi. "A Novel SQL Injection Detection Using Bi-LSTM and TF-IDF," Proc. 7th Int. Conf. Inf. Netw. Technol. (ICINT), vol. 21, no. 6, pp. 16–22, May, 2022, DOI: 10.1109/ICINT55083.2022.00010.

[24] G. Di Gennaro, A. Buonanno, and F. A. Palmieri. "Considerations about learning Word2Vec," J. Supercomput., vol. 2021, no. 1, pp. 1–6, 2021, DOI: 10.1007/s11227-021-03743-2.

[25] D. E. Cahyani and I. Patasik. "Performance comparison of tf-idf and word2vec models for emotion text classification," Bull. Electr. Eng. Inform., vol. 10, no. 5, pp. 2780–2788, October, 2021, DOI: 10.11591/eei.v10i5.3157.

[26] B. Jang, M. Kim, G. Harerimana, S. U. Kang, and J. W. Kim. "Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism," Appl. Sci., vol. 10, no. 17, pp. 5841–5842, August, 2020, DOI: 10.3390/app10175841.

[27] S. Thavareesan and S. Mahesan. "Sentiment lexicon expansion using Word2vec and fastText for sentiment prediction in Tamil texts," Proc. Moratuwa Eng. Res. Conf. (MERCon), vol. 2020, no. 28, pp. 272–276, July, 2020, DOI: 10.1109/MERCon50084.2020.9185369.

[28] R. Kurnia, Y. Tangkuman, and A. Girsang. "Classification of user comment using word2vec and SVM classifier," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 1, pp. 643–648, February, 2020, DOI: 10.30534/ijatcse/2020/90912020.

[29] A. Mallik and S. Kumar. "Word2Vec and LSTM based deep learning technique for context-free fake news detection," Multimed. Tools Appl., vol. 83, no. 1, pp. 919–940, January, 2024, DOI: 10.1007/s11042-023-15364-3.

[30] P. Rakshit and A. Sarkar. "A supervised deep learning-based sentiment analysis by the implementation of Word2Vec and GloVe Embedding

techniques," Multimed. Tools Appl., vol. 202, no. 9, pp. 1–34, April, 2024, DOI: 10.1007/s11042-024-19045-7.

[31] P. Preethi and H. R. Mamatha, "Region-Based Convolutional Neural Network for Segmenting Text in Epigraphical Images," Artif. Intell. Appl., vol. 1, no. 2, pp. 119–127, September, 2023, DOI: 10.47852/bonviewAIA2202293.

Biography

Xinyue Feng is a lecturer and postgraduate. She received her Bachelor's degree in Computer Science and Technology from North University of China in 2012 and her Master's degree in Software Engineering from North University of China in 2015. From 2022 to now, she has been a doctoral candidate in Electrical and Computer Engineering at Maha Salakan University, Thailand, in the research direction of data mining and data analysis. From 2015 to now she has been a full-time teacher at Foshan Polytechnic. She has published 8 academic articles, participated in 6 scientific research projects, authorized 8 patents, published 3 textbooks, and published 2 other academic research and achievements.