Hindawi
Complexity
Volume 2021, Article ID 5529447, 11 pages
https://doi.org/10.1155/2021/5529447

Research Article

Web News Data Extraction Technology Based on Text Keywords

Kun Zhang

School of Communication, Xi'an Peihua University, Xi'an City, China

Correspondence should be addressed to Kun Zhang; zhangkun@peihua.edu.cn

Received 18 January 2021; Revised 2 February 2021; Accepted 1 April 2021; Published 16 April 2021

Academic Editor: Abd E.I.-Baset Hassanien

In order to shorten the time for users to query news on the Internet, this paper studies and designs a network news data extraction technology, which can obtain the main news information through the extraction of news text keywords. Firstly, the TF-IDF keyword extraction algorithm, TextRank keyword extraction algorithm, and LDA keyword extraction algorithm are analyzed to understand the keyword extraction process, and the TF-IDF algorithm is optimized by Zipf's law. By introducing the idea of model fusion, five schemes based on waterfall fusion and parallel combination fusion are designed, and the effects of the five schemes are verified by experiments. It is found that the designed extraction technology has a good effect on network news data extraction. News keyword extraction has a great application prospect, which can provide the basis for the research fields of news key phrases, news abstracts, and so on.

1. Introduction

With the development of Internet technology, online news is growing exponentially, and users can get the latest news at home and abroad in real time through mobile phones and other mobile terminals [1]. But in the Internet, the news content is uneven, there are a lot of induced content, lack of news authenticity, it is difficult for users to accurately find the actual needs of the content in the massive network news. At this time, a kind of network news data extraction technology based on text keywords is designed, which can help users accurately locate valuable news [2]. Keywords in text keywords refer to a kind of refined vocabulary to obtain information. The traditional manual keyword extraction work cannot meet the needs of big data news text at this stage, and automatic text keyword extraction technology is imperative [3]. Online news text automatic keyword extraction technology can save users' reading time and, at the same time, assist users to screen out junk news and quickly obtain news content [4]. In view of this, this research will develop a kind of network news text data extraction technology based on text keywords, so as to enhance the quality of news reading and save users' time.

In order to reduce the dependence of Web text keyword extraction on large annotated text corpus, Campos et al. designed an unsupervised automatic keyword extraction method to extract keywords from a single document through multiple local features [5]. Li and other scholars proposed to infer topic distribution bias labels by the topic model algorithm and construct a weighted graph by using random walk algorithm, introduce offset hash random walk on a weighted graph, and extract text keywords by combining labels [6]. Kolokas and other researchers designed a text keyword extraction method based on a recurrent neural network. The network model can map the text keyword sequence to the whole text and perform continuous re-representation of keywords, which has a good keyword extraction effect [7]. Onan and other scholars have investigated the most commonly used keyword extraction methods, such as the measure based keyword extraction method, word frequency-inverse sentence frequency based keyword extraction method, cooccurrence statistical information based keyword extraction method, eccentric keyword extraction method, and TextRank algorithm. After that, they proposed the combination of keyword based text document representation and ensemble learning, which can significantly improve the efficiency of keyword extraction and improve the prediction performance of the text classification scheme to extend the performance [8]. With the continuous improvement of the level of science and technology, the number of on-board

applications such as car navigation has increased, and the amount of social data of cars has increased sharply. Cloud based vehicle data processing methods have emerged. In order to protect the data from being accessed, Yang et al. proposed a keyword extraction measurement method based on specific text spatial distribution, which requests access through the extracted keyword index, so as to achieve data encryption protection [9].

The progress of digital technology has fundamentally affected the society. Hofmann and other researchers integrate text data from different information sources through text mining technology, process it into an analytical and readable relationship network between technologies, and then study the dynamic system of related technologies [10]. Sapozhnikova and other scholars use the convolutional neural network to classify the news information text of Internet information portal and realize the semantic preprocessing of the text through the open word2vec model. In the network news data extraction, the classification accuracy reaches 84% [11]. With the popularity of social media, users' travel preferences can be obtained from users' historical mobile records on social media. Wen et al. designed a representative travel path framework based on keyword perception, extracted knowledge from users' historical mobile records, and successfully completed the travel route recommendation experiment [12]. Ranjan and Prasad used semantic keywords and BPlion neural network algorithm to automatically classify text. Through experiments on data sets, the results show that the accuracy of text classification can reach 90.9% [13]. In the era of big data, with the rapid increase of network digital resources, short text resources show great vitality. Wang and his team analyzed the classification of Chinese short text under low granularity features (keywords) by comparing the classification ability of different Chinese fragments [14].

These methods have made great progress in keyword extraction, text classification, retrieval, and so on, but there is a lack of relevant research on network news data extraction technology. Although many researchers have designed different types of text keyword extraction methods, the method has strong pertinence and cannot be directly used for news data extraction. In order to save time for users to get the news, this paper designs a kind of network news data extraction technology based on text keywords on the basis of considering the characteristics of network news.

2. Design of Network News Data Extraction Technology

2.1. Keyword Extraction Scheme of Network News Text. With the popularity of search engines and social networks, the way people get information has changed, and the Internet has become an important position for information sharing. Major news portals have launched news mobile clients, resulting in a surge of online news data [15].

As can be seen from Figure 1, the number of mobile Internet news users has increased from 366.51 million in 2013 to 660.2 million in 2019; as of June 2019, the number of Internet news users in China has reached 686 million, an

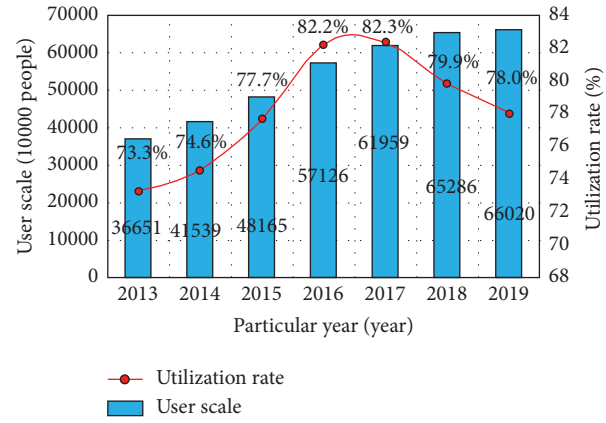


Figure 1: Scale and utilization rate of mobile Internet news users from 2013 to the first half of 2019.

increase of 11.14 million compared with the end of 2018. Through the analysis of user behavior logs, infer user reading preferences, and then push network news to different users so that users have stickiness to the news client. How to accurately extract network news data and achieve "precise and accurate" news push is a powerful tool to save users' time, improve users' reading quality, and improve users' stickiness to news clients [16].

Network news usually focuses on reporting some social events. Generally, only a few keywords are needed to let users understand the main content of the news. Therefore, the extraction of network news data can be summarized as the extraction of keywords in the network news text [17]. Keyword extraction methods are divided into supervised extraction and unsupervised extraction according to whether training samples are needed. This paper mainly studies unsupervised extraction methods, including term frequency–inverse document frequency (TF-IDF) extraction algorithm, TextRank algorithm, and LDA (late Dirichlet allocation) topic model algorithm [18]:

$$\mathrm{TF}_{ij} = \frac{n_{ij}}{N_j}. \tag{1}$$

Formula (1) is the calculation formula of term frequency (TF), where i, j refer to the word and the text corresponding to the word respectively; n_{ij} refers to the number of times the word I appears in text J; the total number of words in text J is represented by N_j:

$$\mathrm{IDF}_i = \log \frac{N}{n_i + 1}. \tag{2}$$

Formula (2) is the calculation formula of inverse document frequency (IDF), the total number of texts is N, and the total number of texts containing word I in the corpus is n_i:

$$\mathrm{TF} - \mathrm{IDF}_i = \mathrm{TF}_{ij} \times \mathrm{IDF}_i. \tag{3}$$

Formula (3) is the calculation formula of TF-IDF. It can be seen that the larger the TF-IDF value of a word i is, the more likely it is to be a keyword of text j. TextRank algorithm

divides the text into several constituent words and constructs the word graph model. Take the automobile network news as an example; see Figure 2 for details.

According to the network diagram shown in Figure 2, the degree of connection between words is explored, the words are scored, and the keywords are obtained by ranking the scores. Set the constructed word graph model as $G = (V, E)$, which is a set of vertices and edges, so the set of all vertices and the set of all edges on the network graph G are represented by V and E in turn. The set of vertices that any vertex v_i points to is $\mathrm{In}(v_i)$, the set that points to other points is $\mathrm{Out}(v_i)$, and $(v_i, v_j) \in E$:

$$H(v_i) = (1-d) + d * \sum_{j \in In(v_i)} \frac{1}{|\mathrm{Out}(v_i)|} H(v_j). \quad (4)$$

Equation (4) is the scoring formula of the vertex v_i of the weighted graph, d is the damping coefficient, and d is 0.85:

$$H(v_i) = (1-d) + d * \sum_{j \in \mathrm{In}(v_i)} \frac{w_{ji}}{\sum_{v_k \in \mathrm{Out}(v_i)} w_{jk}} H(v_j). \quad (5)$$

Equation (5) is the scoring formula when there is a fixed weight between two vertices, where w_{ji} is the weight between v_i and v_j, d is the damping coefficient, and d is 0.85. LDA topic model algorithm combines words and documents which are not directly related by topic to fit the distribution of word text topic.

The original TF-IDF algorithm has the disadvantages of low extraction efficiency and poor extraction accuracy. This paper proposes to introduce Zipf's law and chi-square test to improve the original TF-IDF algorithm, in which Zipf's law is responsible for obtaining weights of different frequencies, and chi-square test is used for keyword extraction. When there are m words in the longer text j, the words with more times appear in the first place and the words with fewer times appear in the second place. The words are ranked by the rank number (word rank) r. When the number of words is n_m, there is $n_m \times r = C$, and C is a constant fluctuating around a fixed value. Most of the online news is in the form of short text, and the frequency of the same word in a single text j is not more than 5 times. The same frequency words are sorted by the maximum method [19, 20]:

$$I_n = r_n - r_{n+1}. \quad (6)$$

Formula (6) is the formula for calculating the number of words I_n with the same frequency, and the value of r_n is the word rank. When the number of words $n_m \le 5$, there is $I_n = (D/n) - (D/(n+1)) = (D/n(n+1)), \quad n \le 5,$ where $D = r_n \times n$, n are word frequency. The proportion of each word frequency in the same text can be counted by I_n/M, and $I_n/M = 1/[n(n+1)]$ and m are approximate constants of the product of word rank and n_m.

It can be seen from Table 1 that, with the increase of n, the value of I_n/M decreases. Because the importance of low-frequency words in short news texts is very low, when extracting network news data, we can first judge whether the word frequency of each word is greater than 1, and if so, calculate the IDF value:

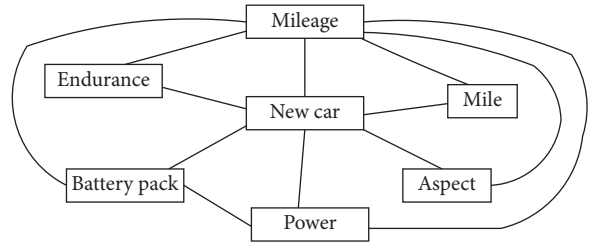


Figure 2: Representation of text graph model.

Table 1: The relation table of I_n/M value with n value.

n	1	2	3	4	5
I_n/M	1/2	1/6	1/12	1/20	1/30

$$\chi^2 = \sum \frac{(A-E)^2}{E}. \quad (7)$$

Equation (7) is the expression of the chi square test χ^2. When χ^2 (degree of deviation) is very small, it is judged as error, where A and E refer to the actual value and the theoretical value, respectively:

$$\mathrm{TF-IDF}-K = \mathrm{TF} \times \mathrm{IDF} \times \log K. \quad (8)$$

Equation (8) is based on the principle of Zipf's law and the chi-square test (tf-idf-k), and there is a chi-square value $K = \chi^2$.

As shown in Figure 3, first preprocess the news text through Jieba word segmentation, then filter the stop words, and count the number of times each word appears, remove the words with word frequency of 1, calculate the TF-IDF value and chi-square value K, multiply to obtain the tf-idf-k value, and arrange them in descending order. The top words are the output results, that is, the text keywords.

2.2. Network News Data Extraction Scheme Based on Model Fusion. Model fusion can significantly improve the accuracy of network news data. Two model fusion schemes are proposed: the first is waterfall fusion and the second is parallel combination fusion.

As shown in Figure 4, waterfall fusion is in the form of cascading multiple algorithm models, using different algorithms for filtering, so as to get the final result. In the process of waterfall fusion, the previous activity is taken as the input, the result filtered by the previous algorithm is taken as the input filtered by the next algorithm, and the candidate results are continuously screened, so as to obtain the final result with less quantity and high quality [21].

As can be seen from Figure 5, parallel combination fusion extracts keywords from the original document through several groups of algorithms and then scores the keywords in the way of parallel voting, so as to select the optimal result. The three extraction algorithms described in Section 2.1 all have defects. TF-IDF relies heavily on corpus and its accuracy is affected by IDF; the TextRank algorithm has too high computational complexity; and the LDA

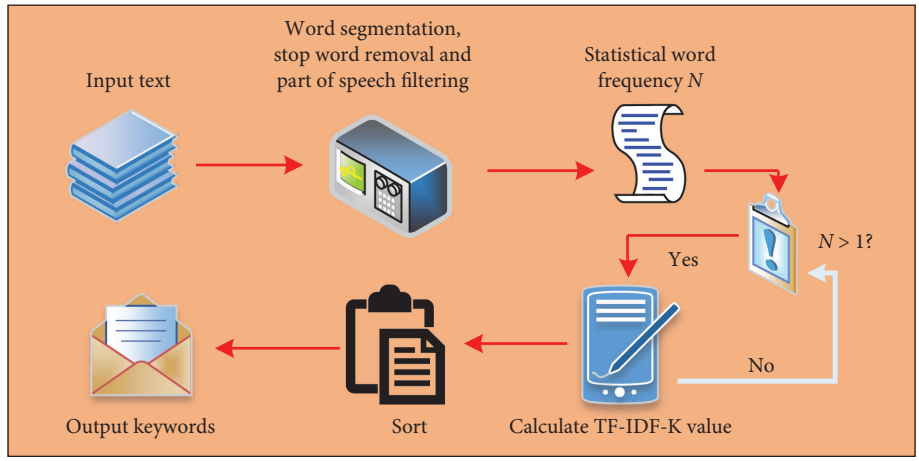


Figure 3: Algorithm flow chart.

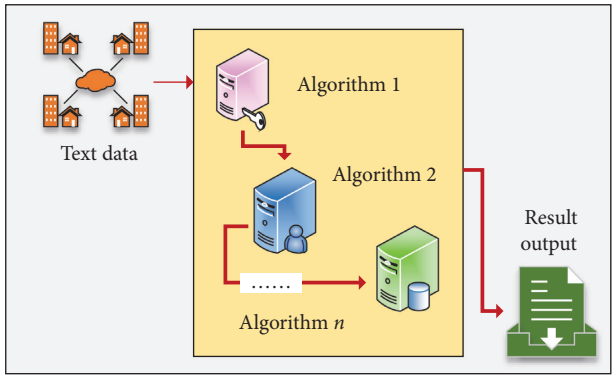


Figure 4: Waterfall fusion flow chart.

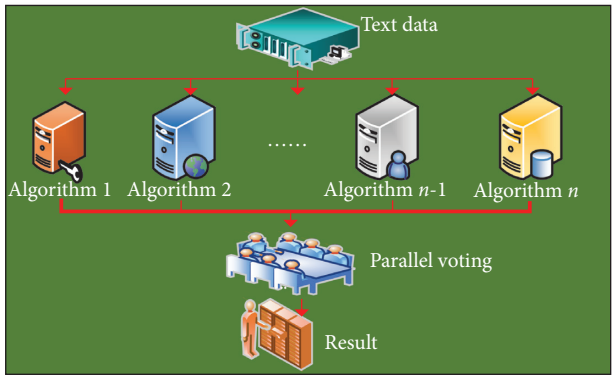


Figure 5: Flow chart of parallel combination fusion.

algorithm cannot feed back accurate document topics. Combined with the characteristics of different fusion models, two kinds of waterfall fusion network news keyword extraction schemes are proposed.

Figure 6(a) is Scheme 1: first, TF-IDF algorithm, and then TextRank algorithm. After word segmentation and stop words removal, the part of speech is filtered and the word frequency n is counted. The TF-IDF value is calculated and sorted according to the TF-IDF value of each word. When the number of times is greater than 50, reorder; when the number of times is not greater than 50, calculate the TextRank value, sort, and output keywords according to the serial number of each word. Figure 6(b) is Scheme 2: first, TF-IDF algorithm, and then LDA topic model algorithm.

The three design schemes in Figure 7 do not consider the sequence. Scheme 1 in Figure 7(a) is the parallel combination of the TF-IDF algorithm and the TextRank algorithm; Scheme 2 in Figure 7(b) is the parallel combination of the TF-IDF algorithm and LDA Algorithm; Scheme 3 in Figure 7(c) is the parallel combination of the LDA algorithm and the TextRank algorithm. The general process of the three schemes can be summarized as follows: input the network news text, process the text by word segmentation, stop words removal, part of speech filtering, and count the word frequency n. The two parallel algorithms sort the words at the same time, list the candidate keywords, output the final keywords by using the parallel fusion method, and complete the key information extraction of network news [22]. In this study, the accuracy rate, recall rate, and F1 value are used to evaluate the effect of online news keyword extraction. The expression of accuracy is shown in

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{9}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{10}$$

$$\begin{aligned} F1 &= \frac{1}{(((1/\text{precision}) + (1/\text{recall}))/2)} \\ &= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}. \end{aligned} \tag{11}$$

Formulae (10) and (11) are the expressions of recall rate and F1 value in turn, where TP refers to the situation where the label is a positive sample and the prediction is a positive sample; FN refers to the situation where the label is a positive sample but the prediction is a negative sample; FP refers to the situation where the label is a negative sample but the prediction is a positive sample; TN refers to the situation

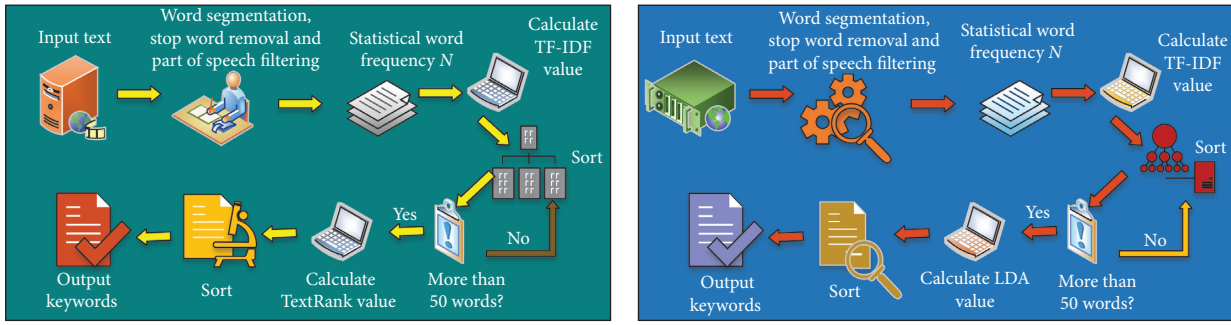


FIGURE 6: Two methods of keyword extraction for online news based on waterfall fusion. (a) Scheme 1 flow chart. (b) Scheme 2 flow chart.

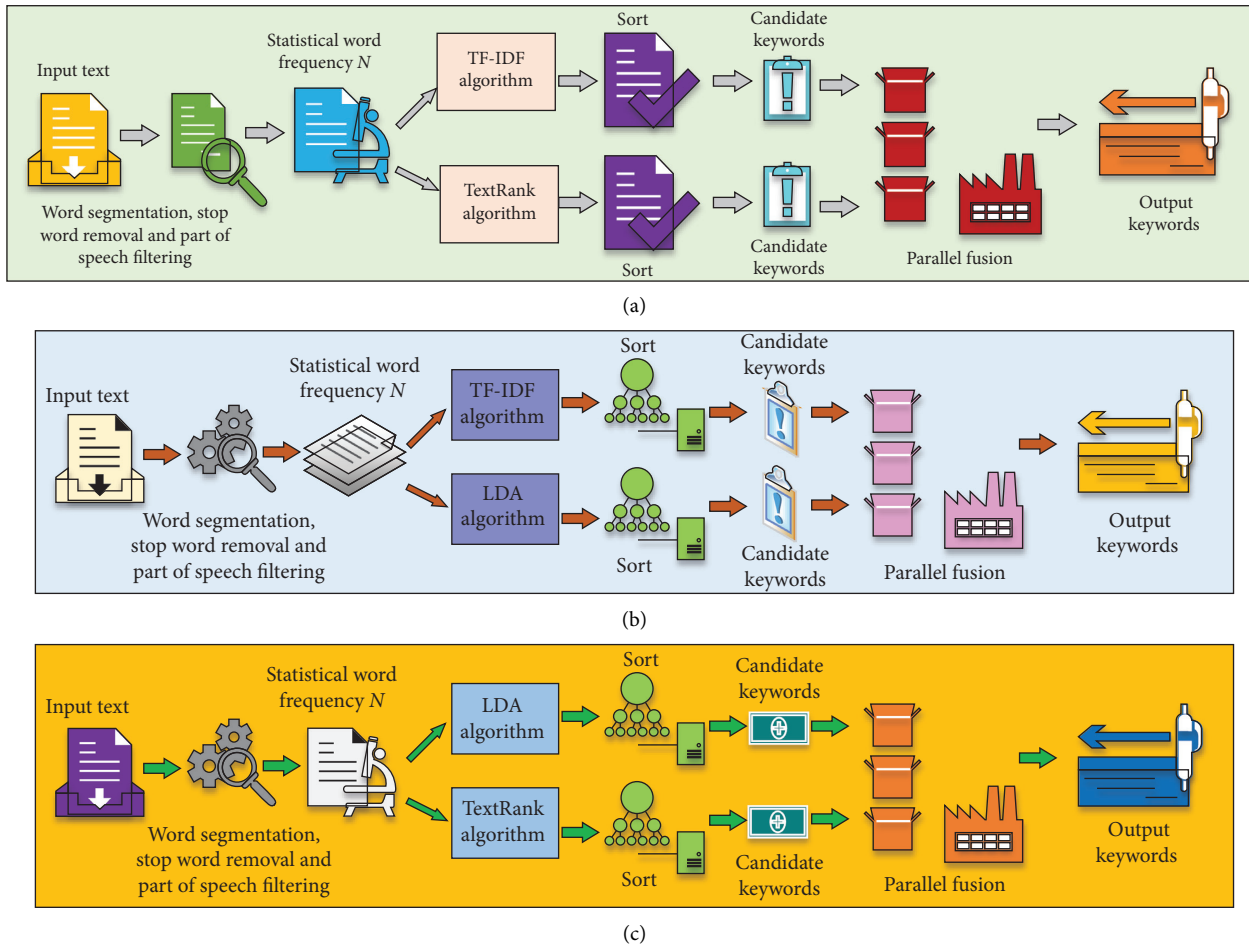


FIGURE 7: Design of three kinds of parallel fusion network news keyword extraction scheme. (a) Scheme 1 flow chart. (b) Scheme 2 flow chart. (c) Scheme 3 flow chart.

where the label is a negative sample and the prediction is a negative sample.

3. Application Effect and Discussion of Network News Data Extraction Technology

3.1. Practical Application Effect of Network News Data Extraction Technology. In order to verify the application effect of the network news data extraction technology proposed in this study, the next stage is the experimental analysis of different schemes. In the experiment, windows 10 system is selected as the experimental operating system, i7 processor is used, and memory is 16g; pychar + python3.6 is selected as the development tool; and 100 network news of ten categories are selected to carry out the experiment.

In the waterfall fusion experiment design, Scheme 1 uses TF-IDF algorithm to extract keywords through the TextRank algorithm; Scheme 2 also uses the TF-IDF algorithm to extract keywords through LDA topic model. The specific experimental results are shown in Table 2.

In Table 2, "Several Private Kindergartens in South Korea Are Involved in Corruption: Embezzling Operating Expenses to Buy Valuable Jewelry" is selected as the extraction object of online news. And through the two schemes for news important information extraction, it can be seen that compared with the standard keywords, it is obvious that the network news keywords extracted by Scheme 1 are closer to the gold standard than those extracted by Scheme 2; that is to say, Scheme 1 has better network news information extraction performance. Select 100 network news texts of ten categories, detect all network news keywords through Scheme 1 and Scheme 2, respectively, and record the time consumed by the two schemes when extracting the key information of the same network news, as shown in Figure 8.

Figure 8 shows the time taken by the two schemes to extract different types of network news keywords. From the analysis of the time taken to extract the key information (keywords) of 10 kinds of network news, the overall time taken to extract keywords in Scheme 1 is much less than that in Scheme 2; the average time taken to extract keywords in Scheme 1 is 43.87 s, and that in Scheme 2 is 138.74 s. The above results show that Scheme 1 has obvious advantages in the time consumption of key information extraction of network news. Ten categories of 100 online news are selected as the experimental objects to compare the recall and accuracy of the two waterfall fusion algorithms. The specific results are shown in Figure 9.

Figure 9(a) shows the recall comparison results of the two waterfall fusion algorithms. On the whole, the recall rate of Scheme 1 is higher than that of Scheme 2; the average recall rate of Scheme 1 is 0.47, and that of Scheme 2 is 0.34, which is 0.13 lower than that of Scheme 1. Figure 9(b) shows the accuracy comparison results of the two waterfall fusion algorithms. On the whole, the accuracy of Scheme 1 is higher than that of Scheme 2; the average accuracies of Scheme 1 and Scheme 2 are 0.38 and 0.31, respectively, and the average accuracy of Scheme 1 is 0.07 higher than that of Scheme 2. The above structure shows that the key information (keywords) acquisition performance of Scheme 1 is better than that of Scheme 2.

In Table 3, online news titled "Japanese College Students' Gathering Led to the Collapse of the Apartment Floor and 30 People Injured" is selected as the experimental object, and three different parallel combination fusion algorithms are adopted to extract the key information of the news. Compared with standard keywords, Scheme 1 (TF-IDF algorithm and TextRank algorithm combined in parallel) is better than Scheme 2 and Scheme 3 in extracting key information of online news. Then, three different parallel combination fusion schemes are used to extract keywords from ten groups of different categories of 100 online news, the specific time consumed by different schemes in extracting key information of different categories of online news is counted, and the average value is calculated. See Figure 10 for details.

Figure 10 shows that the average time consumption of Scheme 2 (parallel combination of LDA topic model algorithm and TF-IDF algorithm) in extracting key information of online news is 92.19 s; Scheme 3 (parallel combination of TextRank algorithm and LDA topic algorithm) in extracting key information of online news is 140.78 s; Scheme 1 (TF-IDF algorithm and TextRank algorithm combined in parallel) extracts the key information of network news, the average time of key information extraction is only 44.77 s. Next, compare the recall rate and accuracy rate of three parallel combination fusion schemes in keyword extraction, and realize the quality comparative analysis of three different parallel combination fusion schemes, as shown in Figure 11.

According to Figure 11(a), the average recall rate of Scheme 2 (parallel combination of LDA topic model algorithm and TF-IDF algorithm) is 0.34; Scheme 3 (parallel combination of TextRank algorithm and LDA topic algorithm) is 0.22; and Scheme 1 (parallel combination of TF-IDF algorithm and TextRank algorithm) is 0.54. From the overall situation of recall rate, the recall rate of Scheme 1 is better than that of Scheme 2 and Scheme 3. Figure 11(b) shows that the average accuracy of Scheme 2 is 0.29, that of Scheme 3 is 0.21, and that of Scheme 1 is 0.35.

In Figure 12, "hot" refers to the popular recommendation method, "$I*$" and "$I*$" refer to the method based on waterfall fusion scheme and the method based on parallel combination fusion scheme, respectively. As can be seen from Figure 12, the effect of popular recommendation is the worst. This is because the popular recommendation method only finds out the popular news list according to the news browsing data of the previous day, filters the news list that users have not browsed, and directly recommends.

3.2. Discussion on Experimental Results. News keyword extraction can help users distinguish junk news and get news content quickly, which has great application prospects. Model fusion is a high-performance classifier composed of multiple classifiers [23]. The research adopts parallel combination fusion technology and designs network news data extraction technology, which can not only realize the accurate extraction of keywords but also greatly shorten the operation time and reduce the budget complexity. For the experimental results of parallel combination fusion extraction, it needs to be carried out from three aspects: intuitive comparative analysis, time comparative analysis, and quality comparative analysis. Among them, intuitive comparative analysis refers to the comparative analysis of the experimental results and the gold standard; time comparative analysis refers to the comparative analysis of the time of extracting network news keywords from the three schemes; quality comparative analysis refers to the respective standard of the three schemes. The accuracy rate and recall rate were compared and analyzed.

From the analysis of the key information extraction time of network news, the average time-consuming of LDA topic model algorithm and TF-IDF algorithm parallel combination is 92.19 s; the average time-consuming of the TF-IDF algorithm and TextRank algorithm parallel combination is

Table 2: Comparison of experimental results.

Title	Standard keywords	Scheme 1	Scheme 2
Corruption Involved in Several Private Kindergartens in South Korea: Embezzling Operating Expenses to Buy Valuable Jewelry	South Korea Kindergarten Corruption related Examination Appropriation	South Korea Examination Appropriation Official Embezzlement	Kindergarten Subsidy Investigate Examination A citizen

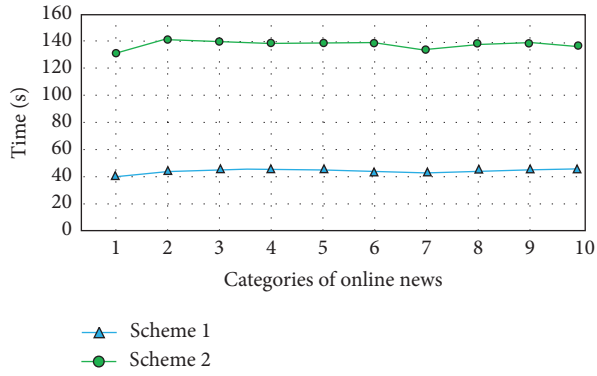


Figure 8: Efficiency comparison chart of keyword extraction.

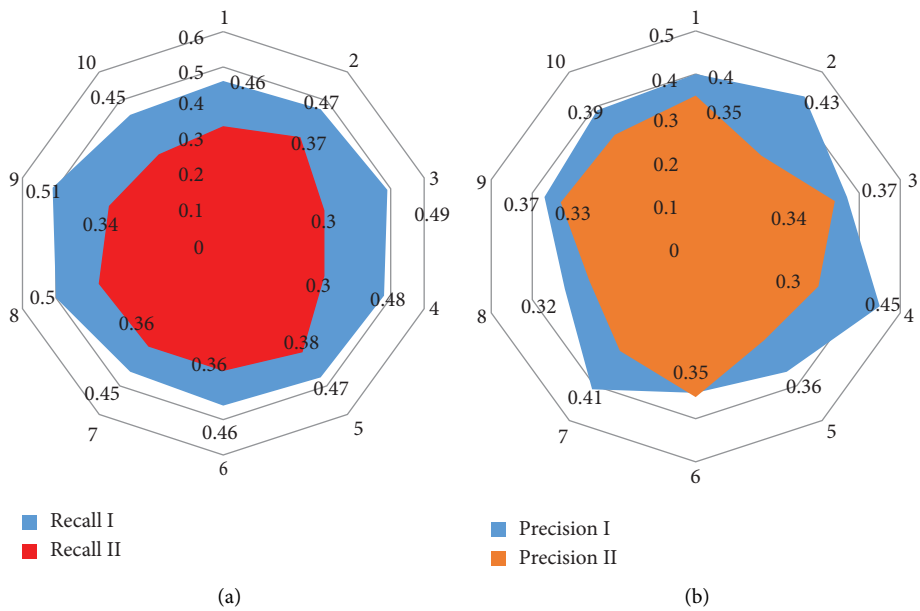


Figure 9: Comparison of recall and accuracy of two waterfall fusion algorithms. (a) Recall comparison. (b) Accuracy comparison.

only 44.77 s. In other words, the TF-IDF algorithm and TextRank algorithm have better time efficiency in network news data extraction. Compared with TF-IDF algorithm, TextRank algorithm, and LDA topic model algorithm, the time consumption of the proposed algorithm on key information of online news is significantly reduced. The average accuracy of the parallel combination of the LDA topic model algorithm and TF-IDF algorithm is 0.29; the average accuracy of the parallel combination of the TextRank algorithm and LDA topic algorithm is 0.21; the average accuracy of the parallel combination of TF-IDF algorithm and TextRank algorithm is 0.35. It can be seen that the accuracy of the TF-IDF algorithm and TextRank algorithm parallel combination is the best, suggesting that researchers can start

Table 3: Comparison of three parallel combination fusion experiments.

Title	Standard keywords	Scheme 1	Scheme 2	Scheme 3
30 People Injured in Apartment Floor Collapse Caused by Japanese College Students' Party	Japan College student University Apartment Accident	University Accident Japan Happen Hold	University Accident Investigation Indoor Student	Accident University Play Happen Party

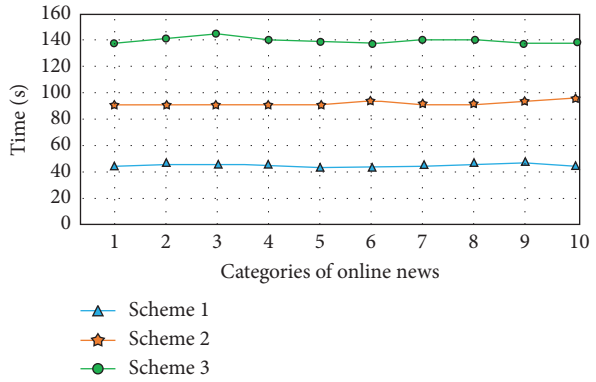
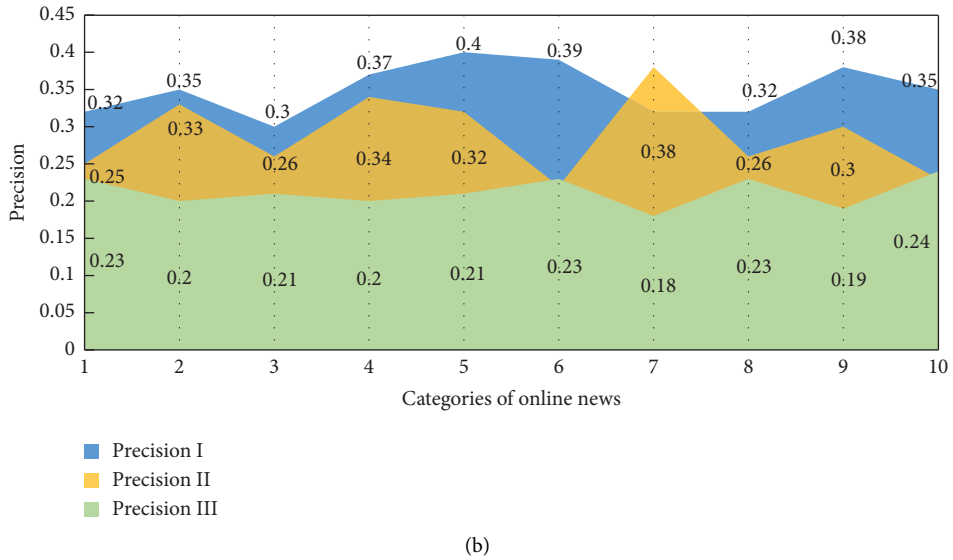


Figure 10: Comparison of three parallel schemes for keyword extraction.

(a)

Figure 11: Continued.



(b)

Figure 11: Quality comparative analysis of three parallel combination fusion schemes. (a) Recall comparison of three parallel combination fusion schemes. (b) Accuracy comparison of three parallel combination fusion schemes.

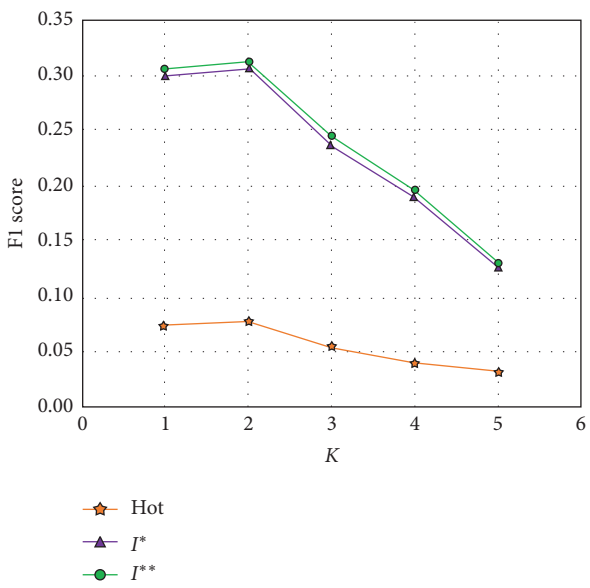


Figure 12: Comparison of news recommendation results under different key information extraction algorithms.

from this combination to further optimize the key information extraction technology of network news [24].

In addition to the consideration of extraction time and accuracy, whether the extracted information meets the needs of users is also very important. In order to compare the application effect of the proposed scheme in the network news data extraction, the research will be based on the above two advantage schemes: first, the TF-IDF algorithm and then TextRank algorithm waterfall fusion scheme, TF-IDF algorithm, and TextRank algorithm parallel combination fusion scheme. The extracted network news keywords are combined with the analysis of different users' historical browsing behavior trajectory, and the corresponding users have analyzed Recommend sex news. And through a recommendation contest platform to score the recommendation effect, compare the news recommendation effect of popular recommendation method, method based on waterfall fusion scheme, and method based on parallel combination fusion scheme and then evaluate the effect of network news data extraction under different methods. From the experimental results, we can see that the popular recommendation method lacks the extraction of network news keywords, and the content lacks pertinence when recommending news to users. Secondly, based on the waterfall fusion scheme, this method can quickly grasp the key information contained in the news through the extraction of network news keywords, improve the pertinence of news recommendation, and effectively shorten the energy consumption and running time in the process of news recommendation. However, the effect of news recommendation based on the waterfall fusion scheme is slightly worse than that based on the parallel combination fusion scheme. This is because in the parallel combination fusion scheme, the two algorithms (TF-IDF algorithm and TextRank algorithm) are not in order, and the corresponding recall rate is high. The more the extracted keywords fit the actual content of news, the more the recommended news is easy to be used as Reading interest.

4. Conclusion

With the rapid development of network news, news content presents an uneven phenomenon, media maliciously exaggerate reports, attract traffic phenomenon is common, and it is difficult for users to quickly obtain the required news content from the massive network news. Network news data extraction technology based on news text keyword extraction has become an effective tool to solve this problem. In view of this, this experiment starts with the

unsupervised keyword extraction method and improves three algorithms based on the analysis of TF-IDF algorithm, TextRank algorithm, and LDA topic model algorithm. The TF-IDF algorithm is improved by yuzipf's law and chi-square test, and five different key information extraction schemes are designed by using waterfall fusion algorithm and parallel combination fusion algorithm combined with the above three unsupervised keyword extraction algorithms. In order to verify the key information extraction effect of different schemes, this paper selects 100 network news of ten categories as the keywords extraction object and verifies the key information extraction effect of network news from three aspects of visual comparative analysis, time comparative analysis, and quality comparative analysis. Finally, through the news recommendation contest, the paper compares the key information extraction effect of network news designed in this study from the side. The results show that the designed extraction technology has a good effect on network news data extraction, and the keyword extraction performance of model fusion is higher than that of traditional extraction methods. Although some achievements have been made in this study, Jieba word segmentation is directly used in the keyword preprocessing step, and the advantages of each algorithm model are not maximized. In the future, a voting mechanism will be introduced to maximize the advantages of each algorithm model, so as to give full play to the advantages of each algorithm model.

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 that they have no conflicts of interest.

References

[1] A. Vanyushkin and L. Graschenko, "Analysis of text collections for the purposes of keyword extraction task," *Journal of Information and Organizational Sciences*, vol. 44, no. 1, pp. 171–184, 2020.

[2] P. Bahad, P. Saxena, and R. Kamal, "Fake news detection using Bi-directional LSTM-recurrent neural network," *Procedia Computer Science*, vol. 165, pp. 74–82, 2020.

[3] P. Symeonidis, L. Kirjackaja, and M. Zanker, "Session-aware news recommendations using random walks on time-evolving heterogeneous information networks," *User Modeling and User-Adapted Interaction*, vol. 30, no. 4, pp. 1–29, 2020.

[4] K. Rohit Kumar, G. Anurag, and N. Pratik, "FNDNet–a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, pp. 32–44, 2020.

[5] T. M. Fagbola, C. S. Thakur, and O. Olugbara, "News article classification using Kolmogorov complexity distance measure and artificial neural network," *International Journal of Technology*, vol. 10, no. 4, pp. 710–720, 2019.

[6] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "Yake! keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020.

[7] L. Li, J. Liu, and Y. Sun, "Unsupervised keyword extraction from microblog posts via hashtags," *Journal of Web Engineering*, vol. 17, no. 1-2, pp. 97–124, 2018.

[8] A. Onan, S. Korukoğlu, S. Lu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, pp. 232–247, 2016.

[9] A. Kolesnikov, P. Kikin, G. Niko, and E. Komissarova, "Natural language processing systems for data extraction and mapping on the basis of unstructured text blocks," *InterCarto. InterGIS*, vol. 26, no. 1, pp. 375–384, 2020.

[10] Z. Yang, H. Yu, J. Tang, and H. Liu, "Toward keyword extraction in constrained information retrieval in vehicle social network," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 4285–4294, 2019.

[11] P. Hofmann, R. Keller, and N. Urbach, "Inter-technology relationship networks: arranging technologies through text mining," *Technological Forecasting and Social Change*, vol. 143, pp. 202–213, 2019.

[12] L. E. Sapozhnikova and O. A. Gordeeva, "Text classification using convolutional neural network," *Information Technology and Nanotechnology*, vol. 2416, pp. 219–226, 2019.

[13] Y. T. Wen, J. Yeo, and W. C. Peng, "Efficient keyword-aware representative travel route recommendation," *IEEE Transactions on Knowledge & Data Engineering*, vol. 29, no. 99, pp. 1639–1652, 2017.

[14] N. M. Ranjan and R. S. Prasad, "Automatic text classification using BPLion-neural network and semantic word processing," *Imaging Science Journal the*, vol. 66, no. 2, pp. 1–15, 2017.

[15] H. Wang and S. Deng, "A paper-text perspective," *The Electronic Library*, vol. 35, no. 4, pp. 689–708, 2017.

[16] H.-M. Kim, "The analysis of characteristics and plan to activate the small wedding reported in Internet news," *Journal of the Korea Entertainment Industry Association*, vol. 13, no. 3, pp. 43–54, 2019.

[17] Z. Wang, K. Hahn, Y. Kim, S. Song, and J.-M. Seo, "A news-topic recommender system based on keywords extraction," *Multimedia Tools and Applications*, vol. 77, no. 4, pp. 4339–4353, 2018.

[18] S. Venkatachalam, L. P. Subbiah, R. Rajendiran, and N. Venkatachalam, "An ontology-based information extraction and summarization of multiple news articles," *International Journal of Information Technology*, vol. 12, no. 2, pp. 547–557, 2020.

[19] H. Mingpan, M. Haigang, and M. Changyong, "Male gibbon loud morning calls conform to Zipf's law of brevity and Menzerath's law: insights into the origin of human language - ScienceDirect," *Animal Behaviour*, vol. 160, pp. 145–155, 2020.

[20] M. A. Helal and M. Mouhoub, "Topic modelling in bangla language: an LDA approach to optimize topics and news classification," *Computer and Information Science*, vol. 11, no. 4, pp. 77–83, 2018.

[21] N. Wen, B. He, Z. Yuan, and Y. Fan, "A object detection algorithm based on pyramid convolutional neural networks (CNN) and feature map fusion model," *Abstracts of the ICA*, vol. 1, p. 1, 2019.

[22] M. Hammad and K. Wang, "Parallel score fusion of ECG and fingerprint for human authentication based on convolution neural network," *Computers & Security*, vol. 81, pp. 107–122, 2019.

[23] X. Chen, L. Ke, Q. Du, J. Li, and X. Ding, "Facial expression recognition using kernel entropy component analysis network and DAGSVM," *Complexity*, vol. 2021, no. 2, pp. 1–12, 2021.
[24] X. Chen, H. Chen, and H. Xu, "Vehicle detection based on multifeature extraction and recognition adopting RBF neural network on ADAS system," *Complexity*, vol. 2020, no. 2, pp. 1–11, 2020.