An Improved TextRank Multi-feature Fusion Algorithm For Keyword Extraction of Educational Resources

Hongyang Zhao\textsuperscript{1,a}*, Qiang Xie\textsuperscript{1,b}

\textsuperscript{1}College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, 211106, China

\*Corresponding author’s \textsuperscript{a}email: zhy393487831@126.com

\textsuperscript{b}email: xieqiang@nuaa.edu.cn

Abstract. In view of the fact that the traditional graph model method which only considers statistical features or general semantic features when extracting keywords from existing massive educational resources, lacks the function of mining and utilizing multi-factor semantic features, this paper proposes an improved TextRank-based algorithm for keyword extraction of educational resources. According to the characteristics of Chinese text and the shortcomings of traditional TextRank algorithm, the improved algorithm featuring multi-feature fusion is developed using the importance of words in the corpus, the location information in the text and the attributes of words. Experimental results show that this method has higher accuracy, recall rate, and F-measure value than traditional algorithms in the process of keyword extraction of educational resources, which improves the quality of keyword extraction and is beneficial to better utilization and management of educational resources.

1. Introduction

With the rapid development of the Internet + education, online education resources are growing explosively. Faced with such a huge amount of information, people need to spend a lot of time and energy on their selection and screening. The concentration of the core content and theme information of educational resources can be demonstrated by keywords. Users can quickly identify the main idea of articles by reading keywords and obtain useful information from massive resources. Keywords extraction plays an active role in information retrieval, topic tracking, text clustering, knowledge map, knowledge base construction, personalized recommendation [1-3], and other aspects. Text keyword extraction is divided into supervised extraction and unsupervised extraction [4]. Supervised keyword extraction requires the pre-marking of corpus and training of corpus by model, which has a high preprocessing cost [5]. Therefore, the existing keyword extraction algorithms mainly adopt unsupervised keyword extraction with strong applicability. Among them is a typical algorithm called TextRank, which, based on the word graph method, describes the relationship between words through a certain algorithm to judge the importance of words in the text.

However, the weight of edges is not considered in the word graph constructed by the traditional TextRank algorithm. To further improve the effect of keyword extraction, Lu et al. [6] combining TextRank with TF-IDF, built a word graph model, counted word frequency and inverse document frequency, and considered the weight of title positioning, so as to extract keywords from the text. Song et al. [7] proposed a new text classification method based on text ranking and Word2vec, integrating the external influence of Word2vec training model and that of position perception weighting of
keywords. Hu et al. [8] proposed a patent keyword extraction algorithm based on distributed Skip-Gram model, which provided an effective method to extract keywords from patent texts for patent classification. Xia[9] introduced two indexes, called internal cohesiveness and boundary freedom, to measure the inner tightness of phrases and the free collocation ability of phrase boundaries respectively, so as to implement the authority calculation of Chinese two-word phrases and the fusion ranking with the extraction results of location-weighted keywords. Wang et al. [10] proposed an extraction method to generate brief topic representation, which took the title in the document title set as the processing object, extracted the common information from different titles, and merged the common information to generate the topic representation of event granularity.

Although there are so many existing keyword extraction algorithms and each has its own advantages, their disadvantages make it impossible for them to fully mine and deeply utilize the multi-factor semantic features, and they can not meet the needs of effective keyword extraction of educational resources which are characterized by a large amount of data, many types of content and diverse forms. To solve the problem, this paper proposes an improved keyword extraction algorithm featuring multi-feature fusion, which fuses the respective weights of the effective frequency of the words in the text set, the position distribution correlation of the words and the word attribute, calculate the weights between the words, then sort the words according to the calculated values and finally extract the keywords.

2. Improve the keyword extraction process of TextRank

![Figure 1. Improved keyword extraction process of TextRank](image)

The keyword extraction process of the improved TextRank algorithm proposed in this paper is shown in Figure 1. step 1: obtain education resource data from network-related education resource platforms; step 2: preprocess the obtained education resource data, and carry out Chinese word segmentation and removal of stop words; step 3: sort the preprocessed education resource data into a text set; step 4, calculate the weight of TF-IDF, word position, and part of speech respectively; step 5: assign corresponding parameters for different weights; step 6, the weight value and parameter value calculated in the previous two steps are brought in to form the score of multi-feature fusion calculation words; step 7: sort the words according to the score calculated in the previous step; step 8, select the first N words as the keywords of the resource.

3. Improve the design of keyword extraction of TextRank

3.1. The Idea Of Improving TextRank Algorithm
The classical TextRank algorithm sets the initial weight of each word node to 1 or 1 / N (N is the number of nodes) by default, and the initial weight of each word is the same. The co-occurrence relationship between words in the corpus document to be processed is used as the edge weight to build
the weight matrix, iteratively calculate the node score, and update the node weight. The potential relationship between words is considered when the weight contribution of word nodes is calculated and the weight is evenly distributed to adjacent nodes. However, the traditional TextRank algorithm has some shortcomings in the application of keyword extraction. It ignores important features such as word frequency, word position, and part of speech, which affect the accuracy of keyword extraction. The importance of each word node is different. For example, the importance of words in different positions in the text set may also be different, which will affect the keyword extraction effect of the whole resource. These correlations are very important, so this paper proposes an improved TextRank algorithm. According to the importance of each word node, different values are assigned as their initial weights, and a multi-feature algorithm for keyword extraction of educational resources is integrated.

3.2 Steps To Improve The TextRank Algorithm

In this paper, the fixed weight in the traditional algorithm is changed into the multi-feature fusion weight of word frequency, position and part of speech, which helps to improve the accuracy of keyword extraction. According to the principle of TextRank, the main steps of the algorithm in this paper are as follows: In the process of keyword extraction, select any word $i$ to calculate the weights of the word $i$.

1. Integrate the comprehensive weight of word $i$, the calculation formula is as follows:

$$ W_{\text{Weight}(i)} = \alpha W_T(i) + \beta W_L(i) + \gamma W_{\text{Pos}(i)} $$  \hspace{1cm} (1)

Among them, $\alpha + \beta + \gamma = 1$, $\alpha$, $\beta$, and $\gamma$ are all greater than 0, which are the TF-IDF algorithm, the position of the word, and the proportion of the part-of-speech feature. $W_T(i)$ is the weight value of the word calculated by TF-IDF, $W_L(i)$ is the weight value of the position information of the word, and $W_{\text{Pos}(i)}$ is the weight value of the part-of-speech feature.

2. Calculate the frequency of word $i$ in the text set. This article chooses the TF-IDF idea to implement it. The calculation formula is as follows:

$$ W_T = tf_{i,j} \times idf_i $$  \hspace{1cm} (2)

$W_T$ refers to the importance of word $i$ in document $j$, that is, word weight. $tf_{i,j}$ represents term frequency, which refers to the proportion of the number of occurrences of term $i$ in document $j$.

$$ tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} $$  \hspace{1cm} (3)

$n_{i,j}$ is the number of occurrences of the word $i$ in document $j$, and $\sum n_{k,j}$ is the sum of the number of occurrences of all words in document $j$. The $idf_i$ in formula (2) is the inverse document frequency of word $i$, which reflects the frequency of words appearing in the overall corpus.

$$ idf_i = \log \frac{|D|}{1 + |D_i|} $$  \hspace{1cm} (4)

Among them, $|D|$ represents the total number of documents in the corpus, $|D_i|$ represents the number of documents containing the word $i$, plus 1 to enhance the robustness of the algorithm.

3. Calculate the position weight of word $i$ in the text set, the calculation formula is as follows:

Combined with the characteristics of text-based educational resources, this paper adds the weight of location information, that is, add different weight information to the words in different positions, to improve the quality of the algorithm. Based on the above principles, the words near the beginning and the end of the document are given greater weights, and different information weights are set according to the position of the words. The document is segmented, assuming that the total number of paragraphs in the document is $a$, and the paragraph where the word $i$ is located in paragraph $b$. Then the weights of all words in segment $b$ are:
In formula (5), the weight of the position information closer to the beginning and the end is greater, on the contrary, the weight of the middle paragraph is smaller and all in the (0,1) interval. In a document, the location information in each paragraph is also different. The first and last sentences serve as a connecting link between the preceding and the following. Therefore, different positions in the paragraph should also be given different weights. The weight of the first and last sentences is more important, and the weight of words in the middle sentence is less. Next, the position information weight $W_{L(i)}$ is normalized to obtain the final position information weight $W_{L0(i)}$.

(4) Calculate the part-of-speech weight of word $i$, the calculation formula is as follows:

The part-of-speech feature is a representation of linguistic knowledge. The part-of-speech of keywords in Chinese text tends to concentrate on substantive words such as nouns, verbs, and adjectives. According to the results of manual labeling, a statistical analysis of the part-of-speech distribution is carried out. The top nouns, verbs, adjectives, and adverbs account for more than 95% of the total number of keywords. Therefore, part of speech is introduced as one of the important features of keyword extraction. The part-of-speech weights are set to 5, 4, 3, 2 according to nouns, verbs, adjectives, and adverbs, respectively, and the calculated values get the part-of-speech weight value $W_{Pos(i)}$.

(5) The calculation formula for the transition probability between nodes is:

(6) Combining the calculation of the above three weight values, the iterative formula of the comprehensive weight is as follows:

In the formula, $W(v_j,v_i)$ is calculated by the formula (6). $d$ is a damping coefficient, usually 0.85.

(7) According to the final calculation of formula (7), the value of $WS(v_i)$ is calculated, and the first N words with larger numerical values are selected as keywords and output.

4. Experiment And Result Analysis

4.1. Experimental Program

The experimental environment in this article is Windows10 x64-bit operating system, Intel Core i5 processor, and 8 GB memory, and the code is implemented in Python language with open source, scalability, and portability. Select 500 educational resources of different lengths, such as thesis, course introduction, audio and video, and picture introduction as the test set.

To test and evaluate the algorithm proposed in this paper, an experiment was conducted to compare it with the commonly used TF-IDF algorithm and the TextRank algorithm. The experiment adopts the precision, recall, and F-measure value that are commonly used in the evaluation of document classification algorithms. The accuracy rate is used to evaluate the accuracy of keywords on the subject information of the original document, and the recall rate is used to evaluate the coverage of the keywords on the subject information of the original document. The F-measure value is the harmonic average of the accuracy rate and the recall rate to weigh the accuracy rate. And the higher the F-measure value is, the better the keyword extraction effect of the algorithm. The calculation formula is as follows:
Among them, A is the number of accurate keywords generated, B is the keywords provided by the data set itself, and C is the number of keywords generated by the algorithm. In the implementation of the algorithm proposed in this article, a maximum of 5 to 10 keywords are generated for a resource through different algorithms.

### 4.2. Experiment Implementation

The experiment uses the natural language processing package to preprocess the texts extracted from each educational resource. First, perform Chinese word segmentation and part-of-speech tagging, then remove stop words in the text, and retain nouns, verbs, adjectives, and adverbs. Afterward, to obtain the best abstract extraction results, this paper sets up a combination experiment of the weighting coefficients $\alpha$, $\beta$, and $\gamma$ of the weight influence factor, where $\alpha + \beta + \gamma = 1$.

The experiment selects different weighting coefficient combinations for keyword extraction. Some representative test data are shown in the table below.

**Table 1. Weighting coefficient values under different ratio combinations**

| Group | $\alpha$ | $\beta$ | $\gamma$ | Group | $\alpha$ | $\beta$ | $\gamma$ |
|-------|---------|---------|---------|-------|---------|---------|---------|
| 1     | 0.8     | 0.1     | 0.1     | 6     | 0.5     | 0.2     | 0.3     |
| 2     | 0.7     | 0.2     | 0.1     | 7     | 0.4     | 0.3     | 0.3     |
| 3     | 0.7     | 0.1     | 0.2     | 8     | 0.3     | 0.4     | 0.3     |
| 4     | 0.6     | 0.2     | 0.1     | 9     | 0.3     | 0.3     | 0.4     |
| 5     | 0.5     | 0.3     | 0.2     | 10    | 0.2     | 0.4     | 0.4     |

Experiment with selected combinations of different proportional weighting coefficients. After repeated experiments, the results show that the 8th group of weighting coefficient combinations, that is, when $\alpha=0.3$, $\beta=0.4$, and $\gamma=0.3$, the text summaries extracted by the experiment have the best results. Therefore, the parameter values in the subsequent algorithm comparison experiments are set according to the above situation.

### 4.3. Experimental Results and Analysis

To verify the effectiveness of the abstract method in this paper, different comparative experiments were set up to extract abstracts.

Three different algorithms are compared in the experiment, namely the classic TF-IDF algorithm, the traditional TextRank algorithm, and the improved multi-feature fusion TextRank algorithm proposed in this article (for convenience, the name is abbreviated as F-TextRank in the following figures). On the test set, set the above parameters and weight coefficients, and compare the three methods under different numbers of extracted keywords. Each method extracts the first 5, 8, and 10 words with the largest weight for comparison. Run these three algorithms separately to compare the values of P, R, and F. The comparison results are shown in the figures below.
Figure 6. Comparison of keyword extraction time of three algorithms

As can be seen from the experimental results in the figure above, compared with the commonly used keyword generation algorithms, the algorithm proposed in this paper has relatively higher accuracy, recall rate, and F-measure value.

According to the experimental data, when the number of keywords extracted is 5, 8, and 10 in the experiment, the F-measure value of the improved TextRank algorithm is improved by 5%, 4%, 6%, 4%, 6%, and 5% compared with the TF-IDF algorithm and the traditional TextRank algorithm. The algorithm proposed in this paper improves the quality of keyword extraction.

With the increase in the number of extracted keywords, the accuracy of the three algorithms decreases, indicating that the words with the highest scores are more likely to be keywords. In terms
of accuracy, the improved TextRank algorithm in this article is better than TF-IDF and Traditional TextRank algorithm; as the number of extracted keywords increases, the recall rate gradually rises. Figure 6 shows the comparison results of the running time of the three algorithms. The improved algorithm adds iterative calculations, so the calculation time is slightly increased, but the increase is small, which means using the improved algorithm to extract keywords does not improve the effect at the cost of increasing operation time. Therefore, the overall fluctuation of the algorithm in this paper is small, which verifies the effectiveness and stability of the algorithm in this paper.

Based on the above experimental data and analysis, the algorithm in this paper is superior to common keyword extraction algorithms such as TF-IDF algorithm and traditional TextRank and has better stability, which improves the quality of keyword extraction to a certain extent.

5. Conclusion
This article analyzes the commonly used TextRank-based algorithms for extracting keywords from educational resources and their limitations, and proposes a multi-feature-fusion keyword extraction algorithm, which takes into consideration word frequency, location information, and part of speech on the basis of TextRank, and modifies the weight of the keywords extracted by the TextRank algorithm so that the keywords extracted by the improved algorithm are not only related to the characteristics of the words themselves but also the position and frequency of the words in the article. The experimental results show that compared with the commonly used TF-IDF, TextRank, and other algorithms, the fusion algorithm proposed in this paper has higher accuracy and stability, which improves the quality of keyword extraction to a certain extent, thus improving the utilization efficiency of educational resources.

References
[1] Chen, W., Chan, H. P., Li, P. J., Bing, L.D., & King, I. (2019) An integrated approach for keyphrase generation via exploring the power of retrieval and extraction. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis. pp.2846–2856
[2] Yang, D. H., Wu, Y. X., & Fan, C. X., (2020) Chinese Short Text Keyphrase Extraction Model Based on Attention. Computer Science, 47(1): 193-198
[3] Collins, A. and Beel, J. (2019). Document Embeddings vs. Keyphrases vs. Terms for Recommender Systems: A Large-Scale Online Evaluation. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). Urbana. pp. 130–133
[4] Huang, R. Z., Huang, D. C., (2019) Words' Relation Based on Uncertain Graph and Automatic Keyword Extraction. Journal of Chinese Computer Systems, 40(2) : 300-304
[5] Duari, S., & Bhatnagar, V. (2020) Complex network based supervised keyword extractor. Expert Systems with Application, 140: 112876
[6] Lu, Y., Zhang, P. Z., & Zhang, C., (2019) Research on news keyword extraction technology based on TF-IDF and TextRank. In: 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS). Beijing. pp 452–455
[7] Song S. Y ., Wang Z. R., Xu S., Ni S., Xiao J. (2019) A novel text classification approach based on Word2vec and TextRank keyword extraction. In: 2019 IEEE fourth international conference on data science in cyberspace (DSC). Hangzhou. pp. 536–543
[8] Hu, J., Li, S. B., Yao, Y., Yu, L. Y., Yang, G. C., & Hu, J. J., (2018) Patent keyword extraction algorithm based on distributed representation for patent classification. Entropy, 20(2): 104
[9] Xia, T. (2020) Extracting Key-phrases from Chinese Scholarly Papers. Data Analysis and Knowledge Discovery, 4(7): 76-86.
[10] Wang, W. Y., Shi, C. H., Yu, X. M., Liu, Y., & Cheng, X. Q., (2021) An extractive topic brief representation generation method to event. Journal of Shandong University (Natural Science), 56(05): 66-75+84.