Research on News Keyword Extraction Based on TF-IDF and Chinese Features

Jiapeng Song[1, a], Rui Hu[1, b], Bingyu Sun[1, c], Yin Gu[1, d], Wenlin Xiong[1, e] and Jianqi Zhu[2, f]

[1]College of Software, Jilin University, Changchun, China

[2]College of Computer Science and Technology, Jilin University, Changchun, China

[a]Songjiapeng99@126.com, [b]240775068@qq.com, [c]sicerain@163.com, [d]2530342524@qq.com, [e]1471901441 @qq.com, [f]zhujq@jlu.edu.cn

Keywords: TF-IDF, Chinese news, keywords composition, weight calculation

Abstract: Keyword extraction technology is the basis of corpus construction, text analysis processing, and information retrieval. For the special carrier of Chinese news text, the traditional TF-IDF algorithm is too dependent on word frequency and cannot handle the drawbacks of Chinese grammar accurately. This paper elaborates on the characteristics of Chinese news text keywords. On the basis of TF-IDF algorithm, it integrates Chinese special features such as part of speech, word length and lexical, and constructs an improved TF-IDF weighting formula that comprehensively considers text features. A scoring method for keyword matching is proposed, and the keywords that are "cut off" by the Chinese word segmentation are reconstituted into formal keywords. Cross-comparison experiments show that the improved algorithm is superior to the traditional algorithm in accuracy, recall and F value, and can correctly and effectively extract Chinese keywords.

1. Introduction

With the continuous development of the Internet, the content of network information has been continuously enriched and gradually complicated. In today's networks with such huge amounts of information, how to effectively manage and find these data has become a focus issue. Firstly, the arrival of the big data makes it difficult for learners to quickly choose the suitable learning resources. Secondly, learning style, learning preference and learning environment of user make their expectations of learning effect inconsistent [1]. For the network news part, the method of using text keyword extraction for classification management has been widely recognized and adopted. However, manual keyword processing requires a lot of manpower and material resources, which has become an impossible issue for the current Internet environment. Therefore, how to use computer to ensure accurate and efficient text keyword extraction has become a natural language processing.

The main job of keyword extraction is to extract words that can represent the core content of the document. Two stages of technology text document retrieval system is a pre-text processing and text representation. Pre-processing of the text consists of many stages, such as tokenizing, stemming, and stop listing. While the text representation stage commonly known as text weighting stage. There have been many previous studies that propose new methods for text weighting. Weighting method which is still commonly used, namely Term frequency - inverse document frequency (Tf-Idf) considering the frequent appearance of the term in the document and the ratio of the length of the documents in the corpus [2]. The methods used first can be divided into two categories, one is based on machine learning, mainly for extracting abstracts of academic papers in digital libraries. The idea of using keywords as the basic unit of extraction for keyword extraction is proposed. Introduce the literature (automatic extraction method based on machine learning for scientific abstracts keywords). Yang Wenfeng improved the PAT tree construction algorithm and proposed to extract the keywords using the largest repeated string in the document [3]. Zhang Kuo et al. proposed keyword extraction based on the SVM model [4].

The accuracy of keywords extracted by these methods is high, but it is necessary to manually build large-scale corpus for training, and the reliance on corpus is high. In the current situation of network information explosion, it is necessary to invest a large number of algorithms to meet the accuracy requirements. Human and material resources, and because online news covers all aspects of the particularity, this cost will be higher. The other is a statistical-based keyword extraction method, such as the TF-IDF algorithm and its derivative improvement algorithm. The TF-IDF method has strong universality. The basis of the judgment is that if a word appears in the article with a high number of times, but at the same time it appears less frequently in other articles, the word is considered to have a certain generality. The more complex algorithms have their own problems. For example, the actual application effect of TextRank is not superior to TF-IDF, and the algorithm is extremely inefficient due to the iterative algorithm involving network construction and random walk. Like TopicModel, the extracted keyword content is too broad and does not reflect the subject of the article well. After pre-testing and literature review, the author believes that the TF-IDF method without prior training costs is more advantageous in dealing with online news. However, due to the particularity of Chinese grammar, there is no natural delimiter between phrases, so there are two serious problems when using this method for Chinese news processing: one is too dependent on word frequency, and the other is too dependent on word segmentation. Accuracy. These two problems make the TF-IDF method less accurate in Chinese keyword extraction. TF-IDF is used majorly to stop filtering words in text summarization and categorization application. By convention, the TF-IDF value increases proportionally to the number of times that a word appears in a document, but is offset by the frequency of the word in the corpus, which helps to control the fact that some,words are more common than others. The frequency term means the raw frequency of a term in a document. Moreover, the term regarding inverse document frequency is a measure of whether the term is common or rare across all documents in which can be obtained by dividing the total number of documents by the number of documents containing the term [5]. Term frequency inverse document frequency (Tf-Idf) is a calculation that illustrates the importance of the word (term) in a document and a corpus. This process is used to assess the weight of term relevance of a document to all documents in the corpus [6].

In this regard, this paper improves the traditional TF-IDF algorithm for the particularity of Chinese language, and calculates the weight by introducing multiple factors such as part of speech and grammar, and finally extracts the key to the high refinement in line with the main purpose of the article by combining the keywords. Word. The algorithm experiment is carried out by using python. The experimental results show that the improved algorithm is simple in calculation and high in accuracy, and can effectively extract and apply Chinese keywords.

2. Chinese news keyword composition analysis

The following functions are mainly implemented in

2.1 Chinese news text analysis

The Chinese news text and the general text are different in content form, which are manifested in the following aspects: 1. The length of the news text is generally short, there is a clear core of the subject, and the key words will be repeated. 2. The headline of the news is often a high-level summary of the entire article, which can reflect the main purpose of the article. 3. The news text reinforces the role of a noun, a named entity, which is often closely related to the subject matter of the article. Due to these unique linguistic features of Chinese news texts, the author made a targeted improvement on the traditional TF-IDF method, and calculated the weights of multiple features of the keyword together with the word frequency to construct a new scoring formula. The importance of the extracted keywords is sorted, and finally effective and accurate keywords are extracted.

2.2 Keyword composition analysis

The keyword should be a word or phrase that can represent the whole article, so that the content of the document can be summarized and conveniently searched, so that the user can roughly understand the content of the document. The results of existing research show that most of the text keywords are composed of phrases, and rarely consist of individual words. For example, the keyword "Leshan Giant Buddha" can be further divided into two words "Leshan" and "Big Buddha", and each word has its own meaning. However, as the key word of the article, "Leshan Giant Buddha" is more general than "Leshan" or "Big Buddha", and is more suitable as a keyword. Similar examples include the "Graduation Thesis" and the "China Ministry of Education".

Keyword extraction technology is the basis of text classification, text clustering, information retrieval and other technologies [7]. Unlike languages such as English and French, there is no obvious delimiter between words in Chinese. We can't implement words separated by spaces like English [8]. Segmentation of Chinese is a requirement for many applications [9]. Chinese text has no obvious word boundaries [10]. This makes the keyword extraction of Chinese texts help with the help of the word segmentation tool. However, with the continuous development of network culture, the amount of network information has exploded, and new phrases have emerged. This has led to the current segmentation tool not being able to update the corpus and train it in time. The literature shows. Even the latest word segmentation system at this stage cannot guarantee the segmentation accuracy of untrained new words. At the same time, due to the particularity of the news carrier, online words and new technical terms often appear in the text. This makes the accuracy of the word segmentation seriously affect the extraction effect of Chinese keywords. For example: "Zero Experience Programmer" will be divided into "zero", "experience" and "programmer", which are regarded as words appearing in the document, which makes the late keyword extraction effect not accurate enough.

3. Main improvement direction

The two parts of the traditional TF-IDF algorithm are mainly improved:

(1). The TF-IDF algorithm only considers the influence of word frequency on keywords, and ignores the influence of other factors such as part of speech on the importance of words. We will make different weight judgments on words with different parts of speech.

(2). Make application adjustments for Chinese news: more accurate extraction of phrases and weights in Chinese keywords.

3.1 TF-IDF algorithm

The TF-IDF algorithm is a weighting algorithm based on statistical angles. The basic idea is to use the Term Frequency and the Inverse Document Frequency to estimate how important a word is to the entire article. The calculation formula is:

$$\mathrm{tf}_{\mathrm{i}}\mathrm{idf}_{\mathrm{i,\,j}} = \mathrm{tf}_{\mathrm{i,\,j}} \times idf_{i} \tag{1}$$

$$\mathrm{tf}_{\mathrm{i},\,j} \times \frac{n_{i,\,j}}{\sum_{k} n_{k,\,j}} \tag{2}$$

$$\mathrm{idf}_{i} = \lg \frac{|D|}{|\{j : t_{i} \in d_{j}\}|} \tag{3}$$

The number of times the word appears in the document / the total number of words in the article; $\mathrm{idf}_{\mathrm{i}}$ = lg (the total number of articles / the number of articles in the corpus containing the word). The tf value represents the frequency at which the entry appears in the document d, and the idf value represents the class distinguishing ability of the word in the corpus.

```
Algorithm.TF-IDF
input: wordlist, filelist, corpuslist
Output: Dictionary containing the results of the article TF-IDF calculation
tf ← 0
idf ← 0
corpuslist ← freqword(wordlist)
For still word in worlist do
    tf ← dic[word]/len(wordlis)
    Idf ← math.log(len(filelist)/(wordinfilecount(word, corpuslist)+1))
    tfidf ← tf × idf
    outdic[word] ← tfidf
End for
Return outdic
```

According to the above formula, when a word appears frequently in a document, but the frequency of occurrence in the entire corpus document is low, the word will get a larger weight via the TF-IDF algorithm, indicating that the word Words can summarize the content of the article to some extent.

3.2 Improvement factor

In view of the characteristics of Chinese news, the factors affecting the final weight calculation are increased, so that the final generated keywords no longer simply consider the word frequency, but the result of comprehensive consideration of the word.ds can summarize the content of the article to some extent.

- Location factor

Due to the particularity of the news carrier, the length of the article is relatively short and the content is compact. The words with summative nature at the beginning and end of the article often have a strong generalization and can reflect the main purpose of the article. Therefore, words appearing at the beginning and end of the title are considered to be more important than others. The position of the word appears to affect its importance to a certain extent. Avoid combining SI and CGS units, such as current in amperes and magnetic field in oversteps. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation. Assume that the position weighting factor is "t_{loc}". Then the formula for position weighting is as follows:

$$\mathrm{loc(wi)} = 1 + \mathrm{tloc} \qquad (4)$$

The position factor t_{loc} is divided into the following cases according to the position where the words appear: 1. the words appear in the title, then t_{loc}=3; 2. the words appear at the end of the article, then t_{loc}=2; in other cases: t_{loc}=0. Enhance its importance by giving the title and the words at the end a higher weight.

- Part of speech factor

Part of speech is the result of grammatical features as the main basis and lexical meaning to divide words. Modern Chinese words can be divided into 14 kinds of words. One type is real words: nouns, verbs, adjectives, numerals, quantifiers, pronouns, distinguishing words, and one type is virtual words: adverbs, prepositions, conjunctions, auxiliary words, interjections, modal particles, onomatopoeia. According to the above analysis of Chinese keywords (in fact, the above mentioned can rarely be added), Chinese keywords are mostly distributed in: nouns, noun phrases, then verbs, and finally decorative words such as numerals and adjectives. Words such as pronouns and conjunctions are not suitable as keywords.

Table.1. Partial word segmentation part-of-speech tagging

code	name
a	*adjective*
ad	*Sub-word*
an	*adnoun*
n	*noun*
ns	*place name*
nt	*Institutional group name*
nz	*Other proper nouns*
nl	*Noun idiom*
v	*verb*
vd	*adverb*
vn	gerund

Set the part of speech factor to t_{pos}, then the part-of-speech weighting formula is:

$$\sin(wi) = wi_{\sin} \times t_{pos} \tag{5}$$

The value of t_{pos} is related to the part of speech:

1. When the part of speech is a named entity, $\mathrm{t}_{pos} = 3$;
2. When the part of speech is a general noun, $\mathrm{t}_{\mathrm{pos}} = 2$;
3. When the part of speech is a verb, $\mathrm{t}_{pos} = 1.5$;
4. When the part of speech is numerals, prepositions, etc, $\mathrm{t}_{pos} = 0$.

Incorporating word-of-speech into keywords can better change the importance of keywords according to the locale.

- Word length factor

The author believes that in Chinese news documents, the longer a word is, the more information it contains, and it is usually more general. However, it cannot be simply assumed that the keyword with a length of 4 is twice as important as a keyword with a length of 2, which is obviously wrong. Therefore, before bringing the final weight calculation formula, you need to unify the word length. The word length formula is as follows:

$$\mathrm{length}(wi) = \frac{len(wi)}{Max(len(w1), len(w2) ... len(wk))} \tag{6}$$

The uniformization of word lengths relatively balances the weight differences of keywords due to their length.

3.3 Comprehensive weight calculation

In news web pages, keywords are usually not composed of a single word, but a phrase consisting of adjacent words. However, due to the Chinese word segmentation mentioned above, keywords are often separated in the process of word segmentation, resulting in an unsatisfactory result of the final result algorithm. Therefore, the author proposes a new keyword matching algorithm. After calculating the candidate keywords according to the improved TF-IDF algorithm, the Chinese words are recombined into Chinese phrases using a specific algorithm, and the importance is judged by giving the keywords a score.

- Combination ideas

The first step is to obtain the importance degree scores of all the words in the article according to the comprehensive weighting algorithm, and select the 10 words with the highest score to form the candidate keyword set.

The second step is to randomly combine the words in the keyword set to generate a multi-phrase. But it is clear that the combination of two keywords does not necessarily produce a valid multi-phrase. It is not easy to assume that a multi-phrase is more important than the two words that make up it. So we need to design a program to measure the validity and importance of multiple phrases. The author proposes the concept of “composition coefficient”, which is the importance score of the multi-phrase calculated by a specific algorithm, and can be directly compared with the importance degree score calculated by the integrated weight of the unary word, so as to facilitate Pick the final keyword from the multi-phrase and the unary word.

The third step is to comprehensively sort the candidate keywords and the grouped keyword scores, and obtain the higher scores as the final officially selected keywords.

Algorithm.Keyword combination
Input:keyword: m,   weight: wei_m, keyword: n, weight: wei_n
Output: Replacing the reduced conjunction paradigm Φ with equivalent variables

```
Word_A ← m + n
Word_B ← n + m
Sum_A ← freq(word_A)
Sum_B ← freq(word_B)
Sum ← max(Sum_A,Sum_B)
score ← max(float(Wei_m),float(Wei_n))
Final_one ← Sum / (freq[m]+freq[n]+Sum)
Final_two ← (1 + final_one) * score
Return Final_two
```

- Composition coefficient and its calculation

Suppose there are two candidate keywords w_i and w_k, and their word frequency in the article is calculated as f_i and f_k. There are two possible situations in which two keyword combinations are generated: the first one is in the form of $w_i + w_k$, and the number of occurrences in the text is $f_{i,k}$; the other is in the form of $w_k + w_i$, which is recorded in the number of occurrences in the text is $f_{k,i}$. Take the larger of $f_{i,k}$ and $f_{k,i}$ as a candidate combination of w_i and w_k, denoted as w_g, and record the word frequency as f_g. If f_g <1, that is, the grouping words do not appear in the text, then the grouping words are not considered to have any meaning to the article, and the combination coefficient and score are not continuously calculated. When $f_g \geq 1$, that is, the grouping word appears at least once in the article, the phrase frequency is converted to the grouping dependency according to the following formula:

$$Rely_{w_g} = \frac{f_g}{Max(f_i, f_k, (f_i + f_k - (k+1) * f_g))} \tag{7}$$

Where f_i, f_k, and f_g are the candidate word w_i, the candidate word w_k, and the candidate group matches the word frequency corresponding to w_g. K is the weighting factor, and $+\infty > K > 0$, K is inversely related to the length of the article. However, the derivation workload of the calculation K is too large, and it is often necessary to perform a large number of experiments for iterative derivation to obtain the relative optimal evolution formula. Therefore, the following formula is used to roughly calculate the value of K:

$$K = \frac{5 * \sum_i^n f_i}{f_{allword}} \tag{8}$$

Where f_i is the word frequency of each word in the candidate keyword set, and $f_{allword}$ is the total number of words in the article.

- Grouping score calculation

In order to make the score of the combined phrase can be compared with the score of the ordinary unary candidate keyword, the author thinks that it is necessary to use the keyword score of the candidate keyword, and after averaging it, combined with the combination coefficient, the final assembly key is calculated. Word rating. The author proposes the following formula for calculating the composition coefficient:

$$Score_g = \frac{1}{2}(\mathrm{Re}\,ly_{w_g} + 1) \times (S_i + S_k) \tag{9}$$

Among them, $Score_g$ represents the group word score finally obtained by calculation, $\mathrm{Re}\,ly_{w_g}$ is the composition coefficient of the above two candidate keywords, and c S_i, S_k respectively represent the keyword scores of the keywords w_i, w_k.

3.4 Experimental results and analysis

- Experimental data and evaluation criteria

This paper selects 120 Chinese news texts crawled from the CCTV news in real time as experimental corpus to ensure that the test texts are randomly sampled. The corpus is divided into group A: 20 and group B: 100 groups, and manually divide the keywords of the article manually, and eliminate special articles with no valid keywords in some articles. The test improves the effectiveness of the algorithm when the corpus is insufficient or sufficient to verify whether the algorithm is superior to the traditional keyword extraction algorithm.

In this paper, the improved algorithm based on TF-IDF, traditional TF-IDF algorithm, and another common keyword extraction algorithm-TextRank algorithm, these three keyword algorithms are cross-correlation experiments, and the results are comprehensively evaluated. The evaluation results were evaluated using the accuracy rate P and the recall rate R and F values.

Among them, TextRank algorithm is also a widely used keyword extraction algorithm. TextRank algorithm is a graph-based sorting algorithm for text. By dividing the text into several constituent units (words, sentences) and establishing a graph model, the voting mechanism is used to sort the important components in the text, and the keyword extraction and abstracts can be realized by using only the information of the single document itself. Similar to the TF-IDF algorithm, the TextRank algorithm does not require prior learning of the document, eliminating the cost of establishing a training set.

a). Precision refers to the ratio of the number of keywords identified by manual extraction and algorithm extraction to the total number of keywords extracted by the algorithm. The accuracy rate can indicate the ability of the algorithm to accurately extract keywords:

$$\text{Precision} = \frac{\text{Correctly extracted entries}}{\text{All extracted entries}} \tag{10}$$

b). Recall refers to the ratio of the number of keywords selected by manual selection and algorithm extraction to the total number of artificially extracted keywords. The recall rate can reflect the ability of the algorithm to capture keywords:

$$\text{Recall rate} = \frac{\text{Correctly extracted entries}}{\text{All correct entries}} \tag{11}$$

c). F-Measure is a comprehensive assessment of accuracy P and recall rate R:

$$\text{F - Measure} = \frac{2\text{PR}}{\text{P} + \text{R}} \tag{12}$$

- Data analysis

In order to comprehensively verify the effectiveness of the improved algorithm, the following three sets of comparative experiments were carried out. In order to reduce the impact of chance on the results, each group of experiments was repeated several times. After removing the highest and lowest results, the average value was calculated to obtain the final P, R, F values. The experimental results are shown in Table 2:

Experiment 1.1 The news keyword extraction was performed on the experimental data of Group A using the traditional TF-IDF algorithm.

Experiment 1.2 Keyword extraction of Group B experimental data using traditional TF-IDF algorithm

Experiment 2.1 Keyword extraction of group a experimental data using TextRank algorithm

Experiment 2.2 uses TextRank algorithm to extract keyword from group B experimental data

Experiment 3.1 The keyword extraction of group A experimental data was performed using the improved TF-IDF algorithm.

Experiment 3.2 Keyword extraction of Group B experimental data using the improved TF-IDF algorithm.

Table.2. The experimental results

Experiment	P	R	F
Experiment 1.1	31.57%	39.21%	0.3497
Experiment 2.1	28.94%	35.94%	0.3206
Experiment 3.1	52.36%	65.03%	0.5801
Experiment 1.2	36.84%	45.75%	0.4081
Experiment 2.2	34.21%	42.48%	0.3789
Experiment 3.2	58.94%	73.20%	0.6530

The above experimental results show that the improved keyword extraction algorithm based on TF-DF algorithm is obviously superior to the traditional method, and it is far ahead of the traditional algorithm in the three criteria of accuracy, recall and F value. At the same time, the author also noticed the following points in the experiment:

a). The improved TF-IDF algorithm can still maintain high accuracy and recall rate when the article has no obvious keyword features or the statement is too long to make the keyword difficult to lock. However, the traditional algorithm is wrong at this time. A meaningless word with a higher frequency is identified as a keyword. And in more than 100 news tests in this experiment, the improved algorithm in 96.58% of the articles, the obtained keywords should be accurate than the traditional TF-IDF algorithm. This proves that the improved algorithm has better stability and can better cope with the characteristics of Chinese news and grammatical complexity.

b). The experiment in this article completely simulates the real news gathering environment. The news uses all the current news obtained from CCTV news, which makes the text frequently appear as a proper name such as the name of the person who is not recognized by the word segmentation module. These proper nouns are often divided into multiple phrases, and traditional algorithms cannot obtain any valid proper nouns with low accuracy. The improved TF-IDF algorithm can reconstitute these fragmented words into valid and specific words, and the accuracy is greatly improved, which can better be compatible with the characteristics of network words and new technical terms frequently appearing in news texts.

4. Conclusion

In this paper, the author discusses the shortcomings of the traditional tf-idf algorithm in extracting Chinese news text keywords, and discusses the reasons for its poor performance, and proposes a set of improvement schemes. On the basis of the traditional tf-idf algorithm, considering the characteristics of Chinese text grammar and part of speech, a keyword scoring weighting formula that comprehensively considers many factors such as part of speech and word length is constructed. Secondly, in order to cope with the problem that the keyword precision is reduced due to the unclear Chinese word segmentation, this paper proposes a combination algorithm of candidate keyword combination and related scoring formula. By using the word frequency according to the word frequency, the position is grouped into a keyword phrase. After the word segmentation, the combination of the words forms the final officially extracted keywords.

The experimental results show that the algorithm described in this paper can effectively improve the shortcomings of traditional TF-IDF algorithm in Chinese news texts. In the absence of manual training, it is also possible to extract sufficiently accurate keywords. However, due to limited experimental conditions, the standard results of the test data in this paper still rely mainly on manual drafting. Although the contingency of the lack of corpus in the experiment process has been reduced as much as possible, the richness and diversity of the test data are still available. Lack of it also has certain limitations in terms of data size. Secondly, in the proposed formula, there are also some coefficients that need a lot of deduction to get the optimal solution. In this paper, only the estimated values are used to prove the validity and rationality of the algorithm, and the optimal solution of these data still needs a lot of experiments. Analysis can be known. Therefore, the author believes that there is still room for further optimization of the algorithm, and the next step will be to further verify and improve the algorithm when conditions permit.

References

[1] L. Yang, R.J. Zheng, J.L. Zhu, MC. Zhang and Q.T Wu, “A Green Cloud Service Provisioning Method for Mobile Micro-Learning”, International Journal of NEW Technology and Research (IJNTR), ISSN: 2454-4116, Volume-4, Issue-4, April 2018 Pages 63-69.

[2] G.Salton, M.Gill, M.J “Introduction to Modern Information Retrieval,” New York: Mc Graw Hill Book. Co; 1983.

[3] K. Zhang, H. Xu, J. Tang. “Keyword extractionusing support vector machine,” Proceedings of the 7th International Conference on Web-Age Information Management, Hong Kong, China, 2006: 85-96

[4] M. Olena, I.H. Witten, “Thesaurus-based index term extraction for agricultural documents,” Proceeding soft he 6th Agricultural Ontology Service Workshop at EFITA/WCCA. VilaReal: IEEEPress, 2005: 11-22

[5] H. Christian, M.P. Agus, and D. Suhartono, “Single Document Automatic Text Summarization Using Term Frequency-Inverse Document Frequency (TF-IDF),” ComTech Vol. 7 No. 4 December 2016: 285-294.

[6] M.N. Saadah, R.W. Atmagi, D.S. Rahayu, and A.Z. Arifin, “Information Retrieval Of Text Document With Weighting TF-IDF And Lcs,” Department of Informatics Engineering, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.

[7] J.N. Zhang, “A Chinese Keywords Extraction Approach Based on TFIDF and Word Correlation,” Information Science, 2012, 30 (10): 1542-1544+1555.

[8] H.H. WANG, “Research on the Application of Chinese Word Segmentation Algorithm in Search Engine,” Management & Technology of SME, 2019 (01): 103-104.

[9] H.G. Chen. “Artificial Intelligence and Chinese Word Segmentation,” China New Telecommunications, 2019, 21 (04): 66-68.

[10] W.H. Xu, Y.K Wen, “A Chinese Keyword Extraction Algorithm Based on TFIDF Method,” Information Studies: Theory & Application, 2008 (02): 298-302.