A TEXT FEATURE WORD EXTRACTION METHOD APPLIED TO ENTERPRISE COMPETITIVE INTELLIGENCE SYSTEM

Zhiwei ZHANG[1,*], Haining ZHANG[2], Guangliang ZHU[3]

To acquire industry feature words, professionals need to collect and analyze the feature word sets from industry sites according to their experience and finally merge the feature word sets at different sites to form industry feature word sets. This method is characterized by a large workload, the difficulty in ensuring the accuracy of text classification, and the necessity to adjust feature words by repeating the above process. To solve the above problems, a universal feature word extraction scheme and a system framework were first proposed in this study based on the actual requirements of the enterprise competitive intelligence system. Then, the key problems involved in the process of feature word extraction were elaborated in detail. Finally, the traditional feature word weight was improved on basis of predecessors' results, and the Sogou lexicon was introduced to correct the word frequency and part of speech. A good classification effect was achieved through experiments with the KNN classifier, which was verified using the vector space model (VSM), and a high average F1 value was acquired.

Keywords: feature word extraction, information extraction, part-of-speech tagging, text classification, competitive intelligence, natural language processing

1. Introduction

Stimulated by the global market-driven economy, information competition has developed into an important aspect in enterprise competition. To timely grasp opportunities and win in the ever-changing competitive environment, enterprises need to master mass reliable, accurate, and valuable information that can provide a reference for their decision-making. Meanwhile, they are bound to keep up with the market conditions, monitor their competitors, and predict the industry development trends so as to make correct analytical decisions and enhance their core competitiveness. One of the feasible and effective means of achieving such a goal is to establish and perfect an accurate and efficient enterprise competitive intelligence system [1]. With the increasingly strengthened Internet openness, an increasing number of enterprises and public institutions have released information

[1] PhD eng., School of Informatics and Engineering, Suzhou University, China, e-mail: zzwloveai@gmail.com
[2] School of Informatics and Engineering, Suzhou University, China
[3] School of Informatics and Engineering, Suzhou University, China

like enterprise profile, product introduction, discount activities, talent recruitment, and technology introduction on the Internet, all of which constitute important data forming competitive intelligence [2].

Even though there is a vast amount of information on the Internet, all enterprises need is valuable text information. Therefore, if the webpages captured from the Internet can't extract and classify the target information in time and effectively, three disadvantages will be brought [3]: (1) Massive webpage information will greatly aggravate the storage burden of the information captured by the system; (2) Massive webpage information will bring great inconvenience to the subsequent information processing and related information processing, which even can't be completed; (3) Due to massive webpage information, enterprise users will be continuously lost in the vast ocean of information, making it difficult to find intelligence information related to themselves. Such intelligence information, if acquired luckily, can't be classified [4].

To solve the above problems, a traditional way is to divide the themes according to the characteristics of the industry or site, and then the content of the webpage information captured from the Internet is manually screened by the information processing personnels [5]. However, this method not only consumes huge manpower and material resources but also suffers from subjective factors. And the classification results are different due to the uneven professional qualities of different periods occupied in this field, and even the same person will have different classification results in different periods [6]. In this study, however, a universal feature word extraction scheme and a system framework were proposed. The main contributions of this paper include:

(1) A new feature word calculation method was put forward based on *TF-IDF*[4] according to the application requirements of enterprise competitive intelligence system.

(2) Chinese word frequency and speech-of-word were corrected by introduce the *Sogou* lexicon.

(3) An industry feature word compounding scheme was proposed.

The reminder of this study was organized as follows: Section 2 introduced the research works related to this paper; Section 3 presented the text feature word extraction framework of the enterprise competitive intelligence system; Section 4 summarized the relevant experimental environment, experimental data, and analysis of experimental results; Section 5 provided the conclusions and the future work direction.

[4] TF-IDF is a technique used to measure the importance of a word in text. Where TF represents term frequency, which is the frequency at which a word appears in a document, while IDF represents inverse document frequency, indicating the prevalence of a word throughout the entire document set.

2. Related Works

2.1. BOW-based text feature word extraction

The BOW[5] model-based text feature representation methods are mostly used in the fields of NLP and information retrieval. Relying on the BOW model, Liu et al. [7] proposed a new model specially designed for XML keyword query, given the fact that the traditional BOW model can't distinguish the roles and relationships of keywords. And, they designed a scoring method based on the proposed model, which can satisfy both query semantics and structural characteristics of XML documents. Qin et al. [8] proposed a character word topic model based on BOWs to consider the relationship between characters and words in topic modeling and designed two types of experiments to evaluate the newly proposed model. Irie et al. [9] studied the BOW representation of word sequences, and the results show that BOW features significantly improve the model performance and efficiency.

In addition, the idea of the BOW model has achieved good results in other fields in addition to text feature extraction. Li et al. [10] put forward a handwriting identification method that quantitatively expresses the shape features of characters by geometric moments, aiming at the high similarity and randomness of handwriting. Liu et al. [11] formed temporal and spatial features into BOW and put forward a sequential BOW model to classify human actions, which captures the sequential structure by dividing the whole action into sub-actions. Experimental consequences indicate that this method is robust and superior to most existing BOW-based classification methods. Kim et al. [12] proposed SymbioLCD to solve the failure of visual BOW to capture the semantic or spatial relations between features by using scale-invariant spatial and semantic matching and achieves good loop closure detection performance.

2.2. Statistical model-based text feature word extraction

As a classical statistical text feature extraction method, the *TF-IDF* has been the most widely employed [13]. Considering the fact that *TF-IDF* neglects the semantic information in texts, Huang et al. [14] analyzed the semantic information of important words in the text based on *TF-ID*F model, first applied natural language processing technology to preprocess the text, and then employed the *TF-IDF* method to find the important words with high *TF-IDF* values in the text. Combining the weighted tree of word similarity and the definition of text semantic similarity, and the clustering experiment was carried out on the benchmark text dataset. The results show that the method proposed by Huang et al. [14] is superior

[5] BOW is the abbreviation for Bag-of-words. The BOW model is widely used in text classification, and the frequency of word occurrences can be used as a feature for training classifiers.

to *TF-IDF* in the aspect of F-measure. Mutual information, which aims to figure out the mutuality between two objects, here measures the statistical independent relationship between feature words and categories.

Aiming at the feature extraction from text information of different industries in the enterprise competitive intelligence system, a complete feature word extraction framework was designed and implemented in this paper.

3. Text Feature Word Extraction Framework for the Enterprise Competitive Intelligence System

3.1. Target feature word extraction

(1) Feature words are supposed to identify the text content. Evaluating whether a feature word is good or not, whether a feature word set is selected reasonably and whether they have an ability of category identification are checked.

(2) Feature words should have the ability to distinguish the target text from others: the intersection between feature word sets is theoretically empty so that the features of each category can be better distinguished and identified.

(3) The number of feature words should be moderate instead of being too large. And the feature words selected should be representative in clustering, aiming to improve the category discrimination.

(4) Feature word extraction should be implementable. The designed feature word extraction algorithm needs high temporal-spatial efficiency, and the actual extraction system should be of simple operation with low demands for operation platform resources.

3.2 Feature word measurement

Based on the traditional *TD-IDF*, the feature word measurements employed in this paper are listed as follows.

(1) Term frequency (TF): the occurrence frequency of a feature word t_i in document j, as shown in Formula (1), which normalizes the number of words to prevent it from being biased to long documents. The higher the TF value of a word, the more important it is, and the greater its weight.

$$TF_{ij} = \frac{n}{N} \tag{1}$$

Where n denotes the occurrence frequency of a feature word t_i in document j, and N stands for the total number of words in document j.

(2) Inverse document frequency (*IDF*): this variable can be obtained by taking the quotient of the total number of documents/the number of documents containing this entry, as shown in Formula (2). The smaller the number of

documents containing the entry, the greater the *IDF* value.

$$IDF_{ij} = \lg\left(\frac{D}{1+d}\right) \tag{2}$$

Where D represents the number of all documents in the corpus and d denotes the number of documents containing the feature word t_i. Then, the TF-IDF value of the feature word t_i can be solved through formula (3).

$$\mathrm{TF}-\mathrm{IDF} = TF_{ij} * IDF_{ij} \tag{3}$$

(3) Term position (*TP*): the position of feature words in the text. This position usually includes the text title, subtitle, abstract, keywords, main body, and other position terms; words in different positions or in different paragraphs of the same position term have different weights.

(4) Part of speech (*TA*): This variable is used to describe the role of a word in the context. For example, a word that describes a concept is called a noun, and a word that quotes this noun below is called a pronoun.

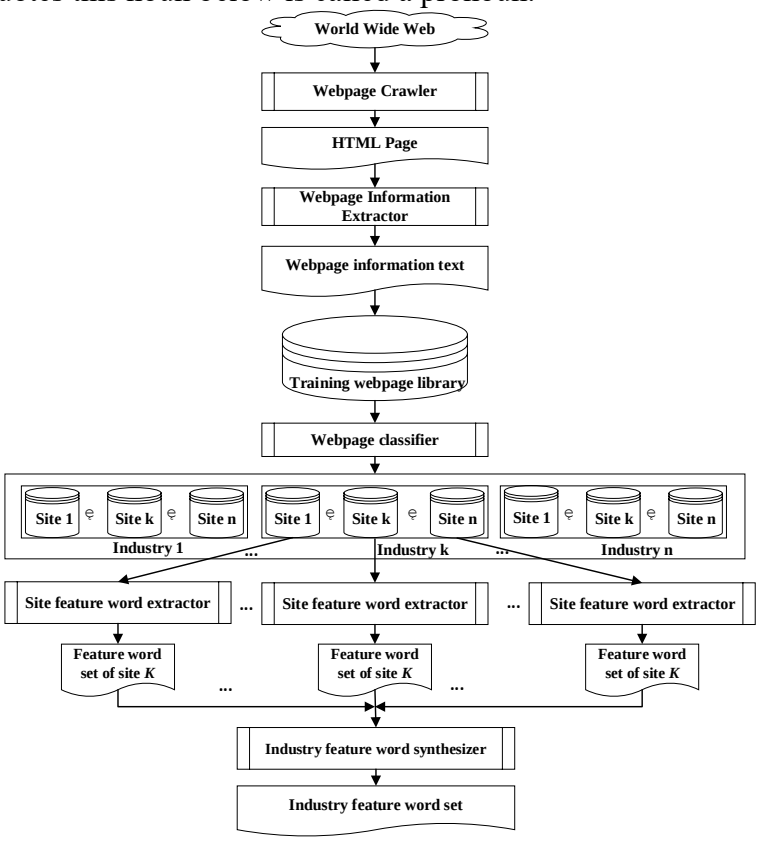

Fig. 1. Overall framework for text feature word extraction.

3.3. Feature word extraction framework

In this section, a framework for feature word extraction from the enterprise competitive intelligence text was designed from the perspective of practical application, as shown in Fig. 1. Then, the webpage grabber and webpage information extractor were used to grab training webpages and extract information from the grabbed webpages, respectively.

3.3.1. Site feature word extractor

The site feature word extractor completes the word segmentation, word frequency statistics, part-of-speech tagging, and word weight calculation of the text belonging to a specific site; and its realization is also one of the innovations of this study, with its framework shown in Fig. 2.

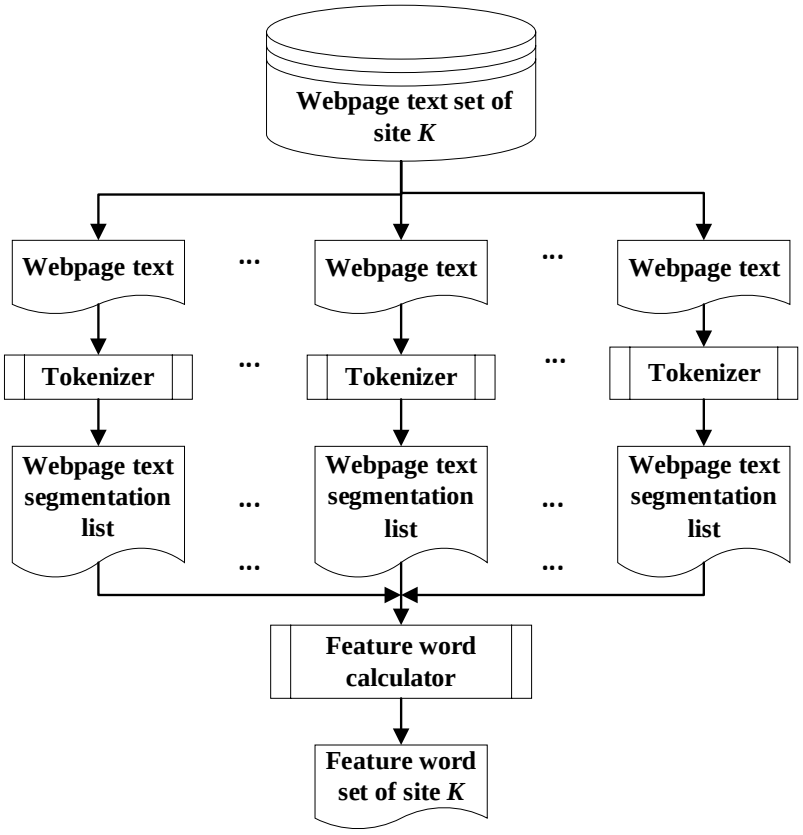

Fig. 2. The extraction framework of industry feature words.

The two most important modules in the framework of the site feature word extraction system are part-of-speech tagging and feature word weight calculation.

(1) *Part-of-speech tagging algorithm*

In this section, the *Sougou* lexicon and an alternative lexicon were mainly used to perform part-of-speech tagging of entries after segmentation. Among them, the alternative lexicon and *Sogou* lexicon had the same structure, both being composed of terms, word frequencies, and parts of speech. During the weight calculation of entries, the quantitative indexes represented by the above parts of speech will be studied in detail in the experimental part below. The formal description of the part-of-speech tagging is shown in Algorithm 1.

Algorithm 1 termAttrDecide(TS)

Input: the text set TS of website S;
Output: the term set TAS, in which the term has the TF, IDF, $Term\ Attribute$, and the $weight$

1. **for** each text TXT in TS
2. Term List TL = splitTermByIKAnalyzer(TXT);
3. Update the records' field TF;
4. **for** each term record T in TL
5. **if** T exist in the ***Sogou*** Term Library
6. TA = findAttribute(T, ***Sogou***);
7. **else if** T exist in the Candidate Term Library
8. TA = findAttribute(T, ***Candidate Term Library***);
9. **else**
10. TA = decideAttributeManly(T);
11. Update the **Candidate Term Library** by using the TA;
12. **endif**
13. **endfor**
14. **if** TA belongs to IA
15. insert TA into TAS;
16. **endif**
17. Update every record's field IDF;
18. **endfor**
19. **return** TAS;

(2) *Feature word weight calculation*

In order to reducing the dimension of feature vectors is the key to term weight calculation, since it can not only improve the time-effect ratio of calculation but also enhance the accuracy of term weight calculation due to the reduction of interference factors. The following formula (4) was used in this study to calculate the weight of word terms.

$$W_i = TF_i * IDF_i * ATTR_i \tag{4}$$

Where TF_i represents the word frequency of the term i, IDF_i is its inverse document frequency, $ATTR_i$ denotes its corresponding part-of-speech weight whose value will be expounded in Section 4.4.

3.3.2 Compounding of industry feature words

The similarity of documents is calculated based on industry feature words in the classification process. Therefore, it is necessary to "merge" the feature word sets of various sites belonging to the same industry to form the feature word sets of the corresponding industries. The strategy adopted in this study was the "generalized union" operation on the feature word sets of various sites in the same industry. The corresponding procedures are summarized as follows:

(1) Set *n* sites $S_1, S_2, \cdots, S_n$ under the industry *I*, and set the feature word set corresponding to site S_i as T_i, in which W_{ij} is the *j*-th feature word term of site S_i; And set IS as the feature word set of the industry *I*;

(2) If W_{ij} appears only once in the feature word sets of *n* sites, insert W_{ij} directly into IS;

(3) If there are several identical word terms named *W*, the following processing is done according to different circumstances: (a)Given different parts of speech, manually tag the parts of speech; (b) If the IDF is different, take the smallest one; (c) If *TF* is varied, take its mathematical expectation. The specific calculation method is: If the word frequencies of *k* words with the same name are $tf_1, tf_2, \cdots, tf_k$, the sum of their word frequencies is $sum = tf_1 + tf_2 + \cdots + tf_k$, and their respective occurrence frequency is $p_i = tf_i/sum$, so their mathematical expectation is expressed as $E = tf_1 * p_1 + tf_2 * p_2 + \cdots + tf_k * p_k$. Finally, the *E* value is taken as the final *TF* value of the word term *W*. The record with a structure of $IR = \{W, TF, IDF, TA, W_{ij}\}$ is inserted into IS;

(4) Sort the records in IS in non-decreasing order according to the weight, and select the top *M* with the largest weight according to the threshold *M* determined in advance as required to form the industry feature word set IS of the industry *I*. Form a feature vector, finally calculate the similarity between the webpage texts, and perform the text classification evaluation.

4. Experiment and Discussion

4.1 Experimental environment

In this study, the corresponding experimental scheme was designed to test the actual running effect of the "Text Feature Word Extraction" system under the following experimental environment: Intel (R) Core (TM) i5-2400 CPU @ 3.10 GHz CPU developed via Java language with the memory of 3 GB. In addition, the IKAnalyzer-3.2 tokenizer in Lucene open-source project was used.

4.2 Experimental data

In this study, metallurgy, medicine, tobacco, education, science and technology were selected as the target industries in the text classification test, and five sites were set in each industry. By manual screening, 2000 webpages were selected for each industry, of which 400 were selected from each site in the industry. A total of 10,000 webpages were chosen from the five industries.

4.3 Evaluation indicators

Without loss of generality, precision and recall commonly used in text classification were selected as the evaluation indicators for the quality of text feature words, as shown in Formulas (5) and (6) respectively.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Where *TP* represents the number of samples correctly classified into a certain category, *FP* and *FN* represent the number of samples wrongly classified into a certain category. In this study, the two indicators were comprehensively considered and evaluated by introducing the commonly used F-measure index, as shown in formula (7).

$$F - Measure = \frac{(\beta^2 + 1) * Precision * Recall}{Precision + Recall} \tag{7}$$

where β is an adjustment factor, which is used to give different weights to precision and recall. In this study, a compromise proposal was adopted to give the same weight to *Precision* and *Recall*, that is, β=1. At this time, F- measure is called the F1 value, as shown in formula (8).

$$\boldsymbol{F1 = \frac{2 * Precision * Recall}{Precision + Recall}} \tag{8}$$

4.4 Feature Selection and Sparsity Processing

Feature selection aims to identify and retain the most relevant features while discarding irrelevant or redundant ones. This process offers benefits such as improved model performance, reduced complexity, and enhanced interpretability. Besides, dealing with feature sparsity is crucial because it can lead to unstable

models and biased results. Thus, this paper will employ the following process for feature selection and sparsity processing. For feature selection, start by performing univariate statistical tests (mutual information) to measure the correlation between each feature word and the target text. Features with high scores are more likely to be informative. Then, recursive feature elimination involves training text classification models iteratively and eliminating the least important feature at each iteration, continuing until a desired number of features is reached and performance stabilizes. As for feature sparsity, we create new features by combining and transforming existing ones, thus then use regularization techniques to mitigate the impact of less informative features and encourage sparsity in classifier coefficients.

4.5 Experimental scheme

Determining the weight measure value of the target speech of word is the key to word set weight calculation of a site or industry by using Formula (4). As or the experimental dataset, 10 part-of-speech weight schemes as shown in Table 1 were selected and used to calculate the weight of the entries of various sites in the industry, and at the same time, the feature word sets of the industry were formed. The webpage texts of the experimental data in Section 4.2 were subjected to classification and test using the KNN and SVM algorithm separately in combination the feature word sets of various industries, and the experimental consequences are shown in Table 2. Finally, the best weight measurement scheme was selected, and the part-of-speech weight measurement in this scheme was adopted, followed by weight calculation of entries in the site via Formula (4) to form the feature word set of the industry. Based on the feature word set of the industry, KNN and SVM algorithm-based classification test was performed using the experimental data of metallurgy, medicine, tobacco, education, and science and technology. The test results were compared with the classification results obtained by the traditional *TF*IDF* feature weight calculation (see the detailed experimental results in Table 3 and Table 4, respectively).

Table 1

Weighting schemes for target parts of speech.

Weighting scheme	Noun	Verb	Adjective	Adverb
W1	0.85	0.06	0.02	0.02
W2	0.85	0.02	0.10	0.01
W3	0.65	0.16	0.09	0.02
W4	0.65	0.05	0.20	0.02
W5	0.55	0.02	0.25	0.02
W6	0.45	0.15	0.25	0.05
W7	**0.45**	**0.05**	**0.35**	**0.05**
W8	0.25	0.25	0.25	0.10
W9	0.32	0.25	0.34	0.10
W10	0.25	0.30	0.12	0.10

Table 2

Classification results under different weighting schemes for target parts of speech.

Weighting scheme	Average precision	Average recall	Average F1 value
W1	73.53%	75.38%	74.44%
W2	78.55%	77.69%	78.11%
W3	82.10%	83.58%	82.83%
W4	88.37%	89.12%	88.74%
W5	91.24%	89.28%	90.24%
W6	89.77%	89.81%	89.78%
W7	**94.77%**	**90.52%**	**92.59%**
W8	86.58%	90.08%	88.29%
W9	87.02%	90.12%	87.31%
W10	89.53%	89.13%	88.42%

Table 3

Comparison of classification effect between traditional *TF*DF* algorithm and the improved TF-IDF using KNN Classifier.

Category	Traditional *TF-IDF*			Improved *TF-IDF* based on part-of-speech		
	Precision	Recall	F1-value	Precision	Recall	F1-value
Metallurgy	76.89%	76.22%	76.55%	94.18%	92.24%	93.20%
Medicine	87.21%	87.70%	87.45%	94.35%	94.08%	94.21%
Tobacco	83.74%	84.34%	84.04%	95.65%	94.34%	94.99%
Science and technology	82.87%	82.55%	82.71%	96.19%	95.33%	95.76%
Education	80.70%	80.03%	80.36%	96.12%	95.14%	95.63%
Average value	82.28%	82.17%	82.22%	95.30%	94.23%	94.76%

Table 4

Comparison of classification effect between traditional *TF*DF* algorithm and the improved TF-IDF using SVM Classifier.

Category	Traditional *TF-IDF*			Improved *TF-IDF* based on part-of-speech		
	Precision	Recall	F1-value	Precision	Recall	F1-value
Metallurgy	80.26%	79.17%	79.71%	85.21%	86.71%	85.95%
Medicine	83.48%	85.20%	84.33%	88.71%	89.14%	88.92%
Tobacco	86.13%	83.49%	84.79%	87.62%	90.21%	88.90%
Science and technology	78.61%	79.36%	78.98%	84.35%	85.76%	85.05%
Education	82.45%	83.13%	82.79%	82.10%	81.92%	82.01%
Average value	82.19%	82.07%	82.12%	85.60%	86.75%	86.17%

4.6 Result analysis and Discussion

It could be seen from Tables 1 and 2 that the average F1 value was relatively low when "noun" was given too high weight in the three schemes W1–W3. However, when the weight of “adjective” was higher than that of “verb” or even

“noun”, the average F1 value was increased, and it reached the maximum in scheme W7. The reason is that, on the one hand, when the webpage text is segmented, the industry-specific terminology database has been used to segment the webpage text, and all the industry terms in the text have been specifically identified, for which part-of-speech tagging has been done, being basically tagged as "nouns". However, the industry-oriented classification method for webpages is adopted in this study, and industry terms often have a high degree of discrimination. Therefore, as industry-specific feature words, the part-of-speech weight of these words should be more important and discriminated than verbs and adjectives. After a comprehensive comparison and evaluation of the vast majority of webpages captured by the web crawling subsystem, on the other hand, it is found by combining the working experience of professionals in the field that nouns, verbs, adjectives, and other parts of speech can often reflect the theme and center of an article, and most of the content of the article describes them.

From the experimental consequences in Table 3 and Table 4, the improved *TF-IDF* feature selection method based on parts of speech increased the average F1 value, which characterized the classification effect, by 12.54% and 14.05% compared with the traditional *TF-IDF*, respectively. This result proves that the part-of-speech weight measurement scheme W7 is favorably feasible in industry-specified classification.

Compare the experimental consequences reported in Table 3 and Table 4, however, it can be seen that there are significant differences in the performance of feature word extraction in different domains and websites with different structures, especially for the categories such as Metallurgy and Education, which indicate that the generalizability of the proposed method in this paper needs further improvement. Moreover, the feature word extraction method proposed in this paper is mainly applicable to Chinese text information, and the above experiments were mainly conducted on texts in Chinese websites. However, the Chinese word segmentation during the feature word extraction process will have a significant impact on the final feature word extraction. Therefore, in summary, it can be concluded that the method proposed in this article has potential shortcomings in different domains, languages and websites with complex structures.

5. Conclusion

In this study, an alternative lexicon with the same structure as the *Sougou* lexicon was designed and implemented. Firstly, the part-of-speech tagging algorithm was designed, and the corresponding tagging subsystem was implemented. Secondly, the whole text feature word extraction system was completed by integrating the improvement of *TF-IDF*. Finally, the text feature word extraction system designed in this study was used to test the performance, which

was found to be good. Besides, the overall framework design and relevant algorithm design were performed for the "Text Feature Word Extraction" system, and the corresponding algorithm were experimentally verified, so as to demonstrate and implement the concrete system functions. The ideas put forward in this study still have a large improvement space and need to be comprehensively compared and analyzed with other methods on large-scale samples. However, the method proposed in this paper still has several potential drawbacks, so there are two directions at least need to further improve in future. One direction is to introduce deep learning technology to achieve automatic learning of features and model parameters. Another direction is to apply the method proposed in this article to websites with different domains, languages and complex structures, further enhancing the generalizability of the proposed method.

Acknowledgements

This work is supported by the Natural Science Foundation of Anhui Province (Grant No. 1908085QF283), the Doctoral Startup Research Fund (Grant No. 2019jb08), the University Synergy Innovation Program of Anhui Province (Grant No. GXXT-2022-047), the Open Research Fund of National Engineering Research Center for Agro-Ecological Big Data Analysis & Application, Anhui University (Grant No. AE202201), and the Scientific Research Projects Funded by Suzhou University (Grant No. 2021XJPT50, 2021XJPT51, 2022xhx004, 2022xhx099, 2021bsk016).

R E F E R E N C E S

[1]. F. Zhi, Z. Peng, J. Meng, Z. Yun, “Competitive Intelligence Needs Analysis of SMEs and Its Implication”, in Information Science, vol. 41, no. 2, 2023, pp. 36-43.

[2]. M. Li, W. He, B. Ding, “Research on the Construction of Competitive Intelligence Service Model for Small and Micro Enterprises from the Perspective of Digital Economy”, in Journal of Modern Information, vol. 43, no. 2, 2023, pp. 126-136.

[3]. H. Zhang, K. Wang, Y. Fan, H. Sun, “Construction and Simulation of Competitive Intelligence Situational Awareness Warning System for New Retail Enterprises”, in Journal of Intelligence, vol. 42, no.2, 2023, pp. 74-81.

[4]. R. Mooney, “Relational Learning of Pattern-match Rules for Information Extraction”, in Proc. of the National Conference on Artificial Intelligence, 1999, pp.328-334.

[5]. S. Soderland, “Learning Information Extraction Rules for Semi-Structured and Free Text”, in Machine learning, vol. 34, no. 1, 1999, pp.233–272.

[6]. J. Fonseca and A.ntónio Grilo, “WeCIM-Web Competitive Intelligence Methodology”, in Journal of Economics, Business and Management, vol. 1, no.1, 2013, pp.112-116.

[7]. X. Liu and C. Wan, "Beyond Bag of Words: A New Model for XML Keyword Query", in 2014 International Conference on Management of e-Commerce and e-Government, 2014, pp. 252-259.

[8]. Z. Qin, Y. Cong and T. Wan, “Topic modeling of Chinese language beyond a bag-of-words”, in Comput. Speech Lang. vol. 40, 2016, pp.60–78.

[9]. K., Irie, R. Schlüter and H. Ney, “Bag-of-words input for long history representation in neural network-based language models for speech recognition”, in Interspeech, 2015, pp.2371-2375.
[10]. X. Li, L. Ayixiamu, T. Yang and W. Xiong, “Handwriting Identification Based on Word Bag Model and Invariant Features of Geometric Moments”, in Computer Applications and Software, vol. 39, no. 7, 2022, pp.154-158
[11]. H. Liu, H. Tang, W. Xiao, Z. Guo, L. Tian and Y. Gao, “Sequential Bag-of-Words model for human action classification”, in CAAI Trans. Intell. Technol., vol. 1, 2016, pp.125-136.
[12]. J. J. Y. Kim, M. Urschler, P. J. Riddle and J. S. Wicker, “SymbioLCD: Ensemble-Based Loop Closure Detection using CNN-Extracted Objects and Visual Bag-of-Words”, in 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021, pp. 5425-5425.
[13]. K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval”, in Document retrieval systems, Taylor Graham Publishing, GBR, 1988, pp.132-142.
[14]. C. H. Huang, J. Yin and F. Hou, “A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method”, in Chinese Journal of Computers, no. 5, 2011, pp.856-864.