Research on Keyword Extraction Algorithm for Public Opinion Monitoring

Lin Xiaojie, Sun Jie

Abstract— **This paper improves the keyword extraction technology in the field of public opinion monitoring and prediction, and proposes a method based on new word detection algorithm and sentiment analysis. Through the application of this technology in the hot events of microblog, keywords can be effectively extracted, which helps to analyze the dynamics of public opinion, and find and guide the wrong and untrue public opinion in time. Through the new word discovery algorithm, we are able to identify new words emerging in social networks and extract them as keywords. This algorithm can automatically capture the hot and influential new words, which provides a more accurate and comprehensive information basis for public opinion analysis. In addition, sentiment analysis also plays an important role in keyword extraction. We used sentiment analysis algorithm to evaluate the sentiment tendency of keywords and help judge the positive and negative emotions expressed by public opinion. This method can automatically identify and filter keywords with emotional color in a large range, and improve the accuracy and efficiency of public opinion monitoring.**

Index Terms—**Keyword extraction, Sentiment analysis, New word detection, Public opinion monitoring.**

I. Introduction

With the rapid development of the Internet, the huge network database stores a large number of valuable text information resources. In this era of information explosion, however, how to quickly and accurately from the vast amounts of network text found valuable information has become a problem to be solved. Public opinion analysis[1] as an important field of study, through effective monitoring and analysis of the social public opinion, insight into the public's attitude, emotion and behavior tendency.

Weibo as one of the most popular social networking platform in China, offer the idea of a large number of users and sound. People can easily share their ideas and views on weibo, to interact with others and discussed. However, there are many different kinds of information on weibo, freedom of speech, which inevitably there are some bad information and rumors. Therefore, how to accurately and comprehensively analyze and monitor microblog public opinion has become a hot research topic, and many scholars and researchers have invested a lot of work to solve this problem.

This article aims to put forward an improved method based on TF-IDF algorithm[2], in order to improve the accuracy and practicability of the microblogging public opinion analysis. Firstly, we use crawler technology to crawl microblog content to obtain large-scale data sets. Then, we use the keyword extraction technology for micro blog this analysis, to extract the keywords. In order to better identify hot new words, we introduce a new word discovery algorithm, which can automatically capture hot and influential new words. At the same time, we also use sentiment analysis algorithm to evaluate the sentiment tendency of keywords to help judge the positive and negative sentiment of public opinion.

Through the improvement of the above algorithm, we can extract the daily keywords more accurately and rank them comprehensively. In this way, users can quickly and accurately understand the hot events of the day, and grasp the social dynamics in real time. At the same time, the algorithm application system can also carry out comprehensive monitoring and analysis of public opinion, help government departments and related institutions to understand the public opinions and emotions, so as to make better policies and decisions.

The research of this paper has important theoretical and practical significance in the field of public opinion analysis. The improved algorithm provides a powerful tool and method to improve the accuracy and efficiency of public opinion analysis. Believe that through our efforts, public opinion analysis field will usher in a more accurate and comprehensive research results, for the social public opinion guidance and optimization to provide better support to decision making.

II. Keyword Extraction Algorithm

A. TF-IDF

The TF-IDF algorithm[3-5] is a weighting technique commonly used in judging the similarity of two articles, information retrieval and data mining. TF is the Term Frequency, and IDF is the Inverse Document Frequency. The essence of the algorithm is to weigh the importance of a word to the target text source. The importance of a word increases in proportion to the number of times it appears in the document, but decreases in inverse proportion to the frequency of its occurrence in the corpus.

The main idea of TF-IDF is: if a word or phrase appears with high frequency TF in one article, and rarely appears in other articles, it is considered to have good category discrimination ability and is suitable for classification. The TF-IDF is actually: TF * IDF, Term Frequency, Inverse Document Frequency. TF said entries appear in the document d frequency. The main idea of IDF is that if the number of

Manuscript received September 03, 2023

Lin Xiaojie, School of Software, Tiangong University, Tianjin, China
Sun Jie, School of Computer Science and Technology, Tiangong University, Tianjin, China

documents containing term t is small, that is, if n is small and the IDF is large, then the term t has good class discrimination. If the number of documents containing term t in a certain type of document C is m, and the total number of documents containing t in other categories is k, it is obvious that the number of all documents containing t is n=m+k. When m is larger, n is also larger, and the IDF value obtained according to the IDF formula will be smaller, which indicates that the category discrimination ability of the term t is not strong. But in fact, if an entry in a class of the frequency in the document, then the entry is a representative of this class, the characteristics of the text should such entry gives higher weights to them, and selected as the text of key to distinguish it from other kind of document. This is the deficiency of IDF. In a given document, term frequency (TF) refers to how often a given word occurs in the document. This number is normalized to the term count to prevent it from being biased toward longer documents. The same word may have a higher number of words in a long document than in a short document, regardless of whether the word is important or not, so this system uses sentiment analysis algorithm to improve TF-IDF.

In the process of system analysis, the number of crawled microblog articles can be regarded as the number of all terms in the text, the frequency of occurrence of term w is counted in each microblog text, TF and IDF values are calculated according to the corresponding formula, and several keywords with higher TF_IDF values are selected as the final result.

The application steps of TF-IDF algorithm in extracting keywords in this system are as follows:

Calculate TF term frequency:

$$TF(w) = \frac{\text{The number of occurrences of the term w in the text}}{\text{The number of terms in the text}} \quad (1)$$

Calculate the inverse IDF document rate:

$$IDF(w) = \log \frac{\text{Total number of texts}}{\text{Number of texts with term w} + 1} \quad (2)$$

The number of texts containing the term w may be 0, so we add 1 to the denominator.

Calculate the TF_IDF value:

$$TF\_IDF(w) = TF(\mathrm{w}) * IDF(w) \quad (3)$$

The process of keyword extraction based on TF-IDF algorithm is as follows: Initialize the text word frequency dictionary, the default word frequency value is 0, traverse all the keyword sets of the text, calculate the frequency of each term in the set, that is, the frequency of occurrence in the document, as a benchmark value, apply TF and IDF calculation formula to each comment word segmentation set, calculate the word frequency and inverse document rate, and further obtain the TF_IDF value. For the word segmentation set corresponding to each microblog article, according to the TF_IDF value of the keyword, the larger value is the keyword of the article.

B. TextRank

TextRank algorithm[6] is adapted from Google's famous PageRank web page ranking algorithm, which mainly uses the graph model to extract the keywords in the article. The main idea of PageRank algorithm is: in the Internet, if a website is linked by other websites, it means that it has universal dependence and trust, so its ranking will be higher. In fact, the importance of each different web page is different, so when voting between web pages, the weight of the higher ranking web page should also be higher. Structure importance ranking the conversion relationship through the above thought, first given an initial weights to each web page, after many times of iteration calculation, according to the type of chain is smooth, the importance of the web site will eventually converge to a steady value.

Google's PageRank algorithm can be mathematically expressed as follows:

$$S_{V_\mathrm{i}} = (1-d) + d \cdot \sum_{j \in \ln(V_i)} \frac{1}{|Out(V_j)|} S_{V_j} \quad (4)$$

Among them: S_{V_i} denotes the importance of the ith website, d is the damping coefficient, generally set to 0.85, is used to prevent the importance of 0, $\ln(V_i)$ is pointing to the case of A collection of web pages, $\ln(V_i)$ is the set of links to web pages of web j. $|\mathrm{Out}(\mathrm{V_j})|$ is the number of elements in the collection. PageRank needs to iterate the above formula many times to get the result.

TextRank formula on the basis of PageRank formula, to chart the importance of introducing the concept of weight:

$$S_{V_\mathrm{i}} = (1-d) + d \cdot \sum_{j \in \ln(V_i)} \frac{W_{ji}}{|Out(V_j)|} S_{V^j} \quad (5)$$

W_ji denotes the weight of the edge from V_i to V_j.

The TextRank model is used on the basis of the constructed graph model. When doing keyword extraction, the initial weight is set to 1, and the importance of each word can be obtained after continuous iteration and stability, so that the threshold can be set for keyword extraction.

III. KEYWORDS EXTRACTION ALGORITHM FOR MICROBLOG TEXT

In microblog public opinion analysis[7], it involves the training of historical data and the extraction of keywords from current microblog data. The recent progress in keyword extraction research rapidly, extraction method is divided into supervised and unsupervised method. Task restructuring belongs to the supervision method, to extract the keywords can be converted to binary classification problems, namely whether candidate word is keyword monitoring method, and bayesian classification method has a simple SVM, decision tree, etc. At present, the keyword Extraction system based on supervised learning mainly includes GenEx system which adopts C4.5 decision tree and genetic algorithm, and KEA(Keyphrase Extraction) which adopts discrete Bayesian method Algorithm) system and Maui system, which is based on KEA and uses multi-feature and bagged decision tree method. However, supervised methods depend heavily on the training set, and the quality of the training set often determines the effect of keyword extraction, and there is no excellent Chinese training set at present. In view of the brevity and diversity of microblog data[8], we use an unsupervised method to extract keywords. Unsupervised learning methods include probability statistics of document content, calculating word weight to sort and extract keywords, such as TF-IDF algorithm, and unsupervised method TextRank based on graph structure.

However, this article studies the microblogging public opinion analysis problem, so the hot issues of words in the corpus is more likely to become the current key words in the microblog information. Therefore, this paper designs a microblog keyword extraction algorithm based on sentiment analysis technology[9].

A. Word segmentation Algorithm based on new word detection

With the rapid development of information technology, it is more and more common for people to express their personal views through microblog. However, most of the microblog content is casual, very colloquial and non-standard, which will produce a lot of network new words. Such as "diaosi", "fat house", "Ollie" and so on. In natural language processing[10], the emergence of new words has brought many adverse effects on the construction of sentiment lexicon, short text orientation analysis, Chinese word segmentation and many other aspects, reducing their efficiency. So how to identify new and efficient become an very important in the process of natural language processing tasks.

In the process of new word discovery in microblog data, traditional rules and statistical methods can not solve the problem of data sparsity in microblog data[11], and the simple rule-based method has poor portability, so that the effect of new word discovery is not ideal. In view of the high probability of new words appearing in hot events of microblog, this paper used a new word discovery algorithm to improve the original algorithm, so as to be able to find and extract new proper nouns and network terms.

At present the definition of new words is not accurate, the unknown word in this article with the new words, that is to say new words in this article is not in the old words in the dictionary.

Definition 3.1: Word split word_split, a word split is a segment that connects multiple consecutive Chinese characters of the document into a segment.

Definition 3.2: aggregation refers to the mutual information of two sub-fragments in a word fragment. When a word fragment exceeds two words, the minimum value of the mutual information of all sub-fragment pairs is selected as the aggregation degree of the word fragment.

Definition 3.3: Adjacency Entropy. Adjacency entropy refers to the information entropy between a word segment and its adjacent single character. The information entropy between a word segment and its left single character is called the left adjacent entropy left_entropy, and the information entropy between a word segment and its right single character is called the right adjacent entropy right_entropy.

Definition: 3.4 weights word_weight into word, word formation probability refers to a word fragments is an independent of the weight value of Chinese vocabulary, in a document, if a word fragments into word weight is higher, its in the current document to the higher the probability of a word.

The calculation formula of the segmentation algorithm based on new word detection[12] is as follows.

$$P_i = Aggregation_i \cdot MIN(Left - Entropy_i, Right - Entropy_i) \quad (6)$$

B. Improved TF-IDF algorithm based on sentiment analysis

The main idea of Senti-TFIDF algorithm is that if a word in a microblog has high attention compared with historical events, and the frequency of inverse microblog is low in the current test corpus, such words can be regarded as words that attract other users to pay attention to the microblog, and such words are more suitable as microblog keywords.

The Senti-TFIDF algorithm is improved by integrating the sentiment analysis algorithm with the keyword extraction algorithm, which is an algorithm that can extract representative keywords in microblog. In the process of system analysis, the use of specific numerical emotional score again and TF-IDF algorithm to calculate the score of final good keywords, namely

$$Senti - TF - IDF = TF - dic * IDF - dic * abs(senti - dic) \quad (7)$$

Aiming at the hot information of microblog, this system uses the sentiment polarity analysis[13-15] based on the dictionary to judge the positive and negative polarity of the microblog text content. From the sentiment location, it extracts the corresponding keywords, counts the word frequency of all keywords, and analyzes the representative microblog keywords with practical significance.

This method mainly calculates the emotional polarity score of the sentence according to the normal grammar rules and the emotional dictionary, and delimits the corresponding score interval. Generally, greater than 0 indicates positive sentiment, less than 0 indicates negative sentiment, and the size of the score reflects the tendency degree of the corresponding sentiment.

The accuracy of sentiment analysis results based on dictionaries is proportional to the richness of the sentiment dictionary set. If the key components of the text content to be analyzed appear infrequently in the sentiment dictionary, the sentiment analysis of this text will produce errors. Therefore, the project relies on the third-party library SnowNLP. Emotional volatility secondary analysis of the text, the results of the analysis and systematic analysis of the results, if the difference is too big, is judged to be system error, the system keep SnowNLP correct analysis result, will this kind of micro blog at the same time, according to the emotional polarity to join the system of the corpus, the next time the analysis of the text type, The error will be compared with the analysis of the primary system. Therefore, the system for error diagnosis and correction process, is also the system itself corpus, sentiment analysis model the process of self-improvement.

Sentiment analysis accuracy on the one hand, rely on the result of the emotional dictionary, on the other hand is dependent on the analysis method. The sources of the sentiment lexicon used by the system are: The model training dictionary of SnowNLP, Baidu stop words list and HowNet sentiment analysis dictionary form the basic dictionary of sentiment analysis of this system, including stop words list, derogatory sentiment dictionary, positive sentiment dictionary, negative word dictionary and adverb dictionary of degree. At the same time, the Jieba word segmentation corpus is used as the basic corpus. Construct negative and positive corpus training database. In dictionary based on human development, according to the common features and experience of the object to be analysed and add some specific corpus ingredients, such as the current popular, such as hot words, feelings will expand the dictionary after a system analysis dictionary library.

With the help of SnowNLP library, the system obtains the sentiment score list of the microblog article, calculates the deviation of the two scores corresponding to the same comment in the two sentiment score lists, and sets the deviation threshold interval (0.3, 0.6). The sentiment score in the deviation interval indicates that the sentiment analysis result of the algorithm is very different from the result of SnowNLP. It shows that when the algorithm calculates the sentiment score based on the current dictionary, the error is too large. The main reason is the lack of the current system corpus. The system retains the sentiment score results of SnowNLP.

In terms of keyword extraction, it is not simple to extract keywords, but to use clustering method to classify synonyms and similar words, and finally get keyword labels, which can avoid keyword redundancy and is more efficient. In terms of keyword screening, the Senti TFIDF algorithm is used to screen, which can meet the characteristics of microblog event topics with a wide variety and new words emerging. The relevant microblog is a small part compared with the microblog as a whole, so the topic keywords with distinguishing ability in microblog information exist in a small part of microblog, and the microblog new words can also be retained.

A statement, the main content of a document, even through the key words in this sentence, the main content of the text of the master, is the key for the understanding of key words in the text. Keywords affect the semantics and the overall sentiment of the text, and different keywords have different weights in the sentence. The most obvious feature is the word frequency. The preposition without emotional color appears frequently in Chinese, but it has no meaning for the extraction of keywords. Therefore, in the design of this system, the extraction object of keyword words is the result set after removing stop words and word segmentation.

IV. Experiment Result and Analysis

The evaluation index is the F1 value of keyword extraction, Precision represents the accuracy rate, Recall represents the recall rate, the number of keywords with the same prediction result as manual annotation is TP, the total number of predicted keywords is NPrediction, and the total number of keywords in the test set is NTest. The specific formula is as follows:

$$Precision = \frac{TP}{Nprediction} \tag{8}$$

$$Recall = \frac{TN}{Ntest} \tag{9}$$

$$F1 = \frac{Precision * Recall * 2}{Precision + Recall} \tag{10}$$

The system uses the Senti-TFIDF algorithm to compare and analyze 967 microblogs. Since microblog articles are relatively short, this experiment set the number of extracted keywords to 5, resulting in 4835 manually annotated keywords, 4835 extracted keywords by TF-IDF algorithm, 4835 extracted keywords by Text-Rank algorithm, and 4799 extracted keywords by Senti-TFIDF algorithm. The comparison of the experiments is shown in Table 1.

Table1 Comparison of experimental results of keyword extraction

Algorithm	Precision	Recall	F1
TF-IDF	67.29%	67.29%	67.29%
Text-Rank	70.37%	70.37%	70.37%
Senti-TFIDF	81.81%	80.71%	81.26%

The experimental results show that the Senti-TFIDF algorithm in this paper is better than TF-IDF algorithm and Text-Rank algorithm in precision and recall rate, and can also show superiority in specific keyword extraction.

Beijing, March 11 (Xinhua) -- Hu Chunhua, a member of the Political Bureau of the Communist Party of China (CPC) Central Committee and head of The State Council Leading Group for Poverty Alleviation and Development, chaired a plenary meeting of The State Council Leading Group for Poverty Alleviation and Development in Beijing on March 11. He stressed that the coronavirus should conscientiously implement the spirit of General Secretary Xi Jinping's important speech at the symposium on poverty alleviation, push forward the fight against poverty with greater determination and stronger efforts, strive to overcome the impact of the epidemic, ensure the completion of the goal and task of poverty alleviation on schedule, and ensure that the building of a moderately prosperous society in all aspects.

The extraction results of the Senti-TFIDF algorithm are as follows:

Coronavirus:1.06740040867
Poverty alleviation:1.9550565763
The State Council: 1.53450906639
Epidemic: 1.73359335274
Well-off society: 1.27613089637
TF-IDF algorithm of extracting results as follows:
Poverty alleviation: 0.1416458608
Politburo member: 0.141645860888
State Council: 0.141645860888
Epidemic: 0.121765808608
Well-off society: 0.141645860888
The extracted results of the Text-Rank algorithm are:
Poverty alleviation: 1.1832186494149877
State Council: 1.1425570196139663
Well-off society: 1.152209632029
General Secretary: 1.1522096320296389
Epidemic: 1.1832186494149877

Senti-TFIDF algorithm can effectively extract "coronavirus", "poverty engines", "disease", "well-off society" and other important keywords, and the TF-IDF algorithm and Text-Rank is ignored the key word "coronavirus", the effect of the two algorithms are easy to be affected by the noise of the complex information in weibo. Senti-TFIDF algorithm, on the other hand, tend to "coronavirus" so sensitive, able to attract men's event keywords extracted.

Therefore, it is found that the Senti-TFIDF algorithm can better extract the sensitive words in the event and better describe the event.

V. CONCLUSION

This paper studies the improved bidirectional maximum matching algorithm to complete Chinese word segmentation. The TF-IDF algorithm is used to complete the keyword extraction task. Using a new word discovery algorithm and emotional analysis algorithm of TF-IDF algorithm was further improved. In the system in practical use, very good solution to the diversity of the Chinese semantic words brought by the complexity of emotional polarity. However, there are still some challenges in keyword extraction. Firstly, it is still difficult for traditional algorithms to accurately identify words with similar semantics but different expressions. Secondly, keyword extraction for long texts is also a challenging task, which needs to consider the context information while maintaining efficient performance. In addition, the selection of evaluation metrics and datasets also needs to be continuously studied and improved in order to evaluate the performance of different algorithms more accurately.

Future research can be carried out in the following aspects: First, further improve the accuracy and efficiency of the keyword extraction algorithm. Such as combining model and knowledge map technology to improve the quality of extraction of key words. Secondly, the keyword extraction method under multi-modal data is explored, and the data such as image and audio are combined for more comprehensive text understanding. In addition, keyword extraction is combined with other natural language processing tasks, such as text classification and information retrieval, to improve the overall text processing ability.

REFERENCES

[1] 夏火松,甄化春.大数据环境下情感分析与决策支持研究文献综述[J].情报杂志,2015,34(2): 1-6.

[2] Alshuraiqi H S, Improved Term Frequency Inverse Document Frequency (TF-IDF) method for arabic text classification[J]. International Journal of Advanced Trends in Computer Science and Engineering, 2020, 9(5): 6939-6946.

[3] 胡亮,夏磊,李伟.Design and Implementation of a Keyword Extraction System Using Optimized TF-IDF%基于改进TF-IDF算法的关键词抽取系统[J].厦门理工学院学报,2017,025(005):67-72.

[4] 张建娥.基于TFIDF和词语关联度的中文关键词提取方法[J].情报科学,2012(10):110-112+123.

[5] Abu-Errub A . Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements[J].International Journal of Computer Applications,2014,93(6):40-45.

[6] 徐立.基于加权TextRank的文本关键词提取方法[J].计算机科学,2019,46(z1).

[7] 许鑫,章成志.互联网舆情分析及应用研究%Research on Application and Analysis of Internet - Mediated Public Sentiment[J].情报科学,2008,026(008):1194-1200,1204.

[8] 余本国.基于Python的大数据分析基础与实战[M].北京:中国水利水电出版社,2018.

[9] 王非.基于微博的情感新词发现研究[J].软件,2015,427(11):12-14.

[10] [美]Steven Bird,Ewan Klein,Edward Loper.Python自然语言处理[M].北京:人民邮电出版社,2014.

[11] 霍帅,张敏,刘奕群等.基于微博内容的新词发现方法[J].模式识别与人工智能,2014,27(2):141-145.

[12] 刘伟童,刘培玉,刘文锋等.基于互信息和邻接熵的新词发现算法%New word discovery algorithm based on mutual information and branch entropy[J].计算机应用研究,2019,036(005):1293-1296.

[13] 李锐, 张谦, 刘嘉勇, 基于加权word2vec的微博情感分类[J]. 通信技术, 2017, 50(03):502-506

[14] Yanqiu L, Zekun D, Chinese Movie Comment Sentiment Analysis Based on HowNet and User Likes[J]. Journal of Physics:Conference Series, 2019, 1229(1):2-9.

[15] 李家俊, 基于多特征加权和混合网络的文本情感分类算法研究[D].西南交通大学,2021.

Lin Xiaojie Graduate student, School of Software, Tianjin Polytechnic University. She graduated from Ludong University with a bachelor's degree, and won academic scholarships for three consecutive years during her undergraduate study.

Sun Jie Associate Professor. He is currently at School of Computer Science and Technology, Tianjin Polytechnic University Network department teachers, long-term engaged in computer-related professional teaching, guidance and related fields of scientific research. Presided over and participated in a number of vertical and horizontal scientific research projects, focusing on the development of e-government application system, The promotion and transformation of big data analysis and system realization have achieved high economic benefits and social effects.