Research And Implementation Of Keyword Extraction Algorithm Based On Professional Background Knowledge

Xuekun Zhang[1,2,3], Jing An[1,2,3], Wen Liu[1,2,3]
[1] Information Engineering School, Communication University of China
[2] Key Laboratory of Acoustic Visual Technology and Intelligent Control System, Ministry of Culture
[3] Beijing Key Laboratory of Modern Entertainment Technology
Beijing, 100024, China
zxk2015@cuc.edu.cn

Abstract: With the development of Internet, Data information is growing at an explosive rate. With the era of big data coming, information social value can only be reflected by people's utilization. In the vast amounts of data, keywords as relatively concise summary of the documentation, its can provide efficient information management methods. Keyword extraction technology (KET)can help people get the data information accurately and quickly, so KET is widely used in the information management system. According to the study of keyword extraction method recent years, the classic TF - IDF algorithm and TextRank algorithm were studied in this paper, TextRank algorithm improved and innovated based on the idea of TF-IDF algorithm, the process of TextRank improved algorithms designed and experiments proved the accuracy of the keyword extraction of the improved TextRank algorithm.

Key words: TF-IDF algorithm, TextRank algorithm, Keywords extraction, Professional background knowledge

I. INTRODUCTION

1.1 Background and significance

Keyword extraction technology is a effective way to convenient to get and use huge amounts of data, improve the efficiency of acquiring related resources and reduce the cost of acquisition too, People can access valid data information by reading and searching for keywords quickly. Keywords have the following functions: As a directory of retrieval, convenient users to find the relevant content of the targets. As a abstract of the full text, users can read and determine the requirements of resources by keywords. Search engines search by keyword, which can greatly improve efficiency, making search results more accurate[1]. Keyword extraction technology is one of the core technologies in the field of information processing, At the same time, it is also essential for information processing. It is widely used in information retrieval, text classification, automatic summary, etc.

1.2 Research status

Keyword extraction technology is an indispensable part of the natural language processing process, and it's the prerequisite of text classification, text clustering, automatic summary generation and information retrieval in data mining. Keyword extraction is mainly through the computer language to establish a given model to deal with the text, The result that is easy to identify and can represent the topic of the article through data processing is the keyword of the text.

Research of keyword extraction methods for text processing start and develop earlier in other countries than China, as the different between Chinese and English, its become more difficult to study. At present, research methods of extraction technology for text keywords in the world mainly includes statistical information, machine learning and shallow semantic analysis, etc[2]. TF-IDF algorithm is a kind of commonly used for information retrieval and text mining weighting algorithm was put forward by Stltion, and Mihalcea and Taraun based on Google's PageRank algorithm research and improvement, proposed a keyword extraction method based on graph model of unsupervised - TextRank algorithm, with the vigorous development of the computer intelligence technology, in the form of machine learning to extract text keywords is becoming a hot spot of current research direction.

This paper mainly studies the text keyword extraction algorithm based on professional background knowledge, on the basis of further study on the theory of key words extraction using background knowledge，the shortcomings of the current key words extraction technology are analyzed. Based on TextRank algorithm, the key words extraction algorithm was improved and a text keyword retrieval system based on football expertise was designed.

II. KEYWORD EXTRACTION ALGORITHM

2.1 Classic TextRank algorithm

The classical TextRank algorithm is an important algorithm for the extraction of key information extracted by Mihalcea and Tarau in 2004, this algorithm is inspired by the famous PageRank algorithm and its widely used in the extraction of key information.

The steps for the TextRank algorithm are as follows:

(1) Identify the nodes in the text and build the map;

978-1-5386-1937-7/17/$31.00 ©2017 IEEE

(2) Assign the same initial value to the importance of each vertex in the graph;

(3) Using a specific formula for each node value of the iterative calculation and update;

(4) Set a threshold in advance, when a certain condition, you can think that the value of the node has been convergence and stabilize, then stop iteration;

(5) Based on the value calculated by each node, and the importance of the node to sort.

Classical TextRank algorithm has word-oriented TextRank algorithm and sentence-oriented TextRank algorithm two, Two kinds of algorithms are similar, The main difference is that the building side.

The word-based TextRank algorithm treats a particular word in the text as a basic element, a node, And then through the artificial definition of the text between the various words of the link, It is possible to establish a directional or non-directional with or without weight between words, Thus building a text within the scope of the word network[3]. For the construction of interfaith links, At present, there are two main methods of judging the relevance of words, One is to build the text within the word relationship, The other is through external tools, The similarity between the two words is judged by the relevance of the part of speech and the meaning of words.

The sentence-based TextRank algorithm treats an independent sentence as a vertex, If the similarity between the two sentences exceeds a certain threshold, It will establish a connection between the two vertices, The similarity between sentences is the weight of the side, The similarity between the two sentences is obtained by calculating the similarity of the pairs of words between the two sentences[4].

TextRank algorithm of extract keywords has the following advantages:

(1) The computation time of the algorithm iteration calculation is short and the running speed is fast.

(2) No need for additional supervision training, it is an unsupervised algorithm.

(3) No need adjust the specific text for the algorithm, and achieve the independence of the text theme.

2.2 Improved TextRank algorithm

Traditional TF-IDF algorithm and TextRank algorithm have some limitations on the method of extracting text keywords. In the process of text feature extraction, as the supervised TF-IDF algorithm only considers the overall situation of the document, and extraction only considered the macro-feature. None supervision TextRank algorithm is only considered a single document in the relationship between words and words, this just a relatively microscopic feature extraction method, and the TextRank algorithm doesn't show represented of the weight values in different words[5]. If the two algorithms are used in conjunction with the method of keyword extraction, it will make up for the drawbacks of these two methods alone.

In the TextRank algorithm, Artificially giving each word the same initial weight, Set the link between the two words, In accordance with the vote, Passing the weight to the next vertex through the connected edge, Instead of this random model, vote on average and jump to the next vertex[6]. For large-scale data corpus for text processing, Different topic content corresponding to the words in the document is not the same degree of importance, So in the improvement of TextRank algorithm for the extraction of keywords, Giving each vertex an unbalanced weight at the beginning, By this setting the initial value is different, And then to the iterative way to calculate the TextRank, So that the results of the formation of each vote more targeted, More close to the main content of the document extracted from the results. Because in the use of TextRank algorithm to take into account the vertices in the figure that the weight of each word, So before the TextRank algorithm to extract keywords, We first on the data set of documents for each word weight calculation, After which the weight value of the word is entered as TextRank for each word, Proceed to the next step.

The specific formula as follows (2-1):

$$W(V_i)=(1-d)W'(V_i)+d*W'(V_i)\sum_{j\in In(V_i)}\frac{W_{ji}}{\sum_{k=V_j}W_{ji}}W(V_j)\ . \quad (2-1)$$

W (Vi) represents the weight of the current vertex. In the technique of extracting the keywords for the overall document, choose the commonly used method TF-IDF algorithm to calculate the weight of each word. But in the process of calculating the document weight, the value of TF-IDF algorithm is relatively small, once enter iterations in TextRank, it will affect the effect of the experiment. So TF-IDF value should be normalized processing, and enter the weight calculation in the TextRank algorithm.

The normalized formula as follows (2-2):

$$W'(V_i)=tf_{ij}*\left[\log\left(\frac{\sum_{i=1}^{M}nt_i}{nt_i}\right)\right]^2\ . \quad (2-2)$$

The main steps of the algorithm:

(1) The data set for cleaning, remove the extra symbols, pictures, etc., saved as txt text form.

(2) Participle processing of documents, add word segmentation dictionary, custom stop word.

(3) Count the words, calculate the value of TF, IDF and TF*IDF.

(4) Normalized the weight of the words.

(5) The weight of each word is mapped to the corresponding words in the document, And then each document with TextRank algorithm for processing, calculate the final weight value for each word.

Finally, the ranking of key words extracted by this method according to the importance, and then the results are obtained.

III. EXPERIMENTAL DESIGN

3.1 Access to background knowledge

Aiming at the problem of keyword extraction based on professional background knowledge, We use the background knowledge data set is football field of football content based on football content extracted from the unstructured text type data, In the background knowledge extraction has a certain degree of particularity. Experimental background text set of corpus data from the Internet data crawl, In the experiment, we use a unified football topic background text for comparative experiments, Through the CCTV, NetEase, Sina, Sohu, Tencent five portal website sports section of the page to crawl, Collected since 2017 on the football game video coverage of nearly 100,000, The video content is converted to text data as a data set for experimentation.

In this paper, because the experimental data is extracted from the web page using crawler software, Data format is not standardized, Content in a variety of forms, If the direct use of the experiment will seriously affect the experimental results, So the data must be used before the text preprocessing, The quality of the text processing will directly affect the results of the experiment is accurate. Preprocessing the text of the data set is to denoise the corpus, Remove the picture in the document, Remove the stop word in the document, And then the content encoding format for a unified specification, So that the form can facilitate the further processing of text mining. One of the most important is to word segmentation and disable word filtering, In the text preprocessing, First of all text for word processing, And then disable the word filter, Get the candidate keyword collection as the input data for the keyword screening phase. Deactivation of word filtering According to the deactivation of the word list on the word after the collection of entries to filter, Filter out irrelevant items to reduce the size of the collection[7].

In the keyword extraction, we use the data mining in the classification of key words to select the candidate words, the candidate words are divided into two categories: keywords and non-keywords. Classification technology in recent decades in the data mining has a very wide range of applications, classification is a set of features with the data mapped to the already defined, non-overlapping categories, the essence is to use the model to determine the type of unknown data type[8]. The classification process mainly includes the training stage and the classification stage, is a supervised learning process. The training phase focuses on the construction of classification models based on data categories, The training set consists of multiple data tuples and their already identified type labels, The classification phase is the stage of using the trained model to predict the data of the unknown class label, The classification model first predicts the set of types, And then use the validation data set to test the classification effect, Finally, the prediction result is compared with the class number determined by the data itself to get the classification accuracy of the model.

The acquisition of professional background knowledge in this algorithm is mainly composed of the following steps:

(1) Using the improved TextRank keyword extraction algorithm for the extracted keywords, get the most important words in the text.

(2) Acquired important words as a query word, query the important word in the football professional background knowledge data, and get the relevant text set with high correlation.

(3) Extracting specific data information in the relevant text, build the football professional background knowledge of the text structure.

3.2 Keyword characteristics calculation

The MUC (Message Understanding Conference) is a Message Understanding meeting. It is one of the most important conferences on information extraction research, and the concept, model and technical specifications defined by MUC play a leading role in the whole information extraction field[9]. MUC is responsible for the organization of the different units from around the world news understanding system for serial evaluation activities, according to certain evaluation index system of evaluating results, one of the main indicators are accuracy, Precision and Recall rate (Recall).

Accuracy is also known as the accuracy rate. In the information extraction, the accuracy rate is the ratio of the correct information in the extracted information, can describe the credibility of the extracted information.

The calculation method shown in equation (3-1):

$$P = \frac{|K_1 \cap K_2|}{|K_1|} \quad . \qquad (3\text{–}1)$$

Recall rate is also called recall rate. In the information extraction, the recall rate calculates the ratio of the information being correctly extracted, which reflects how much information is properly extracted.

The calculation method shown in equation (3-2):

$$R = \frac{|K_1 \cap K_2|}{|K_2|} \quad . \qquad (3\text{–}2)$$

In practical applications, the accuracy rate precision and recall rate recall between 0 and 1 (maximum1), and the accuracy and recall rate are often interdependent, there is some kind of equilibrium relationship. When the extracted examples are less, there is a high accuracy, but the recall rate is low. When the recall rate is high, often get more correct examples.

The total number of instances correspondingly extracted at this time is also increased and the accuracy rate becomes low[10]. When these two indicators need be considered, use F-measure measured. F-measure is another standard commonly used in the field of information retrieval.

The calculation method shown in equation (3-3):

$$F = \frac{2 * P * R}{P + R} . \qquad (3\text{–}3)$$

Among them,K1 represents the set of keywords extracted by the algorithm,K2 represents artificial annotation of the keyword ,K1 ∩ k2 represents the algorithm extracts the correct keyword collection, accuracy (precision), recall rate (recall, R) and F measurement (F-measure, F).For the performance evaluation of the keyword extraction algorithm, Usually marked by artificial keywords as the basis for judging the correct or not, using the three areas of information retrieval indicators commonly used indicators for evaluation.

3.3 Keyword extraction design

The design of information processing module based on keyword extraction includes word segmentation service, keyword ordering, index service, cache, dictionary, improved TextRank algorithm engine, etc.

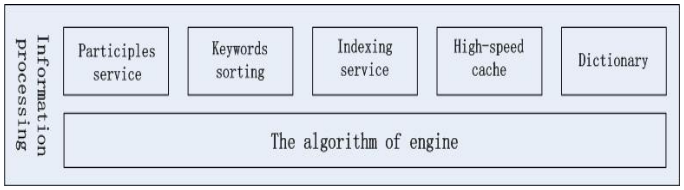

Figure 3-1 Keyword extraction module framework design

Keyword search module design includes database connection, query application, system management, Database covers different types, contains different types of file types, query application has a variety of query functions. System management includes user management, rights management, log management, etc.

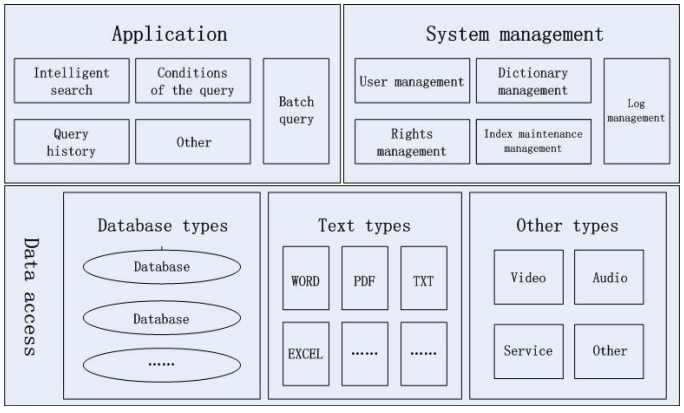

Figure 3-2 Keyword search module framework design

IV. Experiment Results Analysis

4.1 Comparison of three algorithms

Traditional TF - IDF algorithm with TextRank algorithm for text keyword extraction has certain limitation, supervised TF - IDF algorithm only considers the overall situation of the document, the document is macro feature extraction, and unsupervised TextRank algorithm only considers the relationship of the words in a single document, is the micro keyword extraction method. This paper analyzes the TF - IDF keyword extraction algorithm and the shortage of TextRank, combines two algorithm used for the extraction of keywords, the original words of TextRank algorithm, according to its importance in the document, to give different weights, and then iterate calculation, finally will extract the weight value of keywords ranking results are obtained.

In order to comprehensively evaluate the effectiveness of the proposed algorithm, Experiment in this paper using TextRank algorithm based on words, TF - IDF algorithm and three ways respectively in this paper, we study the improved TextRank algorithm for football professional background documents keyword extraction test, and adopts evaluation index to evaluate the results in section 3.2.

The results of the experiment are shown in Figure 4-1:

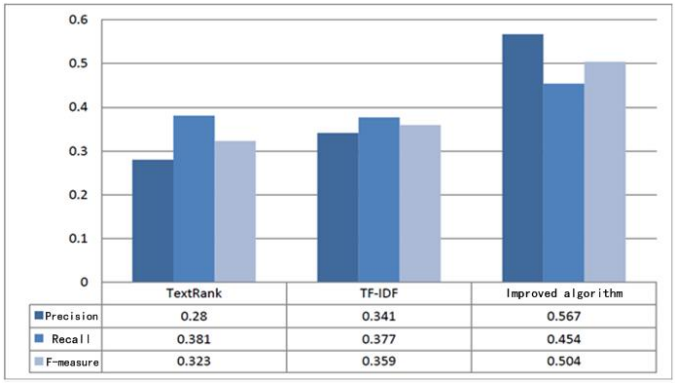

Figure 4-1 Comparison of three keyword extraction algorithms

It can be seen from the figure that the improved keyword extraction algorithm in this paper is better than the other two algorithms in terms of accuracy, recall rate and F-measure.

From the figure 4-1 can be seen, Based on TF-IDF algorithm to improve the effect of TextRank algorithm to extract keywords is better than TextRank algorithm, the main reason is that the TextRank algorithm calculates the weight of the vertex word in the form of the graph model to achieve the weight calculation of the keyword, but as the weight value of the word itself has been set to the same value, so that influenced the keywords ranking, and based on TF - IDF algorithm improved TextRank algorithm ideas, give the word itself different weights, thus accurate keywords ranking.

4.2 Improved algorithm extraction performance analysis

Algorithm in this paper using football professional background knowledge to generate new characteristics, and combining the word frequency and word length, the position of

the word, and other characteristics of the classification of the word to extract text keywords, optimizes the effect of the extraction of keywords, and used as a contrast experiment TextRank algorithm based on words and TF - IDF algorithm by using only football professional background knowledge within the text information, thus causes the keyword extraction effect is not ideal. The results of experiments show that the improved keyword extraction algorithm based on background knowledge in the accuracy and recall rate and F-measure on index is better than the traditional algorithm, illustrate that the proposed improved keyword extraction algorithm is effective and feasible.

V. CONCLUSION

Based on the further study of TF-IDF algorithm and TextRank algorithm on the basis of aiming at the shortcomings of the TextRank algorithm is improved, completed in TF-IDF algorithm thought TextRank algorithm is improved based on keyword extraction algorithm. Experimental results show that the accuracy of the improved algorithm to extract the keywords than TextRank algorithm has greatly promoted, Which also improves the extraction accuracy of keywords based on professional background knowledge, And also improve the algorithm has a good stability and achieve the expected improvement keyword extraction algorithm, improve the purpose of the keyword extraction accuracy based on the background knowledge.

ACKNOWLEDGMENT

This paper is supported by the National Ntural Science Foundation of China (Grant No.61502437) and the National Science and Technology Support Plan 2015BAK22B01.

REFERENCES

[1] Paik J H. A novel TF-IDF weighting scheme for effective ranking[C] Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013:343-352.

[2] YueShan Chang , MinHuang Ho , ShyanMing Yuan . A unified interface for integrating information retrieval. Computer Standards & Interfaces. 201 1(21) : 325-340.

[3] Bohne T, R6nnau S, Borghoff U M. Efficient keyword extraction for meaningfuldocument perception[C] . Proceedings of the l lth ACM symposium on Document engineering. ACM, 2011 : 185-194.

[4] Zhao L, Yang L, Ma X. Using tag to help keyword extraction[C]. Computer and Information Application(ICCIA) , 2010 International Conference on. IEEE, 2010 : 95—98.

[5] Bao Hong, Deng Zhen. An extended keyword extraction method[C]. Proceedings of the 2012 International Conference on Applied Physics and Industrial Engineering. USA : Elsevier,201 2 : 11 20-11 27.

[6] Songhua Xu, Shaohui Yang, Francis C.M. Lau. Keyword Extraction and Headline Generation Using NovelWord Features. Association for the Advancement of Artificial Intelligence.(AAAI). 2010.

[7] M1adenic D.Maehine Learning for Better web Browsing[J]. AAAI Spring Symposium Teehnical RePorts on Adaptive User Interfaces Menlo Park. CA : AAAI Press. 2013 : 82—84.

[8] Bao Hong, Deng Zhen. An extended keyword extraction method[C]. Proceedings of the 2012 International Conference on Applied Physics and Industrial Engineering. USA : Elsevier, 2012 : 1120-1127.

[9] Bekavac, M; Šnajder, J. GPKEX: Genetically Programmed Keyphrase Extraction from Croatian Texts. In Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, ACL, pages 43-47, Sofia, 2013.

[10] Beliga, S; Meštrović, A; Martinčić-Ipšić, S. Toward Selectivity-Based Keyword Extraction for Croatian News. In CEUR Proceedings (SDSW 2014), Vol. 1301, pages 1-14, Riva del Garda, Trentino, Italy, 2014.