Algorithm of Keywords Extraction about Power Documents Based on Hadoop

Wang Tong[1, a], Wang Yongzhi[1, b], Jin Liang[1, c], Li Yongheng[1, d]

College of Instrumentation and Electrical Engineering, Jilin University, Changchun, 130061, China

[a]email: mynameWangTong@163.com, [b]email:iamwangyongzhi@126.com(corresponding author), [c]email:15843008173@163.com, [d]email:1844330774@qq.com

Keywords: Big data, Hadoop, HBase, Keywords, Power documents

Abstract: As the power big data is developing with a wide range of varieties, it is used to manage and analyze data, especially the unstructured data such as power documents. In this paper, HBase database and HDFS are applied to make the document searching easier and faster. Keywords extraction based on Hadoop is use to read the document in an easy way. Above all, it deserves management and design algorithm on keywords extraction based on Hadoop**.**

1. Introduction

In recent years, the power big data increases dramatically. Compared with how it functioned many years ago, the current data include more figures, tables and documents, which become larger and full of varieties. How to manage and analyze those data is a big challenge right now, especially in managing and analyzing the documents. However, using Hadoop can shoot those troubles. In this paper, we deal with power documents through documents management and keywords extraction based on Hadoop to make readers find document easier and read document faster.

2. The model of keywords extraction description

Usually, there are many methods on keywords extraction, such as Semanteme, Machine Learning, Complex Network and Statistical method[1]. Big data can be applied in all these methods above to improve the accuracy and efficiency of keywords extraction. In this paper, we mainly talk about using statistical method to extract keywords.

The model of keywords extraction description is shown in Fig.1. We can see that when we get a power document, first of all we preprocess this document, including removing stop words and stemming. Next we will count the word, and obtain the term frequency data. And then it comes the term weighting, which means differentiating the indicators according to their specific weights. Finally, we sort the result and obtain the corresponding results as keywords[2].

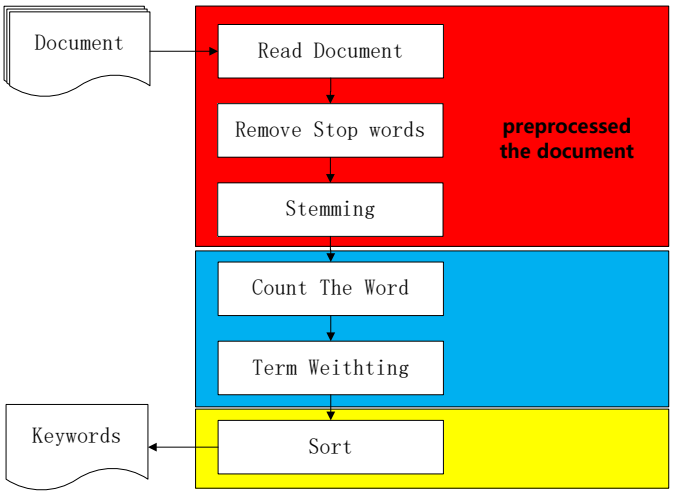

Fig.1. The model of keywords extraction description

3. The Power Document's Management and Keywords Extraction

3.1. Management based on Hadoop about power documents

By using HBase 1.2.2 to manage data, the system requirements can be met. HBase possesses a real-time query ability by integrating the mode of key/value store. Moreover, HBase is capable in offline processing and batch processing by MapReduce. Above all, HBase is able to find results in big data and analyze big data[3,4]. As for using HDFS to manage the document, we can avoid losing documents if the HBase fails to work[5].

In this paper, we try to design and implement two databases, by which users can go to the system immediately. Moreover, users can choose the way of uploading a document, and they can download the source file to the local server. Data can also be managed by users through databases, which enables data searching to be easier and faster.

3.2. Algorithm on keywords extraction based on Hadoop

3.2.1. Word Count

Based on Hadoop, we count the words and obtain terms frequency data. Compare with the pervious principle, the idea to be based on Hadoop can improve the efficiency in counting words because Hadoop will spilt data set into many smaller datasets. Those datasets can process data at the same time while the data will be grouped when data processing is finished. In other words, words will be divided into different parts, which will be counted at the same time. Finally, the result will be grouped and exported. Obviously, counting words based on Hadoop is more efficient and time-saving and it is a better way to get term frequency data as well.

3.2.2. Term Weighting

As we know, we can obtain terms frequency data, which is an important index to get keywords but it’s not the keywords. We need to make the term frequency data different according to specific weights of indicators. We always call this process term weighting. By the research, we usually think about word frequency, location, professional words, inverse document frequency and distance. Above all, we can get an equation:

$$Weight = \alpha \times tf + \beta \times loc + \gamma \times pro + tf \times idf + dis \quad (1)$$

(1)Word Frequency

By counting words, we can get word frequency. If the word possesses high frequency, it means that the word has been mentioned for many times, and it also means that the word is important for

the document. Therefore, this word will probably become keywords.

(2)Location

The location where the word appears is one of the most important indexes to judge the keywords. Usually, within a document there are some locations including keywords like titles, the first sentence in paragraph, the last sentence in paragraph and the summary sentence. In fact, title is the most important thing among all. On the one hand, through the title we can know what the document is mainly about. If the title includes words of technology, then the keywords will usually include them too. If the title illustrates a place, the keywords usually include some positions, and so on. On the other hand, we can judge the type of document through the title, and then further judge the type of keywords. The first sentence and the last sentence in paragraph always summarize the whole paragraph for avoid losing any keywords. For the same reason, we check the summary sentence because it always summarizes or emphasizes the whole sentence or paragraph.

(3)Professional words

Many professional words always come up, such as words of science and technology, words of a place, words of special criterion and so on. In this paper, it is important for keywords extraction to judge words whether they belong to professional words or not. Actually, the professional words appeared on the document are always the keywords.

(4)Inverse Document Frequency

In information retrieval, inverse document frequency is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The equation is:

$$IDF=\log(N/(n+1)) \quad (2)$$

The equation means the logarithm of the number of the document in a collection or corpus is divided by the number of documents containing the words and plus one. It can be inferred that the less appear, the more important. In fact, inverse document frequency is used to distinguish between the common words and the keywords.

In this paper, for power documents, the document possesses different paragraphs, which are independent between each other, like the way a collection or corpus has many documents. In this way, we get the inverse document frequency of the words in the whole document.

(5)Distance

The word span is the distance between the same word. The equation is:

$$dis=(last-first+1)/sum \quad (3)$$

The equation means the first appear place is compared with the last appear place, and then calculate the proportion of documents relative to the document. The word span usually shows range about the word control. The larger the word span, the more important the word to the whole document.

4. Test results

According to the test, we can draw two conclusions as follows:

On the one hand, the compare is on the used time of keywords extraction between the way of hash and the way of Hadoop is shown in Fig.2. It can be seen that the time of Hadoop is stable and consistent while the time of hash tends to be volatile. Therefore, we can infer that if the document is in small size, we should choose the way of hash, conversely we choose the way of Hadoop.

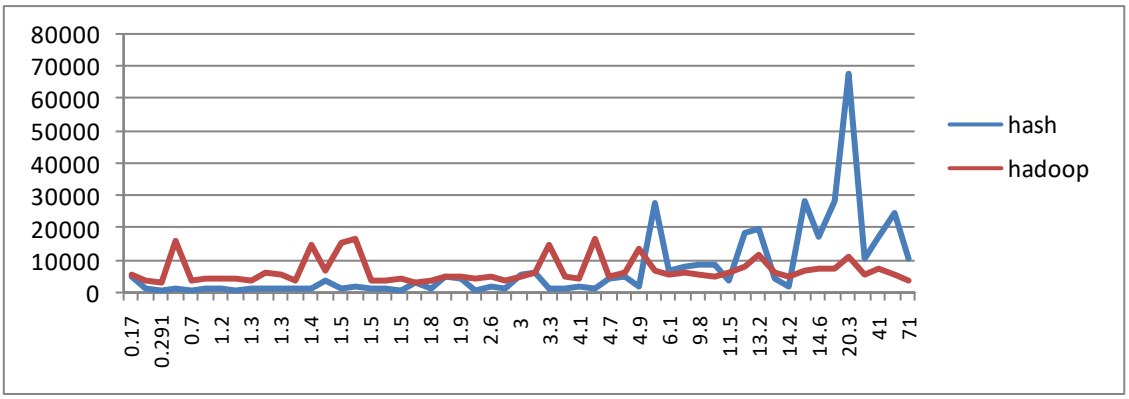

Fig.2. Comparison on the used time between the way of hash and Hadoop

On the other hand, as using the way of Hadoop can introduce more complex and improved algorithm, the keywords extraction based on Hadoop from the document can be achieved and the accuracy of keywords extraction is improved as well. The extraction rate of keywords is shown in Fig.3.

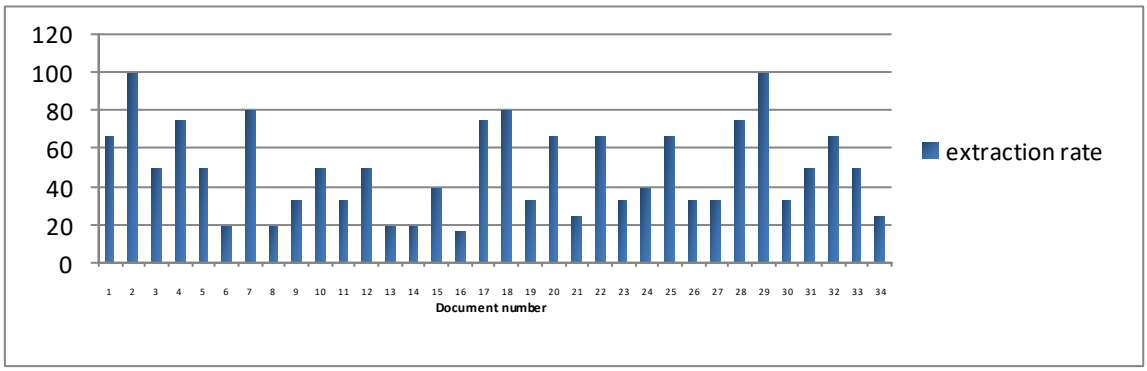

Fig.3. The extraction rate of keywords extraction

5. Conclusions

As we know, the Hadoop is currently widely used and in this paper we manage the power document and extract the keywords about the document based on Hadoop. In fact, by using Hadoop, we achieve management of power documents, and improve the efficiency and accuracy of keywords extraction.

References

[1]LUO Fanming, YANG Haishen. On the Statistical Features-based Information Keyword Extraction Method in Era of Big Data[J]. Information and Documentation Services, 2013,03:64-68.

[2]LUO Yan, ZHAO Shuliang, LI Xiaochao, HAN Yuhui, DING Yafei. Text keyword extraction method based on word frequency statistics[J]. Journal of Computer Applications, 2016,03:718-725.

[3]Rajaraman A, Ullman J. D. Mining of massive datasets[M]. Cambridge University Press, 2012.

[4]YAO Weiguo, ZHAGN Dongbo. Web Text Keyword Extraction Scheme Based on the Hadoop Distributed Platform[J]. Natural Science Journal of Xiangtan University, 2016,02:79-83.

[5]HAO Shukui. Brief Analysis of the Architecture of Hadoop HDFS and MapReduce[J]. Designing Techniques of Posts and Telecommunications, 2012,07:37-42.