Chinese Words Segmentation Based on Double Hash Dictionary Running on Hadoop

Chao Feng[1, a], Baoan Li[1, b*]

[1]Computer School, Beijing Information Science and Technology University, Beijing, China

[a]fengchaohao1@163.com, [b]liba2010@139.com

*Corresponding author

Keywords: Search Engine, Words Segmentation, Hadoop, Double Hash

Abstract. Words Segmentation is an essential stage to establish a search engine, and the quality of words segmentation directly affects the search speed and precision. We have to adopt a word segmentation tool which can deal with a big data when large amounts of data is being segmented, because the traditional single PC segmentation has not been able to meet our needs. This study presents a Chinese words segmentation technology based on Hadoop. In this paper, the adoption of dictionary created by the double hash function, the adoption of the maximum forward successive matching method, and the using of the MR programming realize the parallel words segmentation in the distributed cluster, and it greatly shortens the time and increases efficiency. It provides a convenient and quick method for the words segmentation of a large quantity of text.

1. Introduction

With the high development of the search engine, people rely on search engine more and more, a direct impact on the search engine quality is to establish the index, and the establishment of index of speed and precision directly by the segmentation precision and speed of decision. Up to now, the author has done lots of research in the system of the single computer words segmentation, There have been many kinds of segmentation software that have great precision of Chinese word segmentation, such as Paodingjieniu Chinese word segmentation software [1], Ikanalyzer Chinese word segmentation software [2], Jeasy component software, all of them divided words are based on the dictionary word segmentation model. There are also some other words segmentation models, such as some words segmentation methods that are based on statistics and understanding. Because the limitation of current technology, these methods' theory are still being researched and don't produce any practical value. Although the Chinese words segmentation method that is based on dictionary and very mature, with the Internet entered into the era of big data, the ordinary segmentation software already falls down when we deal with large amounts of data. Therefore, the traditional segmentation tools have been unable to meet a large number of segmentation needs.

Especially for some large search engines need to build the index frequently. In this case, it must be carried out by means of distributed computation [3]. A lot of work must be divided into a number of works to make them work simultaneously on different computers, so we must transplant the dictionary segmentation mechanism to Hadoop mechanism, and handle it with the MR (Map and Reduce) mechanism of Hadoop [4]. Through the Hadoop mechanism, the program can perform in parallel in a hundred computers cluster or even more computers cluster, and the working time can be shortened to a few minutes. By this way, the time problem of segmenting words in large amounts of data is solved, and the mechanism of Hadoop is more suitable for word frequency statistics. The statistics of word frequency can be completed easily by using the parallel principle. While it is almost impossible to handle the word frequency statistics of large amounts of data with the traditional PC, even if you have enough time, space is also a big restricted factor [5].

According to the dictionary word segmentation method, this paper proposes a word segmentation method that segmenting words in the Hadoop cluster. It is proved that the method can automatically segment words and statistic words for a large number of texts.

2. Double Hash Chinese Word Segmentation Based on Hadoop Platform

2.1 The realization of the Map function

The reading and processing of data and the production of intermediate result sets are realized in the Map function. When processing a paragraph of text, the default of Map function is processing by line at a time, which obviously can't meet the requirements for words segmentation, because it only segments the words on the current line if reading by line, and there is no guarantee that whether the word at the end of the current line and the word at the beginning of the next line can compose words or not. Therefore, The Map function reads by paragraph one time.

First read a paragraph, and then splitthe wordsin the map function, word segmentation is completed to produce an intermediate result sets <key, Value>, the current key mode is designed for key= (the article number: word segmentation phrase), Value= 1. The design of Key value needs to account for the differentiation of the number of the same words in different articles. Colon is only used to distinguish the article title and segmentation of words [6], thus, the key can uniquely identify the place of words and can help to extract the segmented words in the later stage, and it will also make the following polymerization method more convenient. The value of the value of 1 is used for word frequency statistics, which only needs simple evaluation and operation for Value in the Reduce function.

While segmenting words for large number of articles, a lot of key-value pairs will be generated in the same Map function, and it may also produce many same key-value pairs, that is to say both the key value and the value's value are the same. All of the generated key-value pairs in the Map function need to be transmitted to the Reduce function. The date in the Hadoop cluster share communication channels, and Map and Reduce are often not running simultaneously in the same host, so the key-value pairs generated in the Map function need to transmit to other machines through the network and the same key-value pairs will be transmitted to one Reduce. If transmitting of large amounts of data at the same time, it will lead to the network congestion, which would greatly reduce the Hadoop data processing capabilities, so this problem must be solved. The proposed solution is to merge in the late stage of Map, that is, merge the same key-value pairs to one in the Map. By this way, the merger of the key-value pairs will reduce the network transmission data volume before transmitted to Reduce. This is just the merging of the same key-value pairs in the same Map function, and the different Map key-value pairs need to be merged in the Reduce function. The storage of HDFS is in order, when reading a paragraph of content, the same content of the article has a great probability of storage in order, and in the same Map function, it has a large probability of reading the same article, so it will be easier to merge the whole article.

The words in the famous novel which is named *how the steel was tempered* turn to corresponding key-value pairs through Map segmentation. Assuming that the number of "*The Making of Steel*" is 1; only parts of the segmented words are listed in the Table1.

Table 1. Key-value pairs

No.	Key	Value
1	01:wounded	1
2	01:hence	1
3	01:open	1
4	01:third	1
5	…	1
6	…	1
7	01: darkness	1
8	01: darkness	1

Table 1 shows the structure of the key-value pairs, at the same time, due to the word *darkness* has reputation, so the key-value needs to be merged in the later stage of the Map function, and merged results is listed in Table 2. By this way, it can reduce the transferring amount of data; of course, there

is a better way to represent, the primary key can be represented by the number of articles; generally, the number is in binary notation representation, which will save more space.

Table 2. The combined value pairs

No.	Key	Value
1	01:wounded	1
2	01:hence	1
3	01:open	1
4	01:third	1
5	…	1
6	…	1
7	01: darkness	(1,1)

2.3 Reduce function processing data

The Reduce principle is to send the Key-value pairs of the same Key value's to one Reduce function, and process the transmitted Key-value pairs according to their own methods. The task of Reduce is to complete the word frequency statistics and according to need to collect the key-value that generated from the Map function. First of all, it is necessary to collect the same key-value pairs, namely, translate all the same key‘s value of value into int, then carry out summation operation. By this way, the table of times of each word appearing in different articles can be made a statistics. Fig.2 shows the Program flow chart.

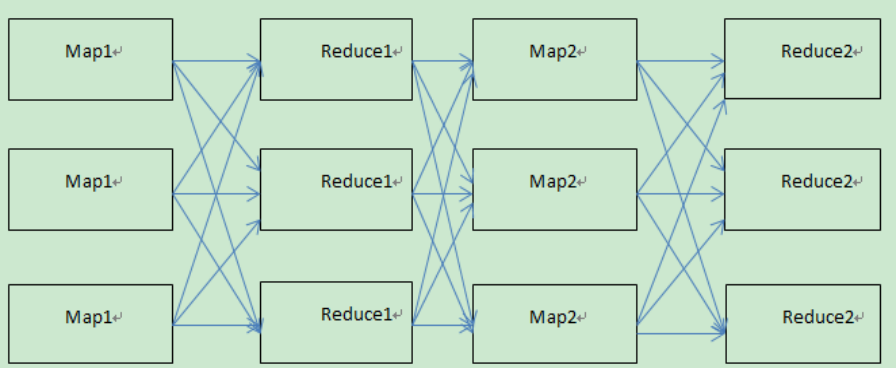

Fig.1. Program flow chart

Key values can also be divided, and establish it’s inverted index, at the same time, provide the appearing frequency of each document, the results shows as in the Table 3, 01 represents the article number, right of the colon represents word, the Value represents the number of the current word appearing in the article. It needs to make two Map-Reduce computations, as shown in Fig. 1, the first Map-Reduce calculation output results as the input of second Map-Reduce computation, at the same time, in the second Map, reading data line by line, separate the read key values as the words needed. The Value’s value adopts the presentation of article number and frequency. Table 4 indicates the intermediate results of the key-value pairs produced by the second Map function. Table 5 shows the final results produced by the second Reduce.

Table 3. Word frequency statistics

NO.	Key	Value
1	01: wounded	123
2	01:open	30
3	01: darkness	50

Table 4. Intermediate results

NO.	Key	Value
1	wounded	123 01
2	open	30 01
3	darkness	50 01

Table 5. Inverted index and frequency

NO.	Key	Value
1	wounded	123 01, 50 03, 30 432…
2	open	30 01, 123 040, ……
3	darkness	5001, 100 987, ……

3. Conclusion

Based on Hadoop, the research of Chinese words segmentation for large data was realized. It combined the double hash dictionary words segmentation and MR programming idea. Build a segmentation model for a large number of texts, under this mode, the word segmentation can be rapidly processed in a large-scale. Distributed processing is the current hot topic [7], and applying the computation into practice is the core of this work [8]. This study stresses the core idea and the content of research steps to achieve a more comprehensive and optimized Chinese word software that is based on Hadoop platform.

Acknowledgements

The work was supported by Project of Construction of Innovative Teams and Teacher Career Development for Universities and Colleges under Beijing Municipality (Grant No. IDHT20130519), and Project of the Specialty Construction and Comprehensive Reform under Beijing Municipality (Grant No. 71M1510818).

References

[1] Dianzhe Sun, Haiping Wei,Yan Chen, Realization and Evaluation of Paodingjieniu Chinese Segmentation in Nuch. Computer and Modernization, in Chinese, 6(53), pp. pp.188-190 (2010)

[2] Jie Chai, Based on IkAnayzer and lucene geocoding research and implementation of Chinese search engine. City Survey, in Chinese, 6(06), pp.46-50 (2014)

[3] Yu Wu, The present situation and the development of Chinese search engine. Modern information, in Chinese, 3(03), pp.40-44 (2003)

[4] Jianjiang Li, Jian Cui, LinYan, Survey of MapReduce Parallel Programming Model, Acta Electronica Sinica, no.11, pp.2636-2642 (2011)

[5] Zhong Zong, Study of Segmentation Algorithm of Dictionary Mechanism Orienting Chinese Information Retrieval. Computer Technology and Development, 24(4), pp. pp.119-121 (2014)

[6] Method of Chinese words rough segmentation based on improving maximum match algorithm. Computer Engineering and Applications, 50(2), pp.124-128 (2014)

[7] Guijing Luo,Hongxiao Fei,Dai Ge, Research of Chinese Segmentation Based on Converse Segmentation Dictionary, Computer Technology and Development, 18(1), pp. pp.81-83 (2008)

[8] Bingyi Zhang, Bo Wei, Cheng Chen, et al, Based on dual coding of the Chinese word segmentation algorithm (J). Journal of Nanjing University of Science and Technology, 38(4), pp.530–526 (2014)