Research on Chinese segmentation algorithm based on Hadoop cloud platform

Chen Hong[1,a]

[1]Computer ScienceSchool of Wuhan DonghuUniversity, 430212, China

[a]chenhong_dh@163.com

Keywords: Chinese word segmentation; ICTCLAS; IKAnalyzer; Inverted descending order; HDFS; MapReduce; Hadoop

Abstract. IKAnalyzer (IK) and ICTCLAS (IC) are very popular Chinese word segmentation algorithms and play an important role in solving text data in a stand-alone environment. In this paper, we compare IK and IC algorithm performance through theory and experiments that reported on experimental work on the mass Chinese text segmentation problem and its optimal solution using the Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and by using parallel processing to process large data sets by using the MapReduce programming framework. The results obtained from various experiments indicate favorable results of above optimized IC and IK algorithms to address mass Chinese text segmentation problems. At the same time, in order to make the large data set after processing is more easily and directly showed, we introduced the Inverted descending order on the segmentation of word frequency in this paper. Through a comparative study of the two kinds of Chinese segmentation algorithm based on Hadoop platform, provides the powerful support for the efficient processing of Chinese mass information.

Introduction

Facing with these massive Chinese information, we hope the result of thetext segmentation first shall be in line with the meaning of the word itself, secondly the length of the word group shall be as long as possible (i.e. the amount of the word segmentation shall be as few as possible). However, Chineseword segmentation merely meeting the two above mentioned conditions is not always what we want, and we often need to consider the accuracy of word segmentation, word segmentation speed and F value and other indicators. In short, there are advantages and disadvantages in various Chinese word segmentation algorithms, and it is difficult to absolutely make the distinction. Thus, most of the time the Chinese word segmentation algorithm needs to be combined with the practical application.

The purpose of this article is to consider the advantages and disadvantages of these word segmentations based on cloud computing platform and then make weighs and selections. Firstly, in section 2, we make a brief review on these two algorithms, and in section 4, the comparison on these two types of word segmentation algorithms is made under four different environments and the corresponding analysis results are given as per the experimental results.

Two Types of Word Segmentation Algorithms

iKAnalyzer

IKAnalyzeris a light weight third party kit of the Chinese word segmentation based on JAVA language. It has experienced three comparatively complete versions since released in 2006, and the current version under application is IKAnalyzer 2012. According to the official statement of the author Lin Liangyi, the IK tokenizer applies the “positive iteration most fine-grained segmentation algorithm”. If we analyze its source code, we can see that the word segmentation utility classIKQueryParser plays an important role and and it adopts layers of iterative search for segmentation on the search keywordfrom the max. word to the min. word with the max. speed of 80W /S (1600KB/S). Thedevelopment language is java thus it is endowed with cross-platform characteristics. The IK tokenizerapplies the multiprocessing model, and the model supports English, figures and Chinese characters, and meanwhile it also is suitable for treating Korean and Japanese. It

optimizes the dictionary storage type and occupies less storage space, and it also supports custom dictionary.

Institute of Computing Technology, Chinese Academy of Science proposes the Chinese Lexical Analysis System (ICTCLAS), mainly including Chinese word segmentation, part-of-speech tagging, identity of named entity and new word recognition, and meanwhile supports user dictionaryand other functions.

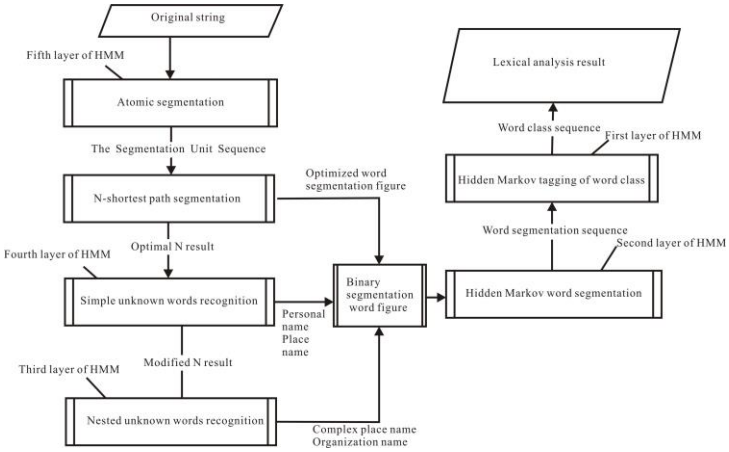

Fig.1 Cascade Hidden Markov Model

ICTCLASproposes the Cascade Hidden Markov Model（CHMM) on the basis of the Hidden Markov Model. The CHMM is a multi-hierarchy simple HMM combination and each layer of Hidden Markov Model is interlinked by the following methods: each layer of HMM enjoys a word segmentation figure (see Fig.1), each layer of Cascade Hidden Markov Model applies the N-Best strategy and sends the several best results generated to the next level for the use of the model with the higher level.

$$W^{\#} = \arg\max_{W} P(W) \tag{1}$$

Given a word segmention atom sequence S, one of the word segmentation result of S is recorded as W=(w1w2...wn), the corresponding category type is recorded as C=(c1c2...cn), and meanwhile we take the word segmentation result W#with the max. probability as the final word segmentation result. Then

$$W^{\#} = \arg\max_{W} P(W) \tag{2}$$

$$W^{\#} = \arg\max_{W} \prod_{i=1}^{N} p(w_i|c_i)\, p(c_i|c_{i-1}) \tag{3}$$

Take the word category as the status, the word as the observed value, and unfold it through first order HMM

$$W^{\#} = \arg\max_{W} \prod_{i=1}^{N} p(w_i|c_i)\, p(c_i|c_{i-1}) \tag{4}$$

(Thereinto c0is the openning tag BEG).

$$W^{\#} = \arg\min_{W} \prod_{i=1}^{N} \left[-\ln p(w_i|c_i) - \ln p(c_i|c_{i-1})\right] \tag{5}$$

For the convenience of calculation, the negative logarithms are usually used for the calculation, namely

$$W^{\#} = \arg\min_{W} \prod_{i=1}^{N} \left[-\ln p(w_i|c_i) - \ln p(c_i|c_{i-1})\right] \tag{6}$$

The word segmentation result searched is finally transferred as the shortest pathfor the word sequence from the initial node to the final node.

System Framework

MapReduce is a programmable model and it is also a subitem of Hadoopwhich can manage and generate super data sets. The user sends a series of key value pairs into the Mapfunction, key value pairs with appointed format will be generated after being processed by the function, then they are sent to the function of Reduce, and such function will put the key value pair with the same key together. The task processing procedure of Chinese wordsegmentation based on the Hadoop platform adopts the above mentioned method in this article. During the process of the system operation, there will be many important operations: data partition, executive scheduling of programs, processingfault machine and management of the communication among the nodes, etc.

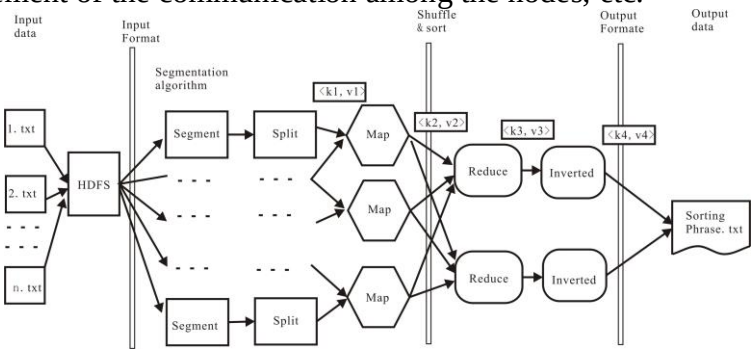

Fig. 2 Chinese Word Segmentation Algorithm Frame Based On HadoopPlatform

See the treatment scheme of the Chinese word segmentation algorithm based on big data in Fig. 2. The treatment of the data set in Chinese text in the figure is divided into two parts: “word segmentation” and “data partitioning”. The treatment of text data set for ICTCLAS and IKAnalyzerword segmentation algorithms is made on the HDFS(Hadoopdistributed file system) and the word segmentation result is given. Hadoop creates a task for every split input (data fragment, generally with the size of HDFSdata block) and treat the data item record of the data fragment in the task in sequence.

Experiment

In order to realize the comparison betweenthe word segmentation algorithms of ICTCLASand IKAnalyzer in the Hadoopdistributed platform, the configuration of Hadoop includes four physical nodes and HDFSdistributed storage. We make the word segmentation algorithms of ICTCLASand IKAnalyzerafter optimization run in the aggregate. First of all, we upload the text data set to be processed to HDFS. Then we make use of the core functionMap and Reduce of Hadoop to transfer the <key,value> pair input into <key,value> pair meeting the requirements as per the rules of the Chinese word segmentation and inverted word frequency. Put the computational node and the data node in the same computer, and the MapReduce frame guarantees to implement the calculation task on the nodes where data are stored.

Nodes with Different Quantities

Record the running time of the text data with the capacity of 1G under the 1，2，3 and 4 nodes separately to compare with the running time of the two algorithms under the distributed Hadoop platform and concentrative single mode environment. On the one hand, it is to test the performance of the Hadoop platform, on the other hand, it is to compare the word segmentation speeds of two word segmentation algorithms under different modes. In order to better imitate the effect of word segmentation of the oversized text set under Hadoop platform,though the text itself takes up little space, we hereprogressive increase twice every time, then the text with the capacity of 1Mis 1G after increasing 10 times.In the experiment, we use financial No. 4382 document (326 characters)in the Sogou corpus, and it will reach 100M after 18 times, 15times, 14times and 11 timesof increasing.We copy the same text 10 times and then obtain a text data of 1G (with 378773957 Chinese characters). We upload the ten text data obtained to the HDFS, and after the analysis and treatment of the ICTCLAS and IKAnalyzer algorithms, we get the following experimental results:

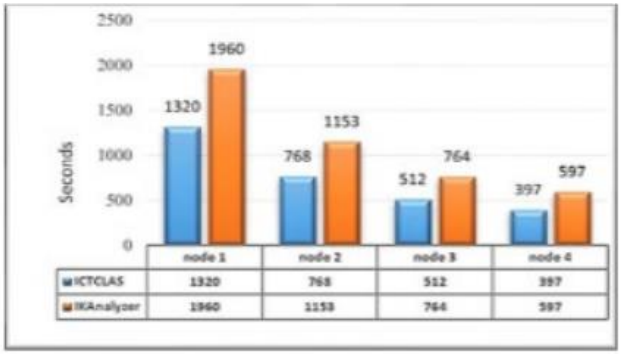

Fig. 3 Comparison of Nodes with Different Quantities

The data in the figure shows that when processing the text withultra-large capacity, the overall time for the word segmentation and inverted orderof IC word segmentation algorithm is obviously shorter than that of the IK word segmentation algorithm. With the increase of the quantity of the nodes, the advantage of the comprehensive processing speed for IC word segmentation algorithmis still quite apparent.

Texts with Different Sizes

For texts with different capacities of 1K，100K，1M，10M and 20M, compare the run time under four nodes with texts with capacities of 10K，1M，10M，100M and 200M. Will the size of the text affect the running speed of ICTCLASand IKAnalyzer on Hadoop platform? The above mentioned texts with different capacities are obtained by the same progressively increase method. The texts are still financial No. 4382 documentin the Sogou corpus. The experimental result is as below:

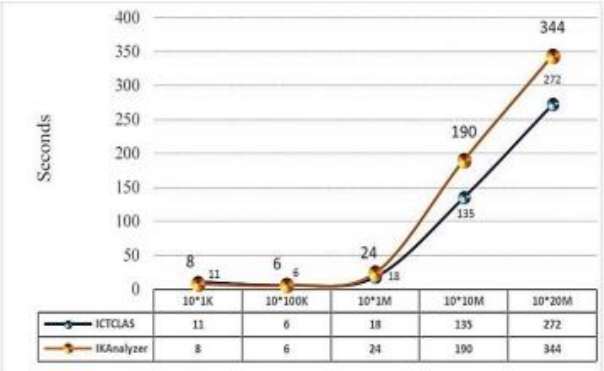

Fig. 4 Comparison of Texts with Different Sizes

The running times of the two algorithms in the above figure in 10*1K are both longer than that of the text with the capacity of 10*100K, and it means undersize text will reduce the processing speed of these two algorithms on Hadoop. The processing speed of the text with the capacity of 10*20M is 800 times of that of the text with the capacity of 10*1K. In addition, the bigger the text is, the more obvious the processing speed will increase. When the capacity of the text is about 10K, IK is superiorto IC. When the capacity of the text reaches 1M, the running speed of these two algorithms is almost equal; when it exceeds 10M, the processing speed of IC is obviously superior to that of IK.

In brief, when these two word segmentation algorithms are processing smalltext data, they both will occupy a large amount of time due to the start of Map task and make the word segmentation algorithms in a low efficiency within unit time. thus, word segmentation algorithm based on Hadoop is not suitable for processing smalltext data set.

Texts with Different Quantities

Under the condition that the overall text capacity is fixed as 100M, The comparison of the running time in single node amongtexts with different quantities of 1， 10，100 and 1000 is used to compare whether the quantity of the text will affect the running speed of ICTCLASand IKAnalyzer on Hadoop platform?The texts with the capacities of 100M，10M，1M and 100K are obtained by the same method. The experimental result is as below:

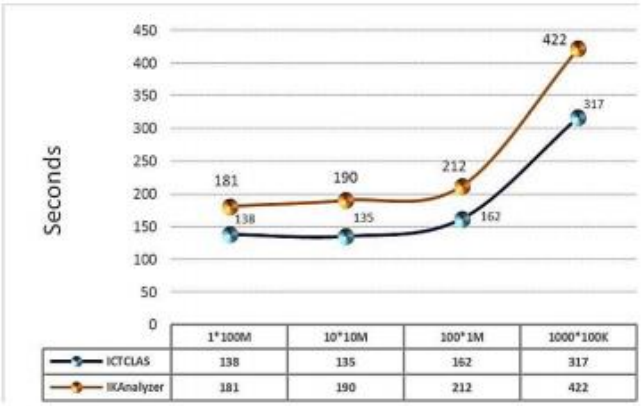

Fig. 5 Comparison of Texts with Different Quantities

HDFSis the distributed file system providing high throughput and it is designed to visit large files in the beginning. However, most of the news data are small files with the capacity of severalKB. Themetadatagenerated from the amounts of small files will send excessive read-write requests to the host node to increase the time of data transmission and request response in the network and finally decrease the performance of the whole platform.

Therunning times of the two algorithmsin the above figure in 1000*100K are both longer than that of the text with the capacity of 1*100M, and it means overmuch undersize text will reduce the processing speed of these two algorithms on Hadoop.In addition, the more the quantity of the text is, the more obvious the processing speed will decrease. When the quantity of the text exceeds 1000, the processing speed of these two algorithms will decrease obviously and IK will decrease more seriously. When the quantity of the text is 10, the processing speeds of these two algorithms are comparatively ideal.

Conclusion

In this article, we find that when the word segmentation algorithms of IK and IC deals with massive text data, they have very good performance, and it will make optimization in the aspect of undersized word segmentation when IC processes the massive texts in the later period. In addition,as the word segmentation method based on understanding is emerging in the aspect of Chinese word segmentation. This kind of word segmentation method isto achieve the effect of word recognitionby making the computer imitate people’s understanding of sentences. The basic idea is to make syntactic and semantic analysis while make the word segmentation, and make use of syntactic information and language information to deal with ambiguity.If it can be successfully introduced to the Hadoop platform, it should have good performance.

References

[1] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters [J]. Commun ACM 2008, 51:107-113.

[2] Chuck Lam. In Book Title: Hadoop in Action [M].

[3] Y. Geng, J. Chen, K. Pahlavan, Motion detection using RF signals for the first responder in emergency operations: A PHASER project[C], 2013 IEEE 24nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), London,Britain Sep. 2013.

[4] S. Li, Y. Geng, J. He, K. Pahlavan,Analysis of Three-dimensional Maximum Likelihood Algorithm for Capsule Endoscopy Localization, 2012 5th International Conference on Biomedical Engineering and Informatics (BMEI), Chongqing, China Oct. 2012 (page 721-725).