Research on Campus Network Hot Topics

Li Qiong
School of Computer Science and Technology
Hankou University
Wuhan, Hubei, China, 430212

Yan Lijuan*
School of Computer Science and Technology
Hankou University
Wuhan, Hubei, China, 430212

Chen Li
School of Computer
Central China Normal University
Wuhan, Hubei, China, 430079

***Abstract*—With the advancement of education informatization in China, using IT technology to pay attention to campus network hot topics, and timely grasping students' dynamic information has gradually become the normalcy of campus management. In order to solve the problem of lower computing power and poor storage capacity in network hot topic mining system, this paper proposes a new method for applying cloud computing technology to the network hot spot mining system. By means of paralleling mining process, the proposed method can improve the mining speed, reduce the mining time consumption effectively, so as to help university administrators to timely grasp the hot topics that students care about and take precautionary measures as soon as possible.**

Keywords—Education informatization; Campus network; Hot topic; Campus management

I. INTRODUCTION

As the rapid development of Internet and communication technology, it produces massive amounts of data in the world every day. According to Forrest Research, more than 80% of these massive data are in the form of text. Subject extraction of network text corpus refers to the process of extracting implicit and potentially useful subject knowledge from these massive text data, it belongs to the category of text mining.

At present, text mining is widely used in commercial, military and education fields. For many enterprises, public institutions and government agencies, document management is a tedious but extremely important work. Text mining can help managers to manage documents effectively. For example, automatic document clustering division, automatic document summary for easy management and use, etc.

In addition, there is also a lot of potentially useful information hidden in a large number of patent literatures. By means of text mining technology, it can mine relevant competitive intelligence to assist enterprises in strategic decision-making. Similarly, in the military field, text mining can help people quickly discover potential knowledge patterns hidden in massive text information and provide intelligence, so as to assist the military decision-making [1]. For colleges and universities, they can timely grasp the hot topics that students care about by the method of extracting subject of network text corpus, and do a good job of active guidance and preventive measures.

In the process of analyzing massive text data, the text mining system generally has the disadvantages of low computing capacity and poor storage capacity. It cannot meet the requirements of users on timeliness and massive storage. Cloud computing is developed on the basis of distributed processing, parallel processing and grid computing. It is an internet-based supercomputing mode. It doesn’t need high-performance server, but it can provide massive storage capacity and supercomputing capacity through running on a large number of cheap PC clusters [2]. By combining cloud computing technology and text data mining technology, it can obtain higher data mining speed by using lower data processing cost, and it can also reduce the time consumption from massive text data to auxiliary decision making.

This paper mainly analyzes and discusses how to apply cloud computing technology to network hot topic mining extraction and analysis： Through paralleling text subject extraction process, it can improve the text subject extraction speed, so as to speed up the text subject extraction process, reduce the time consumption of hot topic mining, and then quickly help teachers and administrators in colleges and universities to make timely decisions and measures.

II. RESEARCH STATUS AT HOME AND ABROAD

The subject extraction of network text corpus refers to the process of extracting and mining various subjects’ information from all kinds of Chinese text information existing on the internet. It belongs to the fields of Chinese text information process.

In 1995, Feldman formally proposed the concept of text mining, and since then, KDD international academic conference and other international conferences have discussed a large number of articles and reports on text mining every year in just a few decades. In the study of text mining, the representation model of text experienced has developed from the initial word-frequency vector notation to deep semantic notation, such as ontology and concept. Also, the type of text mining has expanded from the initial text structure analysis, text clustering/classification, to trend prediction, pattern discovery and other aspects, etc.

At present, text mining has been applied in many fields, such as digital information library, automatic analysis and filtering of emails, automatic analysis of salesmen's daily reports, deeply mining and discovery sales trends, and knowledge discovery of biomedical reports.

Abroad, many government agencies, such as the FBI, already use text-mining software to evaluate and analyze intelligence data. Many famous companies in the world have developed many practical text mining tools successively, such as "Intelligent Miner for Text" developed by IBM and "Text Miner" developed by SAS [3].

The research on text mining started late in China, but the government, enterprises and institutions and academic circles all pay great intention to the study of text mining theory and technology. In 1998, the project "image, voice, natural language understanding and knowledge mining" which has been implemented in the first batch of key basic research development planning in China includes text mining. Since then, a large number of domestic researchers began to carry out knowledge mining analysis and research oriented to Chinese text. Because Chinese is different from English, Chinese scientists and technicians have done groundbreaking work on the word segmentation algorithm, and the designed algorithm has high word segmentation speed and precision. For example, The ICTCLAS developed by the Chinese academy of sciences is one of the most widely used Chinese word segmentation methods. In terms of text representation model, it develops from the early "word bag method" to the semantic representation of text, such as CNKI knowledge network, etc. With the development of the Internet, the analysis and mining of the web network has gradually become the current research hotspot, such as the log mining of the web network and the discovery of user interest mode. In terms of the depth of text mining, there are still many text-based retrieval, clustering and classification methods.

However, association analysis, distribution analysis, trend prediction, subject tracking, and network association discovery, have gradually developed and become the research focus of text mining analysis in recent years[4-5]. In the aspect of processing large-scale and massive network text data mining analysis, the current text mining system is still generally solved by optimizing serial algorithm and high-performance computer, and the research on parallel mining through distributed technology is still relatively few.

Parallel mining is a good method to solve massive text data mining analysis at present. In 2007, Google pioneered the cloud computing technology, because its powerful parallel processing capability and massive storage function, it can meet the mining and analysis demand of massive text data from both computing and storage. At the same time, applying cloud computing to massive network text data for analysis and mining has become a research hotspot. For example, Yang Liu et al. studied distributed latent semantic indexing technology based on MapReduce; it can cluster the input network text matrix into K modules through the parallel k-means algorithm, and then conduct text mining and analysis for each cluster by using latent semantic indexing technology. Cheng Miao et al. studied MapReduce based user preference path mining algorithm [6-7].

In addition, many domestic scientific research institutions, universities and companies are committed to designing and implementing text mining system based on cloud computing. To a certain extent, they have provided decision makers with fast and accurate auxiliary decision information.

However, in the network text mining system, the parallelization of text classification process has not been well realized.

III. RESEARCH ON THE EXTRACTION OF CAMPUS NETWORK HOT TOPICS BASED ON CLOUD COMPUTING

The subject extraction system of network text corpus is mainly composed of three parts: text preprocessing, mining analysis and visualization, as shown in figure 1.

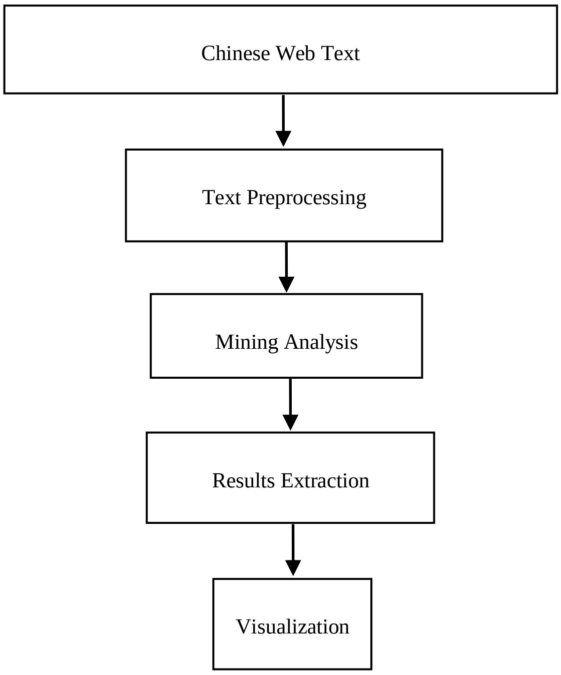

Fig. 1. Network text corpus subject extraction system.

This paper mainly studies the mining and analysis of massive text data. The main contents of the research include: process of text mining, major difficulties of text classification algorithm in mining analysis, analysis of cloud computing core technology and construction of Hadoop cloud platform. The text mining system model based on cloud computing is analyzed and designed to realize the parallelization and high efficiency of text classification process.

The main technologies involved in the subject extraction system of network text corpus include: cloud computing technology, preprocessing technology, text mining technology and visualization technology.

A. Cloud Computing Technology

Cloud computing technology, which is based on the Internet, integrates the characteristics of distributed processing and parallel processing; it can provide massive information storage and super computing capacity. The distributed parallel computing environment based on cloud computing is built on a large number of cheap PC cluster, because of its low cost, it can greatly reduce the cost of mass data processing. In addition, under the condition of parallel computing, nodes can be added conveniently and flexibly, thus, it can give the mining system strong fault tolerance [8]. Applying cloud computing technology to massive text mining can effectively meet the demand of time consumption and storage capacity in massive text data mining.

B. Text Preprocessing

Text preprocessing mainly includes two key techniques: word segmentation and feature extraction.

First, word segmentation technology: automatic word segmentation is the basis of Chinese information processing, is the essential text preprocessing. The word segmentation technology is based on dictionaries and rules, combined with the language model method based on probabilistic analysis to improve the accuracy of automatic word segmentation. At the same time, according to the different application environment, it customizes the individual character segmentation algorithm for specific requirements.

Second, feature extraction technology: using the vector space model to represent the characteristics of the network text vector, due to network text eigenvector dimension is higher, it usually cause a larger influence on the subsequent analysis of the data mining. As a result, combining feature extraction was needed to reduce network dimension of text feature vector and to reduce the time of machine automatic learning, improve the system of mining speed. Feature extraction algorithm refers to constructing an evaluation function to evaluate the features of each network text, and then queuing the features according to the evaluation score, finally, selecting the features having the higher scores as the optimized feature vectors.

C. Mining Analysis

Using text classification technology to realize text mining analysis .Text classification is to let the computer learn a classification model or function, the classification model can map the text vector to a certain category which is in a number of existing categories, so that data retrieval or query become more faster, and more accurate.

In text classification algorithms, the Support Vector Machine (SVM) method is a good classification method based on the principle of minimization of structural risk, which can take into account both classification accuracy and generalization ability. So it is suitable for text classification analysis of large scale and complex structure. Kernel function is a good way to solve the nonlinear separable problem of original vector space effectively [9]. Cloud computing technology can provide powerful data storage capacity and super computing capacity [10].

In the process of massive text mining, problems such as large text data size, complex data structure and nonlinear and gradable vector space exist. Using the binary tree to construct SVM multi-level classifier to identify the text category analysis, combining the kernel function to improve the performance of classifiers mining, meanwhile, adopting cloud computing to paralleling classification process, so as to speed up the classification, shorten the recognition time, finally, improve the speed and efficiency of text mining, and help policymakers accurately and timely make strategic decisions.

D. Visualization

Visualization refers to using computer graphics and image processing technology to transform network text data into graphics or images, which are displayed on the computer screen. It can be interactive processing with users.

E. Key Techniques and Methods

- Cloud computing environment: using open source Hadoop to build the cloud platform.
- Subject extraction of network text corpus: using support vector machine (SVM) to extract and analyze the subject of web text.
- Model establishment of network text corpus subject extraction system: using cloud computing technology to parallel topic extraction algorithm, so as to improve the traditional topic extraction process and the speed of subject extraction.

F. Experimental Environment and Means

- Experimental text data: Hankou University website text data.
- Analyzing and designing the model of network text corpus subject extraction system based on cloud computing.
- Building Hadoop cloud platform.
- Parallelizing the subject extraction algorithm.
- Digging out the hot topics that students care about quickly.
- Helping college teachers and administrators to make decisions and countermeasures timely.

IV. CONCLUSIONS

The in-depth promotion and popularization of modern education information technology makes the use of information technology to assist the teaching and management of university education becoming the normalcy in the management mode of education in today's universities.

This paper first introduces the importance, practical significance and function of using modern information technology to focus on the hot topic on campus network. Then, it mainly discusses how to apply cloud computing technology to campus network hot topic mining extraction system. By paralleling network text subject extraction process, the proposed method can improve the text extraction speed, so as to speed up the text subject extraction process, reduce the time consumption of hot topic mining, and then quickly help teachers and administrators in colleges and universities to timely grasp the hot topics that students care about, and eventually, to do a better job of actively guide and preventive measures.

ACKNOWLEDGMENT

This study is supported by the Science and Technology Research Planning Foundation of Education Bureau of Hubei Province of China (No. B2015507).

REFERENCES

[1] Zhou Yao. Research on Text Mining Technology Based on Cloud Computing [D]. Changsha: National University of Defense Technology, 2011. (In Chinese)

[2] Liu Peng, Cloud Computing [M]. Beijing: Publishing House of Electronics Industry, 2010. (In Chinese)

[3] Cao Zewen. Design and Implementation of JP Algorithm Based on MapReduce. http://wenku.baidu.c, 2015. (In Chinese)

[4] Liu Y, Wang Z L, Huang Y L. Research on Classification of Large-scale Text on GPU Platform [J]. Computer Engineering and Applications, 2012,vol. 48(8), pp.141-143.

[5] Chen Guangjing. Study and Implementation of Hadoop Small File Processing Technology [D]. Nanjing: Nanjing University Of Posts And Telecommunications, 2013. (In Chinese)

[6] Lu Tianwen. Deep Analysis and Prospect of Data Center [J]. Power of the World, 2013, pp.6-8.

[7] Viktor Mayer-Schonberger. Big Data: A Revolution That Will Transform How We Live, Work, and Think [M]. Hodder & Stoughton. 2013.

[8] Ren Z, Xu X, Wan J, et al. Workload Characterization on A Production Hadoop Cluster: A case study on taobao [C]. Workload Characterization (IISWC), 2012 IEEE International Symposium on.IEEE,2012:3-13.

[9] Duan Y. Application of SVM in Text Categorization [J]. Computer & Digital Engineering, 2012, vol.40(7), pp.87-88+149.

[10] Cai Ruicheng. Performance Optimization and Improvement of HDFS Based Small File Processing and Related MapReduce Computing Model [D]. Jilin: Jilin University, 2012. (In Chinese)