A Distributed Approach For Chinese Micro-blog Hot Topic Detection

Zhang Xiang
College of Information and Control Engineering
Xi'an University of Architecture and Technology
Xi'an, China
Zhangxiang1001@126.com

Dong Lili
College of Information and Control Engineering
Xi'an University of Architecture and Technology
Xi'an, China
donglilixjd@163.com

Lin Ruitao
College of Information and Control Engineering
Xi'an University of Architecture and Technology
Xi'an, China
torrylin88@gmail.com

Wang Ru
College of Civil Engineering
Xi'an University of Architecture and Technology
Xi'an, China
wangru@xauat.edu.cn

Abstract—In consideration of the features of micro-blogging content such as short text, sparse feature words and the huge scale, a method to detect micro-blogging hot topic was proposed in this paper based on MapReduce programming model. This method first employs the hidden topic analysis to solve the problem of short micro-blogging content and sparse feature words. Then the CURE algorithm is used to alleviate the problem that the Kmeans algorithm is sensitive to the initial points. Finally, the hot topic clustering results are obtained through the parallel Kmeans clustering algorithm based on the MapReduce programming model. The experimental results show that proposed method can effectively improve the micro-blogging hot topic detection efficiency.

Keywords- Micro-blog; MapReduce; Kmeans clustering; Hidden topic model

I. INTRODUCTION

With the development of Internet and the emergence of Web2.0, micro-blog has gradually become an important tool for people to communicate with each other and make speech. As of 2012 December, the number of Sina Weibo's registered users has reached 503 million, daily active users was 46.2 million, in 2012 February, the daily micro-blogging number has more than 100 million. Other domestic micro-blog websites, such as Tencent micro-blog (t.qq.com), NetEase micro-blog (t.163.com) etc, the number of whose registered user and micro-bloggings also grow quickly. These micro-blog services all have the characteristic of cross platform, which allows users can use various terminals to express their views, record their life knowledge and get to know the relatives and friends of the latest whenever and wherever possible. Micro-blog possess the strong characteristic of real-time, lead to the spread speed of emergencies and hot news in social reality in the micro-blog, significantly faster than traditional media, newspapers, radio and television. Therefore, timely finding a hot topic in the micro-blog is a hot issue in the public opinion monitoring, information security and other fields at present, which is significant to maintaining social stability and harmonious development..

Traditional topic detection and tracking (TDT) technology laid a foundation for the research of discover the hot topics in micro-blog, which main research object is the Newswire, radio, television, the Internet and other media, what news reports may be longer, which research data mainly uses the TREC conference TDT corpus, contained a smaller scope. while micro-blog have the characteristics of short content, small and sparse feature words, large scale and others, so the traditional TDT technology can't effectively applied to micro-blog news.

According to the characteristics of micro-blog news, the research of discover the micro-blog hot topics need to introduce some new handling methods [1-3]. The literature [2] is proposed a method of finding topics based on Latent Semantic Analysis (LSA), effectively solving the problems of high dimensionality and synonym polysemy in the traditional vector space model. The literature [3] presents a large-scale Twitter data sets, taking the advantages of implicit theme analysis technology, preferably solving the problem of micro-blog short text data sparsity, however they all have not come up with a method to resolve the clustering speed effectively.

This paper put forward a method of disposing the micro-blog data and finding the hot micro-blog topics, which combined with the Latent Dirichlet Allocation (LDA) model and the MapReduce programming model, in condition of ensuring the clustering precision, effectively improved the efficiency of clustering algorithm.

II. DISCOVER MICRO-BLOG HOT TOPIC BASED ON MAPREDUCE PROGRAMMING MODEL

A. Method of thought and basic framework

In order to find hot topics from the mass of micro blog short text, we need to solve two problems: first, reasonably modeling the micro-blog short text, in order to calculate the similarity between the micro-blog; second, accurately and fast clustering the mass micro-blog data, in order to timely discover the purpose of the hot topics.

We first carry on the pretreatment to the micro-blog, namely word segmentation and get rid of the stopword. The word segmentation system, this paper employs the ICTCLAS50 word segmentation system which is developed by Chinese Academy of sciences. Then latent themes modeling the micro-blog data fully excavated the latent themes, in order to reduce the influence that micro-blog data's sparsity to the calculation of text similarity. Next, using CURE clustering algorithm, we make the preliminary clustering to the micro-blog data, where we obtain the input parameters of K-means algorithm, means the clustering number and the corresponding initial class centers. Finally, using the K-means clustering algorithm based on the MapReduce programming model, fast clustering of data, we can get the hot topics. The process is shown in figure 1.

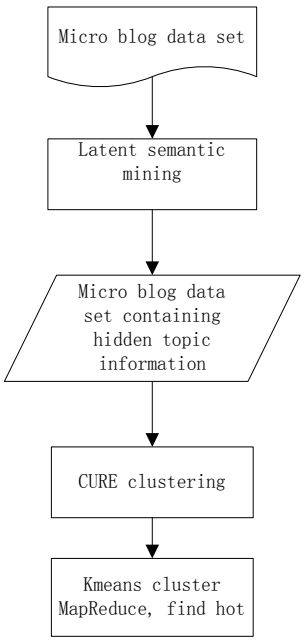

Figure 1. Processing flow chart

The introductions of the 3 steps mentioned are as follows.

B. Data preparation and pretreatment

Since at present there has not a universal data set in the micro-blog data mining, this paper use the Tencent micro-blog provided open API, randomly caught 22471 users, from April 13, 2013 to 2013 April 21,totally 9 days, published 331065 micro-blogging data.

After preliminary observation of the data, which was found to contain a lot of very short text, mainly includes the expression micro-blog message, exclamation micro-blog messages, there are also many messages related to Tencent games, these micro-blog messages have no practical significance in finding the hot topics of the micro-blog. So, in order to improve the data validity, we simply clean up the data, namely remove messages which length is less than 4 and themed to Tencent game, after the processing, we obtain 182162 micro-blog message texts; then the paper used the ICTCLAS50 segmentation system developed by the Chinese Academy of Sciences computing method, processing the micro-blog data word; and then make stopword removal to the results of the word segmentation.

C. Latent themes mining

1) LDA mode:

LDA (Latent Dirichlet Allocation) is a kind of three layer tree Bayesian probabilistic generative model, which consists of document layer, theme layer, word layer, as shown in figure 2. It is based on this assumption: 1, there is a number of K independent theme of each other in document layer; 2, each theme is the multinomial distribution over words; 3, the document by the K theme random mixture; 4, each document is the multinomial distribution of K themes; 5, each document in the subject of probability distribution by a Dirichlet distribution.

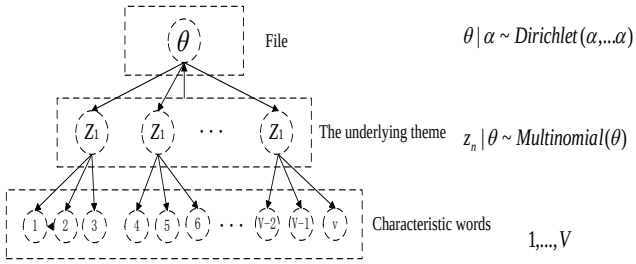

Figure 2. LDA Bayesian probabilistic generative model

The three arborescence Bayes probabilistic generative model, θ is a K dimensional vector, indicated the multinomial distribution on the topic document collection. Set $\theta = [\theta_1, \theta_2, \ldots, \theta_k]$, then $\theta_1 + \theta_2 + .. \theta_k = 1$, and $0 \le \theta_k \le 1$, $1 \le k \le K$; α shown as in the ducument layer that corresponding to θ K dimensional Dirichlet hyper-parameter, set $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_k]$。The LDA generation model can be shown in figure 3 with Bayesian network diagram.

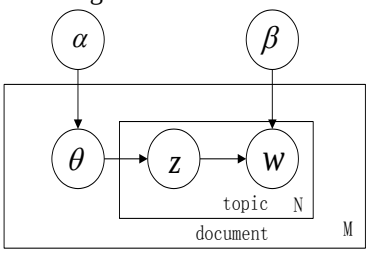

Figure 3. The LDA graph model

For a document collection, document set parameters α and β are respectively sampled one time; for each document, document level parameter θ is sampled one time; for each word in a document, words parameters Z and W are respectively sampled one time. In this way, the document generation process of the LDA model is [7]:

Step 1, extract N words, N obeys the Poisson distribution;

Step 2, extract θ (the vector of subject probability), θ obeys the Dirichlet distribution;

Step 3, for every w_n in N words,

(a) Generate the theme z_n, z_n obeys the multinomial distribution:

(b) Generate w_n from $p(w_n \mid z_n, \beta)$, w_n obeys the subject z_n conditions.

The β is a $k \times V$ matrix, denoted as $\beta = p(w^j = 1 | z^i = 1)$, which recorded a subject condition to generate a probability of a word, here is seen as a consistent with variable. More description about LDA model please reference [4-5].

2) LDA parameters estimation and modeling results

About parameter estimation for LDA model, this paper adopts Gibbs[6] sampling methods, using GibbsLDA++ to model the above micro-blog short text data, according to reference[3], the latent theme number is 200, the initial hyper parameters choose 0.25, β choose 0.1. Through the operation, we can obtain the following 5 files:

- *.others：This file contains the input parameters of LDA model；
- *.phi ： This file contains a word-topic distribution matrix Φ, means a $K \times V$ matrix, K means the number of latent topics and the micro-blog data set, V represents all the different word number in a micro-blog data set. Each meaning of the value is $P(words_w | topic_t)$, namely the frequency of each latent themes generated for each word;
- *.theta ： The file contains a theme-document matrix Θ, which is a $M \times K$ matrix, M represents the total number of micro-blog data text, namely the sum of micro-blog messages; K represents the number of latent topics micro-blog data set. Each meaning of the value is $P(topic_t | document_m)$, which is the probability that micro-blog data set on each text generation latent themes;
- *.tassign：The content of the file is the subject of distribution of training data set of words, each row is a list consisting of $< word_{ij} > : < topic\_of\_word >$, namely belongs to each word in a micro-blog hidden themes.
- *.twords ： This file contains the information, which is after the GibbsLDA++ modeling, all the generated subjects and these vocabulary included.

D. A preliminary clustering of modeling results

CURE[8] algorithm is a clustering algorithm based on hierarchy, which is different from the traditional method that use one object to represent the cluster, instead of using multiple objects, and through the shrinkage factor to regulate cluster shape, as a result, it can deal with non-spherical distribution or size of uneven distribution of cluster, and has a good effect on treatment of isolated points, so in this paper micro-blog data use the CURE algorithm to initializing clustering, to get the input parameters of K-means algorithm: the number of clusters and the corresponding initial class centers, so as to reduce the clustering fluctuation problem of randomness and apriori of K-means initial clustering center. The procedure is shown as Algorithm 1.

Algorithm 1 CURE clustering

Input: the topic-document distribution matrix Θ

Output: the number of cluster K, the vector of K clustering center value

Step:

1）Obtained from the previous step of the theme - document matrix Θ, then extract sample S;

2) Divided the sample S into isometric n parts, locality cluster each partition of it;

3) Through random sampling to remove isolated point, slowly growth or no growth clusters;

4) Cluster the locality clusters;

5) Marked the corresponding cluster with the corresponding cluster labels;

6) Calculate the average samples value of each category to obtain the corresponding class center.

E. Clustering of modeling results

1) The basic idea of MapReduce

MapReduce[9-11] is a Google development which is a parallel programming model for processing and efficient task scheduling model for large scale data set, which has been widely used in the field of log analysis, mass data sorting, data searching etc. MapReduce mainly through Map and Reduce -- the two steps to concurrently handle the massive data. Map is a process of decomposition, firstly divided the large data sets into hundreds of small independent data set (splits), then each (or some) data set is allocated to one nodes in the cluster (usually mentioned of a general computer) for processing; while Reduce is a complicated process, it will integrate separated data into together and return the output. The MapReduce model is shown in figure 4[12].

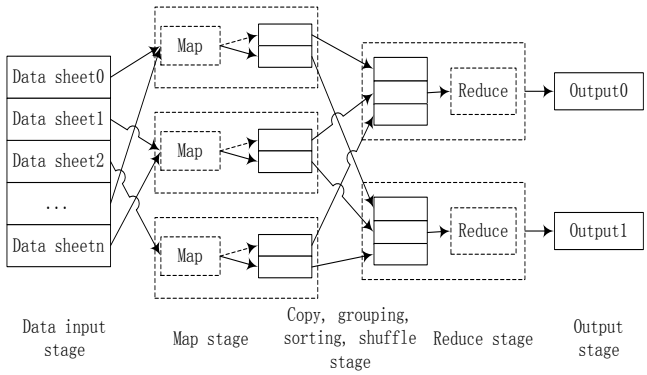

Figure 4. Mechanism of MapReduce

Firstly MapReduce process divided the input file into M parts; then the main control program--Master distribute each of data block to the free Worker node so as to conduct Map process, every Map is independent of each other and highly parallel, which uses the form <Key, Value> to write to the local hard disk; next the Master distribute the Worker node to conduct the Reduce process, calling the remote procedure to read the intermediate results of all Map node in local hard drive in form of <Key, Value>, and then sorted the results according to the key words, collected the information of the same key,

following the Reduce machine transfer the keywords and related intermediate results set to the Reduce function which is defined by users, finally Reduce function composite the summary data to an output file. When all Map and Reduce task is completed, the Master node will return all Reduce results to the user program.

2) K-means clustering based on MapReduce programming model

The parallel idea of K-means algorithm is: Once the algorithm iterated, we start a MapReduce calculation process, namely to realize goal that proceed the parallel computation at each iteration inside, the main task of Map function is calculated the distance for each record to the class center and mark it or anew label the category of it. The main task of Reduce function is to calculate the center point of the new class, according to the obtained intermediate results, and then put the center point set to the next MapReduce process. The algorithm's procedure is shown as the following steps:

Algorithm 2 K-means clustering based on MapReduce programming model

Input: the topic-document distribution matrix Θ and the clustering center and number of cluster type K after CURE method.

Output: K cluster C_i

Steps:

1) The K center point cluster after CURE algorithm serve as the initial cluster center O_i;

2）Repeat

3) Execute Map function, calculate distance of each point to the center of the cluster, label or anew label their categories;

4) Execute the Reduce function, calculate the new center of the cluster, and then use the new center of the cluster instead of the original one;

5) Calculate the sum of squares D of the distance of two center of cluster;

6) Until D is less than a given threshold

The input of the Map function is the data of all waiting clustered and the last round cluster center of iteration (or initial cluster), input the data <Key, line Value> corresponded form <the line number of the theme-document matrix Θ, record attribute vector>; each Map function read the file that describe the clustering center, then calculate the distance of this recording point to all of the center, find out the shortest distance, which is served as the new category of this point; output the intermediate results <Key, Value> corresponded form <clustering category ID, record attribute vector >.

The Reduce function's input data<Key, Value> corresponded form is <clustering ID, {the set of record attribute vector}>; all the same records Key(that is, the record of same category ID) send to a Reduce task – accumulate the point number of same Key and the sum of each record component, calculate the mean of each component, get a new descriptive file of clustering center; output the result<Key, Value> corresponded form < clustering category ID, mean vector >.

III. EXPERIMENTAL RESEARCH AND ANALYSIS

The experimental environment is consisted of 6 common PC based on the cluster of Hadoop platform, which configuration is as follows:

CPU: 4 * Intel Core I7-2600 @3.40G
Memory: 4G
Hard disk: 1T 7200 RPM IDE
Operating system: 32 Ubuntu 11.04
Hadoop version: 1.0.1
The network configuration: 1000Mbps

Experiment 1: Filter the micro-blog messages in April 20, 2013, a total number of 21324, according to the method of this paper to conduct cluster, as the results in Table 1, the left of table 1 are some hot topic in a micro-blog website at the same time, while the right side are the automatic extracted hot topics using the algorithm in this paper. As can be seen from table 1, automatic extracted hot topics and online collection of hot topics are not totally same, may be the different between this paper and their ranking rules, or due to some hot topic is initiated by their own web site, however it can accurately reflect the intraday hot topics of micro-blog. Maintaining the Integrity of the Specifications

TABLE I. IN APRIL 20, 2013 THE ONLINE COLLECTION OF HOT TOPIC AND AUTOMATIC EXTRACTED HOT TOPICS

ranking	*online collection of hot topics*	*automatic extracted hot topic*
1	Earthquake in Ya 'an sichuan province	Si'Chuan, Ya'an, Earthquake, Die, Lushan
2	Zhigen Zhu injured Yang Sun again	Si'chuan, Ya'an, cheer supplication, safety, love wishes
3	Poisoning case in Fudan University has been solved	Spread, cold, radix isatidis H7N9, virus, specialist disinfect
4	Explosion's suspect on the run in Boston were arrested	Fudan University, Poisoning, postgraduate, jealous, roommate
5	Seven work days in next week	Boston, explode, sorry, malathion

Experiment 2：After simply cleaning, during the pretreatment, the LDA modeling and CURE clustering of the 182162 micro-blog text, then use the 1 to 5 nodes test the distributed Kmeans' text clustering efficiency based on MapReduce programming model, the results as shown in Figure 5, 6. As you can see from Figure 5, with the increased number of nodes in the cluster, its running time reduce gradually. Seen from Figure 6, with the increased number of nodes in the cluster, the speed-up ratio become larger gradually, but at the same time the running time become slow gradually, mainly because the amount of communication between nodes and the spending has became bigger. Data analysis shows that the Kmeans algorithm based on the MapReduce programming model can provide a high clustering efficiency, and has good speed-up ratio.

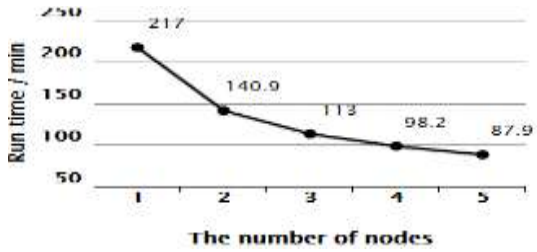

Figure 5. The running time of different number of machines clusters to do the text clustering

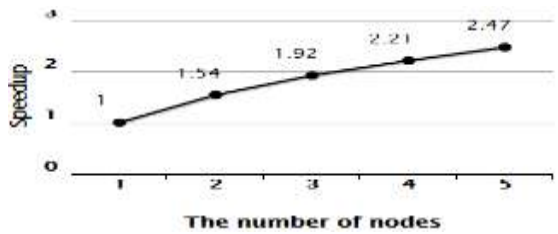

Figure 6. The speed-up ratio of different number of machines clusters to do the text clustering

IV. CONCLUSION

In this paper we studied how to discover the hot topics in mass micro-blog messages rapidly and precisely, in this paper, using the method of modeling latent topics, effectively solve the problem of short text data's sparsity, and then use the CURE algorithm, effectively solve the sensitive issue that K-Means algorithm choose the initial points, finally use K-means algorithm based on the parallel MapReduce algorithm, partly improve the efficiency of cluster, but also cannot give full play to the performance of MapReduce. The next work mainly concentrates on the following two aspects: first, solve the performance problem in MapReduce, give the full play to the performance of MapReduce; second, make full use of performance of the sub node, add the GPU to the calculation.

ACKNOWLEDGMENT

The work is supported by the National Natural Science Foundation of China under Grant No. 51278400; Natural Science Foundation of Shaanxi Provincial under Grant No. 2012JM8042; Natural Science Foundation of Shaanxi Provincial Department of Education under Grant No. 12JK0940; Xian Project to Promote Technology Transfer under Grant No. CXY1348-(1) ; Science and Technology Plan Projects of Yulin city under Grant No. 12-2-07.

REFERENCES

[1] Zhang Chenyi, Sun Jianling, Ding Yiqun, Topic Mining for Microblog Based on MB-LDA Model [J]. Journal of Computer Research and Development,2011,10:1795-1802.

[2] Deng Yigui, Ma Wen Wen. Micro-blog topic detection method based on Latent Semantic Analysis. [J]. Computer engineering and Applications, 2012.

[3] Lu Rong, Xiang Liang, Liu Mingrong, Yang Qing. Discovering News Topics from Micro-blogs Based on Hidden Topics Analysis and Text Clustering[J]. Pattern recognition and artificial intelligence,2012,03:382-387.

[4] Blei D M, Ng A Y. Latent Dirichlet Allocation [J]. The Journal of Machine Learning Research, 2003,3:993－1022.

[5] Shi Jing, Li Wanlong. Topic Words Extraction Method Based on LDA Model [J]. Computer engineering,2010,19:81-83.

[6] Thomas L. Griffiths, Mark Steyvers. Finding scientific topics[J].Proceedings of the National Academy of Sciences of the United States of America,2004,101(suppl 1):5228-5235.

[7] Li Bo. Analysis Model of Medical Text and Image based on LDA and LSA and its Application [D]. Journal of Jilin University,2012

[8] Guha S et al. CURE: An efficient clustering algorithm for large databases. In: Proc of the ACM SIGMOD Int' l Conf on Management of Data. 1998.

[9] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters [J]. Communications of the ACM, 2005, 51(1): 107-113.

[10] Wu Xue Jiang, Zhang Jing, Wang Zhiming. Study on Parallel Programming Framework Model Based on MapReduce[J]. Microelectronics & Computer,2011,06:168-170+175.

[11] Xu Xiaolong, Wu Jiaxing, Yang Geng, Cheng Chunling, Wang Ruchuan. Mass data processing system based on large-scale low-cost computing platform [J].Application Research of Computers,2012,02:582-585

[12] Zhang Xueping, Gong Kangli, Zhao Guangcai. Parallel K-Medoids algorithm based on MapReduce [J]. Journal of Computer Applications,2013,04:1023-1025+1035.