MICROBLOG MINING BASED ON CLOUD COMPUTING TECHNOLOGIES: MESOS AND HADOOP

[1]KONG XIANGSHENG
[1]Department of Computer & Information, Xinxiang University, Xinxiang, China
E-mail: victor_kong@163.com

ABSTRACT

The content analysis of microblogs has increasingly become a focus for academic research. By microblog mining, microblogs provide either first-person observations or bring relevant knowledge from external sources in emergency situations such as Wenchuan earthquake in China. We introduce the architecture for microblog mining based on cloud computing and present Mesos which shares microblogging resources in a fine-grained manner and an improved Apriori algorithm about association rules mining. The proposed design and implementation provide valuable reference for the implementation of microblogging cloud storage system and microblog mining.

Keywords: *Microblog Mining; Association Rules; Apriori Algorithm; Text Mining Algorithm*

1. INTRODUCTION

Microblog is a relatively new phenomenon in the web services world of user generated content. It is a new form of lightweight chat through which users can describe things of interest, post and exchange short messages, express attitudes that they are willing to share in short posts (such as Twitter). These posts are distributed via instant messages, mobile phones, email, or the Web [1]. These micro-branding posts are immediate, ubiquitous, and scalable. Since they are online, they are also typically accessible by anyone with an Internet connection. There are also archival in the sense that these microblogs permanently exist and are searchable via web search engines and other services [2].

Early work of microblog mainly focused on its user relationship and community structure. Others such as Krishnamurthy studied user behaviors and geographic growth patterns of Twitter. Only a few researches on content analysis of microblogs were proposed recently, that is mainly because traditional text mining algorithms, which are suitable for old corpora, cannot model microblog data very well without considering its inner structured information on social network [3].

Although most posts are conversation and chatter, they are also used to share relevant information and report news. By microblog mining, microblogs provide either first-person observations or bring relevant knowledge from external sources in emergency situations such as Wenchuan earthquake in China. Information from official and reputable sources is regarded as valuable and hence is actively sought and propagated. Other users then elaborate and synthesize this pool of information to produce derived interpretations [4]. Microblog is becoming a valuable tool in disaster and emergency situations as there is increasing evidence that it is not just a social network, it is also a news service.

2. RELATED WORKS

Mesos is a cluster management platform for distributed applications and frameworks. This is complementary to our work. Our focus on microblogging resource isolation and two-level microblogging resource management should provide better support for Mesos and similar frameworks. One is System Resource management, the other is Cluster management [5].

Mesos decides how many microblogging resources to offer each framework, based on an organizational policy such as fair sharing, while frameworks decide which microblogging resources to accept and which tasks to run on them. While this decentralized scheduling model may not always lead to globally optimal scheduling, we have found that it performs surprisingly well in practice, allowing frameworks to meet goals such as data locality nearly perfectly. In addition, microblogging resource offers are simple and efficient to implement, allowing Mesos to be highly scalable and robust to failures. Mesos also provides other benefits to practitioners [6]. First, even organizations that only use one framework can use Mesos to run multiple instances of that framework in the same cluster, or multiple versions of the framework.

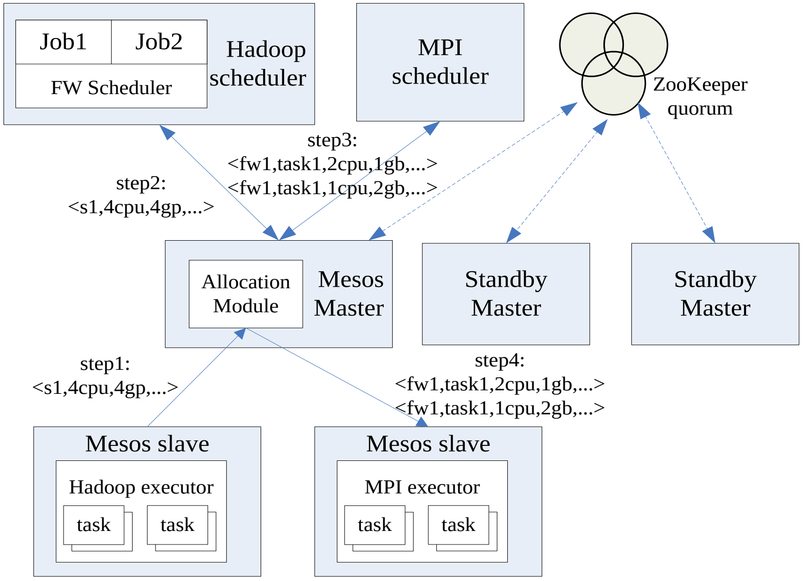


Fig.1 Mesos And Hadoop Resource Offer Example

As shown in Fig.1 we use Mesos internally as an end-to-end framework for deploying some of application services in microblogging system. Using Mesos for some of their services appealed to microblogging system for many reasons, including:

Flexible deployment: Statically configuring where services should run makes it difficult for different teams within microblogs to operate autonomously. By leveraging Mesos, engineering teams can focus on doing code deploys against a generic pool of resources, while the operations team can focus on the operating system and hardware (e.g., rebooting machines with new kernels, replacing disks, etc).

Increased utilization: Many services within the cluster are sharded for better fault-tolerance and do not (or cannot) fully utilize a modern server with up to 16 CPU cores and 64 GB of memory. Mesos enables microblogs to treat machines as a pool of resources and run multiple services on the same machine, yielding better overall cluster utilization.

Elasticity: Certain services might want to “scale up” during peak or unexpected events when traffic and load has increased. Using Mesos, it’s easy for different services to consume more or less resources as they are needed [7].

3. ARCHITECTURE

Based on cloud computing technologies Mesos and Hadoop, we designed the architecture of microblog mining which is divided into five modules: System Resource Layer, Cluster Management Layer, Resource Service Layer, Data Mining Layer and Web Access Layer (shown in Fig. 2).

(1)System Resource Layer: This is the most fundamental part of the architecture. This layer provides storage resources for microblog mining. Due to processing large-scale and distributed resources, we choose HDFS which is the file system component of Hadoop to store file system metadata and application data separately.

(2)Cluster Management Layer: An important issue to be addressed is how to effectively organize different frameworks distributed. Unfortunately, sharing a cluster efficiently between two or more of these frameworks is difficult. Mesos which achieves these goals is a platform that enables fine-grained, dynamic resource sharing across multiple frameworks in the same cluster.

(3)Resource Service Layer: Resource Service Layer is the important and difficult part for implementation of the architecture. This layer provides application services such as text preprocessing, microblog indexing, text extraction from HTML pages, and an API allowing the user to easily implement data persistence [8]. Resource Service Layer uses different techniques including microblog AIP and web crawler to implement efficient and reliable resource service. This API provides an interface to extract relevant data from the social network sites, such as Twitter, or performing complicated interactions with AJAX-based applications, or identifying Web objects in particular Web application [9]. Web crawler is responsible for extracting the textual content of posts, indexing texts for the search process and removing the information considered irrelevant to the microblog analysis.

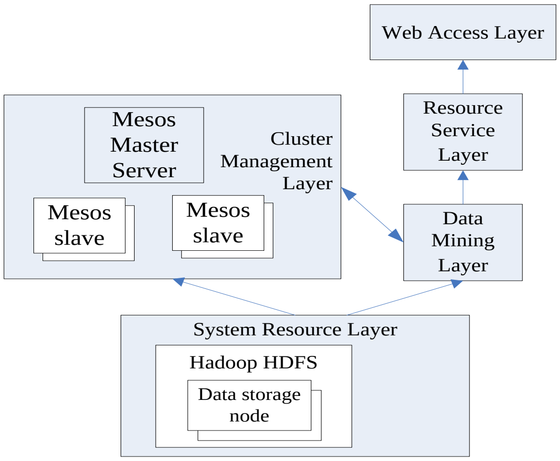


Fig.2 The Architecture Of Microblog Mining Based On Cloud Computing

(4) Data Mining Layer: Data Mining Layer is the core part and also the most difficult one for implementation of the architecture. This layer is responsible for discovering potential rules, useful knowledge method and the technology from the massive microblogging data. The Hadoop programming model is designed for the processing and generation of large data sets and allows a high degree of parallelism [10]. Based on data partitioning method is generally adopted. First of all the large data sets are divided into appropriate sub-blocks, and then each sub-block is processed using traditional mining algorithms (such as Apriori algorithm), finally the results of each sub-block is merged into the database.

(5)Web Access Layer: Web Access Layer is the gateway for users to implement a flexible and easy-to-use WEB access interface [11]. The results of data mining are stored into the database, and Web Access Layer directly interacts with database. Any authorized user could access different services provided through Web Access Layer.

4. MICROBLOG MINING DESIGN AND IMPLEMENTATION

In this Section, we introduce a approach that integrates the Hadoop and Mesos with scalable data mining techniques in Microblogging, and Apriori association rules mining algorithm, a popular data mining algorithm. Association Rules can help to uncover hidden or previously unknown associations. A rule in the form of A => B, denotes an implication of element or item B by item A. Association rules have successfully been used in a wide range of domains, e.g., market basket analysis, law enforcement, biotechnology [12].

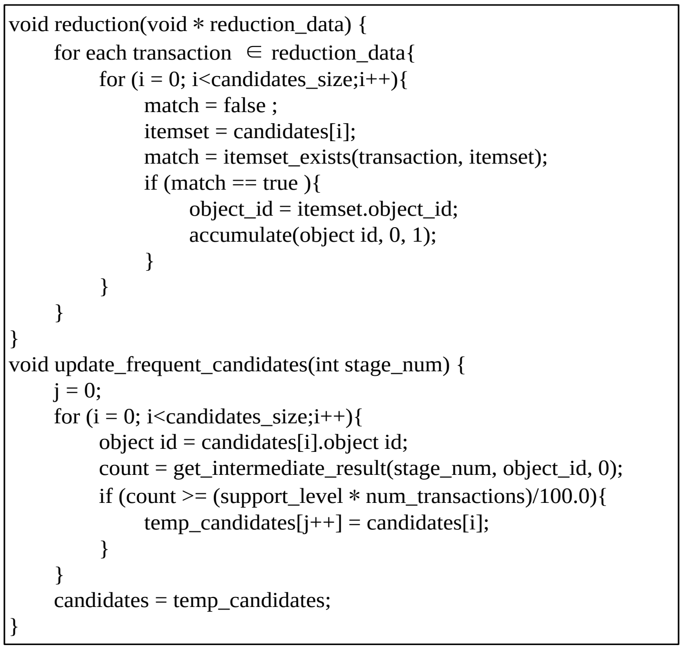
Association Rules Mining Algorithm. Association rule mining that implies a single predicate finds interesting association or correlation relationship among a large data set of items. In data mining it is one of the techniques used to extract hidden knowledge from datasets, which can be put to good use for business profit and is referred as a single dimensional or intra dimension association rule, since it contains a single distinct predicate with multiple occurrences (the predicate occurs more than once within the rule). The terminology of single dimensional or intra dimension association rule is used in multidimensional database by assuming each distinct predicate in the rule as a dimension [13].

The input microblogs may be filtered according to the different annotation types. After pre-processing, the system allows for Frequent Pattern Mining (FPM) via the Hadoop Map Reduce framework before we extract the actual rules.

In general, Apriori is used an influential algorithm for mining frequent itemsets for generating Boolean (single dimensional) association rules. A classical Apriori algorithm exits in association rules which show that the condition when the attribute or value in assigns data that frequently appears together. Apriori algorithm which is a kind of data mining algorithm uses a circular system to rake though one gradation in turn to complete frequent itemsets mining. Apriori Algorithm can be described as follows (shown in Fig. 3):

5. EXPERIMENT AND RESULTS

The association rules mining is the base of the data mining models built. The quality of this algorithm lies in the level of efficiency. To mine the massive data not only need the optimization algorithm, but also with considerable hardware conditions to complete the work. Here, we take testl, test2 four data sets four example, each contain 1000,4000,15000,20000 records, the minimum support of 3%, mining frequent 5 - item set. To use of these two algorithms, the time spent is shown in Fig.4.



```
void reduction(void ∗ reduction_data) {
    for each transaction  ∈ reduction_data{
        for (i = 0; i<candidates_size;i++){
            match = false ;
            itemset = candidates[i];
            match = itemset_exists(transaction, itemset);
            if (match == true ){
                object_id = itemset.object_id;
                accumulate(object id, 0, 1);
            }
        }
    }
}
void update_frequent_candidates(int stage_num) {
    j = 0;
    for (i = 0; i<candidates_size;i++){
        object id = candidates[i].object id;
        count = get_intermediate_result(stage_num, object_id, 0);
        if (count >= (support_level ∗ num_transactions)/100.0){
            temp_candidates[j++] = candidates[i];
        }
    }
    candidates = temp_candidates;
}
```

Fig.3 Pseudo-Code For Improved Apriori Algorithm

It can be seen from Fig.4, when the data is small, the improved Apriori algorithm to improve execution time greatly. Traditional methods often read the database, take a lot of system resources. Improved algorithm without traversing the database computing support, but the algorithm complexity increases, while the data is large, taking up considerable memory and processor resources, computation time is not significantly improved. The result showed the high performance of our Hybrid Recommender System after training. In addition, it was also surprising to see the backoff model perform worse than the unigram multinomial model, although this is understandable since bigram counts were too sparse with this training set.

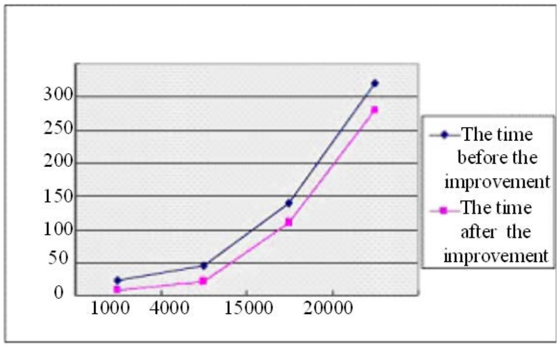


Fig.4 Apriori Algorithm Comparison Chart Before And After

6. SUMMARY AND FUTURE WORK

The paper aims for approach to microblog mining design and implementation based on cloud computing and discusses the association rules and the Apriori algorithm. In this paper, we have presented Mesos, which currently run Hadoop for fast in-memory parallel computing, fine-grained sharing of clusters among different storage resources for microblogs. And we have improved the efficiency of mining frequent itemsets when using mining association rules algorithm, puts forward an improved Apriori algorithm. In future work we will look into building personalized microblog retrieval models, for which geolocalized information on microblog consumption might serve to incorporate cultural species.

REFRENCES:

[1] Bernard J. Jansen,Mimi Zhang,Kate Sobel,Abdur Chowdury, Micro-blogging as Online Word of Mouth Branding, Proceedings of the 27th international conference extended abstracts on Human factors in computing systems. 2009. 3859–3864.

[2] Bernard J. Jansen, Mimi Zhang,The Commercial Impact of Social Mediating Technologies: Micro-blogging as Online Word-of-Mouth Branding, Proceedings of the 27th International Conference on Human

Factors in Computing Systems, ACM, 2009. 3859-3868.

[3] Chenyi Zhang, Jianling Sun, Large Scale Microblog Mining Using Distributed MB-LDA, Proceedings of the 21st international conference companion on World Wide Web,2012, 1035-1042.

[4] France Cheong, Christopher Cheong, Social Media Data Mining: A Social Network Analysis of Tweets During The 2010-2011 Australian Floods, Proceedings of Pacific Asia Conference on Information Systems, 2011, 46-62.

[5] Barret Rhoden, Kevin Klues, David Zhu, Eric Brewer, Improving Per-Node Efficiency in the Datacenter with New OS Abstractions, Proceedings of the 2nd ACM Symposium on Cloud Computing, 2011, 200-207.

[6] Benjamin Hindman, Andy Konwinski, Matei Zaharia, Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center, Proceedings of the 8th USENIX conference on Networked systems design and implementation, 2011, 22-35.

[7] BENJAMIN HINDMAN, ANDY KONWINSKI, MATEI ZAHARIA, ALI GHODSI, ANTHONY D.JOSEPH, RANDY H.KATZ, SCOTT SHENKER, ION STOICA, Mesos Flexible Resource Sharing for the Cloud, Mesos, 2011, 37-45.

[8] R. Ferreira, RJ Lima, II Bittencourt, DM Filho, O. Holanda, E. Costa, F. Freitas, L. Melo. A framework for developing context-based blog crawlers. Proceedings of the IADIS International Conference on WWW/Internet, 2010, 120–126.

[9] Guandong Xu, Yanchun Zhang, Lin Li, Web Content Mining, Web Information Systems Engineering and Internet Technologies, 2011, 71-87.

[10] Philipp Berger, Patrick Hennig, Justus Bross, Christoph Meinel, Mapping the Blogosphere —Towards a Universal and Scalable Blog-Crawler, IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, 2011, 672-677.

[11] Mr. Yogesh Pingle, Vaibhav Kohli, Shruti Kamat, Nimesh Poladia, Big Data Processing using Apache Hadoop in Cloud System, International Journal of Engineering Research and Applications, 2012, 475-480

[12] Robert Neumayer, George Tsatsaronis, TRUMIT: A Tool to Support Large-Scale Mining of Text Association Rules, Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases, 2011, 646-649.

[13] Debajyoti Karmaker, Hafizur Rahman, Mohammad Saiedur Rahaman, Md. Kamrul Bari, A Fine Grained Technique for Viral Marketing based on Social Network: A Machine Learning Approach, International Journal of Science and Technology, 2011, 89-95.