BC-PDM: Data Mining, Social Network Analysis and Text Mining System Based on Cloud Computing

Le Yu
Beijing University of Posts and Telecommunications
Beijing 100876, China
yuleyule1987@gamil.com

Jian Zheng
Beijing University of Posts and Telecommunications
Beijing 100876, China
zhengjian8972@gmail.com

Bin Wu, Bai Wang
Beijing University of Posts and Telecommunications
Beijing 100876, China
wubin,wangbai@bupt.edu.cn

Chongwei Shen
Beijing University of Posts and Telecommunications
Beijing 100876, China
wsscwqx135@sina.com

Long Qian
Beijing University of Posts and Telecommunications
Beijing 100876, China
qianlonglarry@qq.com

Renbo Zhang
Beijing University of Posts and Telecommunications
Beijing 100876, China
zhrbocool@gmail.com

ABSTRACT

Telecom BI(Business Intelligence) system consists of a set of application programs and technologies for gathering, storing, analyzing and providing access to data, which contribute to manage business information and make decision precisely. However, traditional analysis algorithms meet new challenges as the continued exponential growth in both the volume and the complexity of telecom data. With the Cloud Computing development, some parallel data analysis systems have been emerging. However, existing systems rarely have comprehensive function, either providing data analysis service or providing social network analysis. We need a comprehensive tool to store and analysis large scale data efficiently. In response to the challenge, the SaaS (Software-as-a-Service) BI system, *BC-PDM* (Big Cloud-Parallel Data Mining), are proposed. *BC-PDM* supports parallel ETL process, statistical analysis, data mining, text mining and social network analysis which are based on Hadoop. This demo introduces three tasks: business recommendation, customer community detection and user preference classification by employing a real telecom data set. Experimental results show that *BC-PDM* is very efficient and effective for intelligence data analysis.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database application-Data Mining

General Terms

Algorithms, Design, Experimentation

KDD'12, August 12–16, 2012, Beijing, China.

Keywords

Parallel data mining, Hadoop, SNA, Text mining

1. INTRODUCTION

Telecom BI(Business Intelligence) system consists of a set of application programs and technologies for gathering, storing, analyzing and providing access to data, which contribute to manage business information and make decision precisely. The telecommunication operator generates huge count of information and data. The study focus on mining potential significant information from massive mobile communication data efficiently. Current application mode has the following problems:

- Data storage and access: A middle level branch of CMCC(China Mobile Communication Corporation) will be 12-16 TB in a year[1], and there are many other data including network data, customer data and so on. Traditional data storage uses commercial database or data warehouse system, such as Oracle. When facing with large scale dataset, these storage solutions are unable to ensure high scalability and availability.

- Data analysis and process: SAS Enterprise Miner and SPSS Clementine are traditional BI systems. However, these tools are low scalability and high cost. With the Cloud Computing development, some parallel data analysis systems have been emerging[2]. Mahout [3] is a collection of scalable machine learning algorithms based on Hadoop, but it lacks an user-friendly interface and its learning cycle is long. Radoop[4] integrates HDFS, Hive and Mahout functionalities in the RapidMiner environment. PEGASUS[5] and Giraph[1] utilize the parallel computing framework to implement graph mining. However, these projects have single function and they cannot be used as BI system. Even it lacks the basic function of BI system, ETL.

In response to these challenges, the *BC-PDM* is proposed which is based on Hadoop. *BC-PDM* provides access to

[1]http://incubator.apache.org/giraph/

large telecom data and business solutions for telecom operators. The system is an important project of Big-Cloud(BC) which is proposed by CMRI(China Mobile Research Institute). It combines ETL, data mining (DM), social network analysis (SNA), statistical analysis (SA) and text mining (TM) with MapReduce. This demo has the following contributions:

- *BC-PDM* provides cloud-based data warehouse. Large-scale datasets are stored in HDFS and HBase.
- Computing model for meeting the capability of data intensive computing and nearly linear speedup of ETL, DM, SNA , SA and TM algorithms based on MapReduce and BSP (Bulk synchronous parallel)[6]. According to the characteristics of the algorithms, they are implemented in different computing model. *BC-PDM* also provides brilliant visualization for the presentation of results.
- Novel methods are proposed to assist operator marketing in real-world application for telecom industry. Moreover, *BC-PDM* has established Business Application Templates (BATs), such as statistical template, business recommended model, customer churning, community detection, etc.

This demo focuses on three tasks: business recommendation, customer community detection and user preference classification.

This demo introduces the following points. Section 2 shows the architecture and the core algorithms of *BC-PDM*. Section 3 experiments the performance and scalability of *BC-PDM* using large-scale datasets. Section 4 presents how the system solves the business tasks in telecommunication using novel data mining, social network analysis and text mining techniques.

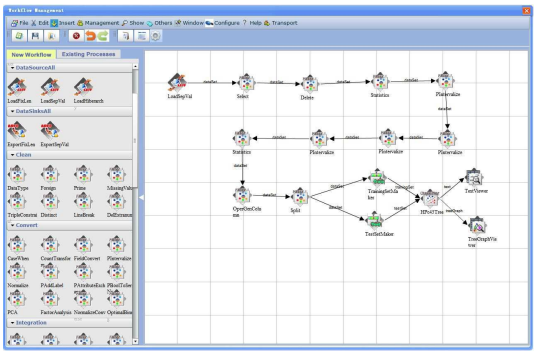

Figure 1: User Interface of *BC-PDM*.

2. SYSTEM OVERVIEW

The overview architecture of *BC-PDM* is presented in Fig.2. The system consists of three main layers. The function and feature of each layer is described as followed:

- *Cloud Computing Layer* provides storage and computational support. This layer uses HDFS, HBase and Hive as distributed file system.
- *Core Algorithms Layer* is the key of *BC-PDM* which includs ETL, DM, SA, SNA and TM algorithms.

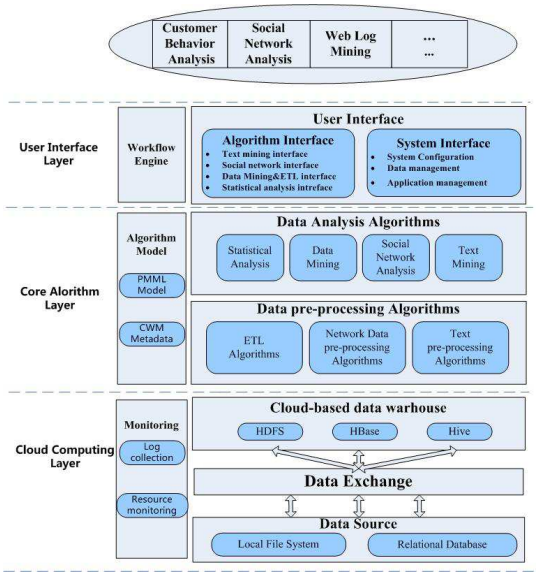

Figure 2: The Architecture of *BC-PDM*.

- *User Interface Layer* provides a Flex-based user interface. The functions of data storage and analysis are published as web service

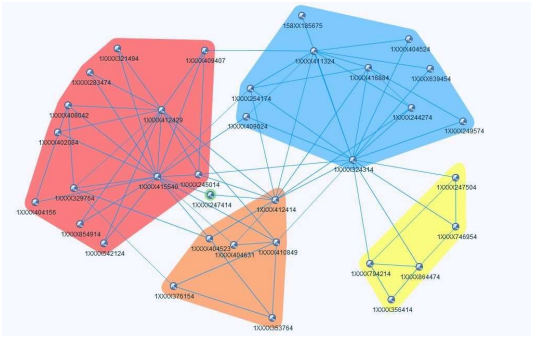

Figure 3: The community structure in a small mobile communication network.

Fig.3 shows the community structure of mobile communication network. Social network analysis provides many basic graph metrics including: degree, cluster coefficient, PageRank, network density, shortest distance and network diameter. Community detection and Community evolution analysis are also comprised in *BC-PDM*. The SNA has the following features:

- *Parallel Computing Framework: BC-PDM* employs MapReduce and BSP to implement the CPM (Clique percolation method)[7] which is an overlapping community detection method. The experiment indicates that BSP is more suitable for iterative graph algorithms than MapReduce.
- *High Applicability:* When analyzing small-scale data, serial computing could achieve good performance because the initiate time of Hadoop cannot neglect. *BC-PDM* offers serial algorithms in SNA as an alterna-

tive choice to guarantee the high practical applicability. System can select the most appropriate computing model for users according to the data size. Experiments have found the suitable inflexion point of the data size.

- *Message Passing Mechanism:* In order to improve the parallelism efficiency of graph algorithm and enable SNA to be flexible enough to express graph algorithms, *BC-PDM* build a multi-source message passing mechanism in *BC-PDM*. The message's state is modified to control the execution process of algorithm makes full use of every parallel computational job and avoid great amount of iteration.
- *Balanced Data Distribution: BC-PDM* proposes a novel high efficient parallel multi-level stepwise partitioning algorithm to reduce the communication and speed up the system.

Text mining refers to the process of deriving high-quality information from text. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. Typical text mining tasks include text categorization, text clustering, entity extraction, production of granular taxonomies, sentiment analysis, document summarization and entity relation modeling. *BC-PDM* provides text preprocessing algorithms, Bayes and Center Vector classification algorithms. Meanwhile, it also provides clustering algorithms such as Partitioning Methods and Canopy.

BC-PDM implements many clustering, classification and association rule algorithms. Moreover, *BC-PDM* creates and compares a set of different models based on the specified options. Then it ranks the best candidates according to the according to users' criteria. So users can choose the best approach for a given analysis. The data mining model are stored as PMML (Predictive Model Markup Language) and users can use PMML to migrate data mining model.

Parallel ETL can extract, transform and load variety kinds of data. The large-scale data are stored in HBase and HDFS. Corresponding metadata are kept at metadata repository. Technology metadata of each ETL component are conducted in accordance with the CWM (Common Warehouse Metamodel).

3. PERFORMANCE

The Chinese Software Testing Center (CSTC) has tested the *BC-PDM* in function, reliability, usability, efficiency, maintainability and portability. The evaluation has been performed on the 10 nodes Hadoop clusters, and each node consists of 4 Intel Xeon E5504 CPU, 8 GB main memory and 1024 GB hard drive. The evaluation data set is 403G.

Fig.4 shows the performance of typical algorithms. The average response time of 100 concurrent users accessing systems is 3.32 seconds. This experiment compares the performance of Groupby and MissValue component in different scale datasets. MissValue component is used to process the numerical incomplete problems. The result shows that the execution time presents linear speedup. Also,the social network algorithms have excellent performance. PageRank algorithm needs 17 minutes and 2 seconds to process 200,000 nodes and 800,000 edges, while Community Detection algorithm needs 1 hour 10 minutes and 17 seconds.

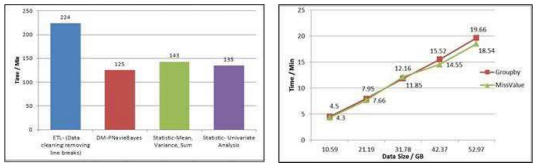

Figure 4: The performance of the DM and ETL algorithms.

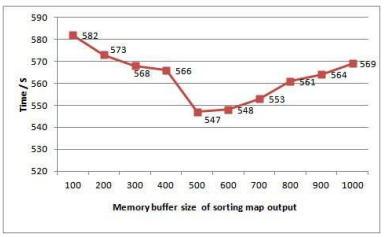

Figure 5: The effect of parameters.

Hadoop has many configuration parameters. Adjust parameters can take different effect. *BC-PDM* focuses on how to set the optimum parameters to improve system performance.The experiment results as follows (Cluster with 8 nodes, data sets 8GB):

- Memory buffer size: The left part of Fig.5 shows the execution time varying with memory buffer size. With the increase of memory buffer size, the execution time increases first and then decreases.
- Maximum of merge streams: The maximum of merge streams is the maximum number of streams to merge at once when sorting files. Fig.6 shows the execution time varying with maximum of merge streams. The execution time decreases with the increase of the maximum of merge streams and tend to certain stable value.

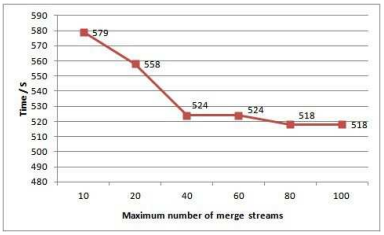

Figure 6: The effect of parameters.

4. APPLICATION SCENARIOS

BC-PDM analyzes user influence, user behavior, user preferences, network quality and the possibility of customer churn, such as telecommunications business or Internet search department.

4.1 Business recommendation

With rapidly growing of telecommunication users, users' consumption behaviors present some specific patterns. Discovering these patterns is conducive to business promotion.

Traditional analysis algorithms meet new challenges. *BC-PDM* uses the CDR (Calling Detail Records) and user information to establish the recommended model. First, *BC-PDM* selects long-distance calls, local calls, short message service and other fields from the CDRs as features. *BC-PDM* divides users into six categories using cluster algorithm. *BC-PDM* uses uses' category as the classification attribute and users' basic information to build the recommendation model. The experimental results show that the accuracy reaches 89.03%. *BC-PDM* provides business recommendation to 100 million users in 2 hours using the above model.

4.2 User preference classification

Reading on mobile phone is becoming part of everyday life. *BC-PDM* gets the hobby of user from the log which records the surfing of a specific user during a period of time based on the idea of the action of user. Browsing log records consists of much information such as IP address, Timestamp, encoding, page title, URL, etc. Through analyzing users' accessing page contents, *BC-PDM* can find users' preferences. Then, user preferences classification is converted into web page classification problems.

First, *BC-PDM* collects tens of thousands of pages as the original corpus for training classification model. The well-trained text categorization model has more than 30 categories and 73 sub-categories. Second, system uses the RMBoost classification algorithm which is proposed by *BC-PDM* and the accuracy rate can reach 90%. RMBoost is a centroid-based algorithm by calculating the centroids of different categories. The core of RMBoost algorithm is the centroid adjusting strategy. Centroid is the identifier of every category. RMBoost pushes the loose centroids farther while brings the compact centroid closer.

Meanwhile, it requires several iterations to make sure that the relative margin is greater than a given margin. Finally, the type of pages which users access for the most times and have the longest on-line time is the users' preference.

4.3 Overlapping customer community detection

A large account of complex systems can be modeled using complex networks. Mobile social networks consist of mobile users who communicate with one another using mobile phones. The structure of customer communication network provides us a way to understand customers' relationships. Mobile communication network is comparatively more common than other communications in everyday lives. Community structure is an important feature of mobile community detection. Generally, a community is a sub-graph whose nodes are densely connected within it but sparsely connected with the rest of the network. *BC-PDM* can also detect hierarchical community from mobile communication network.

This system employs CPM algorithm to detect hierarchical community. By combining with parallel computing framework, system has realized the parallel edition of CPM. Fig.7 shows the community structure in telecommunication network.

5. CONCLUSION AND FEATURE WORK

This demo introduces *BC-PDM* based on cloud computing. It has the ability to analyze the huge scale telecom data with data mining, social network analysis, text mining and statistical analysis and the high performance with linear speedup ratio. The *BC-PDM* shows its effectiveness on real world mobile communication data.

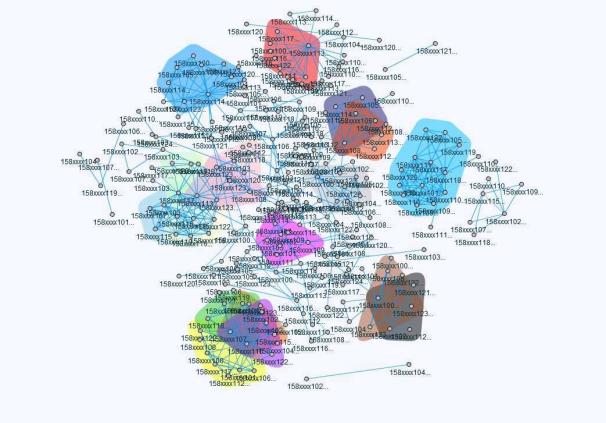

Figure 7: The result of community detection.

6. ACKNOWLEDGMENT

We would like to thank Dr. Deng Chao from China Mobile Research Institute for providing strong support on the project. This work is supported by the National Natural Science Foundation of China (Grant No.60905025, 90924029, 61074128). And the authors would like to acknowledge the anonymous reviewers.

7. REFERENCES

[1] M. Xu, D. Gao, C. Deng, Z. Luo, and S. Sun. Cloud computing boosts business intelligence of telecommunication industry. *Cloud Computing of Lecture Notes in Computer Science*.

[2] A. Ghoting, P. Kambadur, E. Pednault, and R. Kannan. Nimble: a toolkit for the implementation of parallel data mining and machine learning algorithms on mapreduce. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 334–342. ACM, 2011.

[3] S. Owen, R. Anil, T. Dunning, and E. Friedman. *Mahout in action*. Manning Publications Co., 2011.

[4] Z. Prekopcsk, G. Makrai, T. Henk, and C. Gspr-Papanek. Radoop: Analyzing big data with rapidminer and hadoop. In *Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM 2011)*, 2011.

[5] U. Kang, C. Tsourakakis, and C. Faloutsos. Pegasus: A peta-scale graph mining system implementation and observations. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 229–238. IEEE, 2009.

[6] L. Valiant. A bridging model for parallel computation. *Communications of the ACM*, 33(8):103–111, 1990.

[7] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.