Send Orders for Reprints to reprints@benthamscience.ae

Open Access

Research on Improved Distributed Association Rules Mining Algorithm in Hadoop Cloud Platform

Fenghua Liu*

Network Information Center, Shandong Women's University, Jinan, Shandong, 250300, China

Abstract: In this paper, data mining of association rules, data mining of association rules on distributed databases and distributed encrpytion techniques are introduced. Secondly, existing algorithms of data mining of association rules on distributed databases are analyzed in detail, and then they are improved on aspects of efficiency and security, whereafter the algorithm of EP_ DMA is proposed, later some examples are given and then the merit of EP_DMA can be seen. Finally, a frame of system of distributed mining of Association Rules based on EP_DMA is proposed.

Keywords: Distributed Association Rules, Cloud, Hadoop Platform, Mining Algorithm.

1. INTRODUCTION

Nowadays in the big data era, people are overwhelmed with the sheer volume of information. According to the statistics released by the authority in 2011, the total amount of global data every two years will be doubled, and it is expected that the amount of data human beings have in 2020 will reach a staggering 35 trillions GB. Facing with such vast oceans in data, how to extract valuable information using data mining techniques highlights strong vitality. Through data mining, users can extract potential useful information, rules, high level information from huge, random, vague and noisy data sets, thus it can be of great importance to the field of scientific studies, business decisions, etc. [1].

While big data brings great opportunities, it also presents challenges for effective data management and utilization. Cloud, the Emergence of New Forms of improves. Cloud computing is service based and can provide computing pattern with dynamic scalable virtualized resource. It is also able to efficiently seek out useful information in the large amount of data, thus making plenty of new applications flourish in the cloud environment. Utilizing its advantages in distributed processing and virtualization, this paper conducts a study in the following three aspects [2].

Data mining has been researched and applied widely in the near few years, and data mining of association rules which is highly demanded in the area of commerce decision-making is one of the most important and fundamental problems in this area. Now most of big corporations have many branches that are self-governed, and people have paid more and more attention to network security, so existing sequential algorithms can't content demand [3]. In this paper, encryption techniques are utilized and an algorithm of privacypreserving distributed data mining of association rules is proposed, which is named EP_ DMA, finally a frame of the system of distributed mining of association rules based on EP_ DMA is proposed.

2. RELATED WORKS

Data mining is an important area in KDD, and mining association rules in large databases is a critical aspect of data mining researches. The rapid development of Internet or Intranet makes a great progress in database applications. Since the security and cost of communication and efficiency of the applications, collecting and integrating a large amount of data from Internet/Intranet sites are not practical ways [4]. The problem of mining association rules in distributed databases arises from this situation.

This dissertation proposes the C-DMA (center-distributed mining association rules) algorithm in star structure, and a method of mining multiple layers association rules in distributed databases, a method of mining multiple layers association rules using meta-learning and adjustable method in distributed databases, based on analyses and introduction of the basic concepts and algorithms of mining association rules and mining association rules in distributed databases. After analyzing the quantitative association rules and interestingness of association rules which are encountered often in distributed association rule mining, the dissertation proposes the methods of changing the quantitative attributions into bool attributions using FCM and Gene algorithm [5].

Step 1:

The FDM and CD are main stream algorithms for mining association rules in distributed databases. These two algorithms all work on net structure networks. However, in practical applications, considering cost in constructing the networks or management in networks, users prefer staring networks to net networks which do not meet their requirements. The dissertation proposes the C-DMA algorithm to solve this problem, based on FDM and CD. Experimental results show

1874-4443/15 2015 Bentham Open

that the performance of the C-DMA is available and extendable.

book	I0	I1	I2	I3	I4
customer	(DBS)	(DS)	(C-programming)	(Computer Network)	(OS)
T1	1	1	1	1	0
T2	0	1	0	1	1
T3	0	1	1	1	0

Step 2:

In the process of mining association rules, the quantitative attributes exit in databases. How to handle these attributes affects the mining results and the efficiency. The dissertation proposes the methods of changing the quantitative attributes into bool attributes, so that many algorithms can be used, based on enhanced FCM and the genetic method.

book	I0	I1	I2	I3	I4
customer	(DBS)	(DS)	(C-programming)	(Computer Network)	(OS)
T4	1	1	1	1	1
T5	1	0	1	0	1
T6	0	1	1	0	1

Step 3:

In practical applications, multiple layers concept association rules mining are often encountered. The dissertation proposes the multiple layers concept rules mining algorithm in distributed databases, based on designing and analyzing the algorithm of mining association rules in single database.

book	I0	I1	I2	I3	I4
customer	(DBS)	(DS)	(C-programming)	(Computer Network)	(OS)
T7	1	0	1	1	0
T8	1	1	1	1	0
T9	1	1	0	1	0

It is important to enhance the efficiency in mining association rules in distributed databases. The dissertation proposes adjustable meta-learning algorithm in mining association rules in distributed databases, based on the Sampling algorithm.

How to evaluating the association rules mined from large databases is very critical in applications. The dissertation proposes a method to processes the association rules mined, which combines the Klementtinen theory and similarity theory based on analyzing methods about the interestingness of association rules as shown in Fig. (**1**).

Since the concept of association rules is proposed by Agrawal in 1993, research of mining association rules has been one of the most active aspects of defaming area. Presently, research of mining association rules in centralized system has been well done and the relative theory is becoming perfect. But mining association rules in distributed system is a topic that has just been proposed and the relative theory is not as so much. With the development of Internet and the distributed-database, a great deal of data is stored in the distributed nodes of the web and it is impossible to be stored in one single node on account of communication-efficiency and security, which makes it important to find algorithm of mining association rules in distributed system and makes the meaning of our research in the thesis [6] (Fig. **2**).

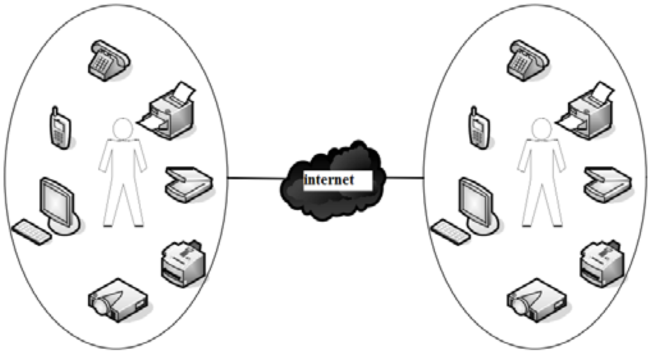


Fig. (2). Cloud terminal.

In this paper, an algorithm AprTidRec based on Apriori is proposed. Based on the algorithm of mining association rules in distributed system-CD and DD, two ways of mining association rules together with their architectures in distributed system are provided. In the end of this paper, the system of mining association rules in distributed system is implemented. The system based on C/S(client to server, Fig. (**3**)) mode is composed of local and global modules. After run of the local module, we get association rules based on local database and we get association rules based on global database after run of the global module [7].

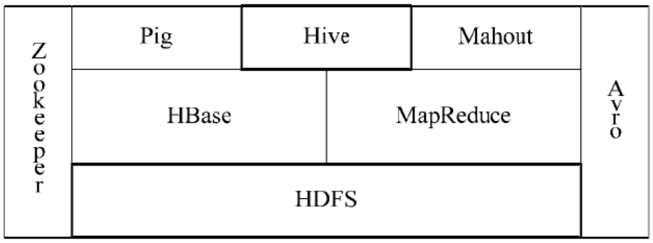


Fig. (3). The relationship between hadoop major projects.

To verify the validity of the algorithm, we test the system of mining association rules by means of experimental design.

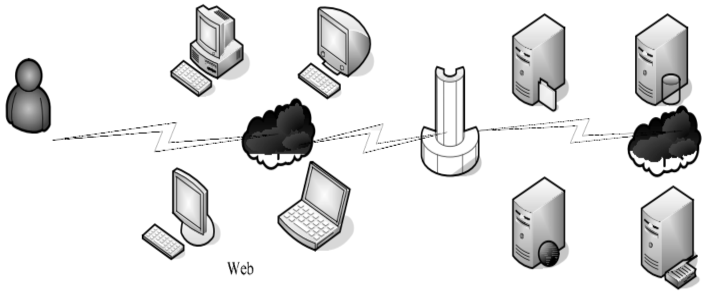


Fig. (1). Cloud computing service mode.

The results of the experiment indicate a perfect application and a high efficiency of the algorithm.

Association rule mining is an important sub-branch of data mining, which mines interesting association or correlation relationships among a large set of data items. Association rules are considered interesting if they satisfy both a minimum support threshold and minimum confidence threshold (Fig. **4**). Association rule mining has become a hot research topic in recent years, and it has been used widely in selective marketing, decision analysis and business management. Association rule mining algorithms are core contents in the area, and there are several famous typical algorithms [8]. This dissertation does some research on these algorithms, proposes a new algorithm and applies it to distributed data mining.

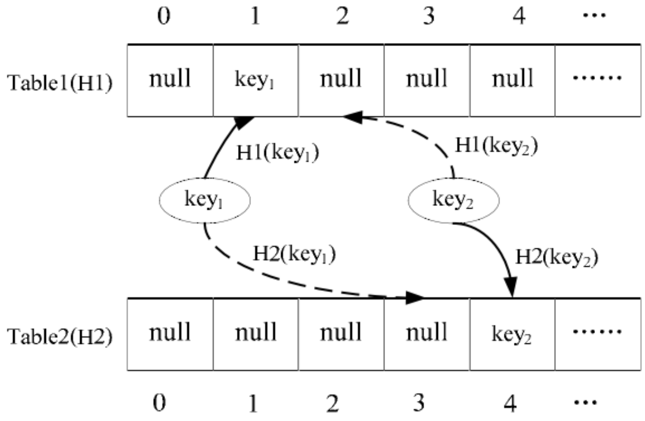


Fig. (4). Cuckoo hash table.

The algorithm of association rules is an important research direction in the field of data mining with its simple, efficient solution and wide application has become one of the hot spots of research in the field of data mining. Distributed association rule mining has broad application background; many practical problems can be solved using the data mining algorithm. This paper studied and analyzed the association rules algorithm and distributed association rules algorithm, focuses on the data mining algorithm in distributed system. This paper uses the mature Hadoop distributed platform presents two improved algorithms of distributed association rules algorithm. In view of the existing shortcomings of Map Reduce distributed association rules algorithm presents a global pruning strategies and frequent matrix storage strategy. (Fig. **5**) With the frequent matrix storage and pruning has certain advantages in efficiency.

First of all, based on association rules algorithm development and research status of the comprehensive understanding, according to the existing association rule cannot deal with large-scale data and distributed data, the paper presents the realization of rule and knowledge discovery process using Hadoop platform. Application of Map Reduce computing model can effectively solve the problem of data block; can solve many computers cooperative processing of massive data problem. It makes the problem difficult to solve simple as shown in Figs. (**6** and **7**).

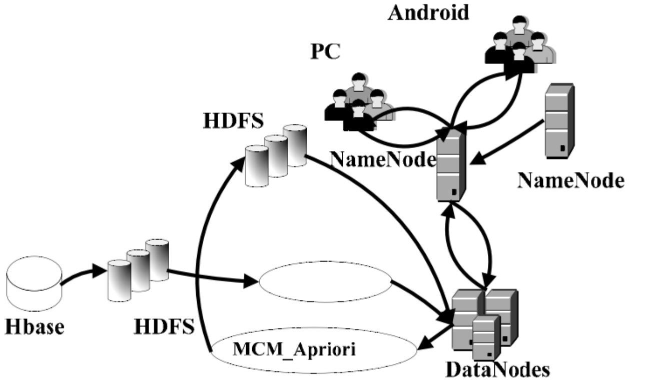


Fig. (6). The frame of association mining section.

Secondly, in view of the current distributed association rule algorithm, MPAOR algorithm is proposed for large-scale data distributed processing, the algorithm adds a global pruning techniques in the realization of the existing MPApriori, which makes the calculation of frequent item sets count again reduced, at the same time frequent matrix storage is applied to distributed association rule algorithm, calculation method is put forward by frequent matrix storage Map Reduce computing model. Experiments show that the proposed algorithm improves the efficiency of the algorithm, and save

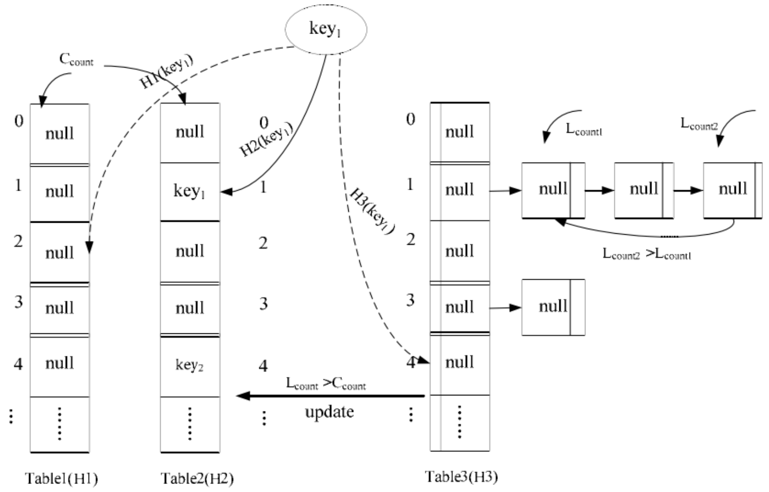


Fig. (5). The improved hadoop algorithm.

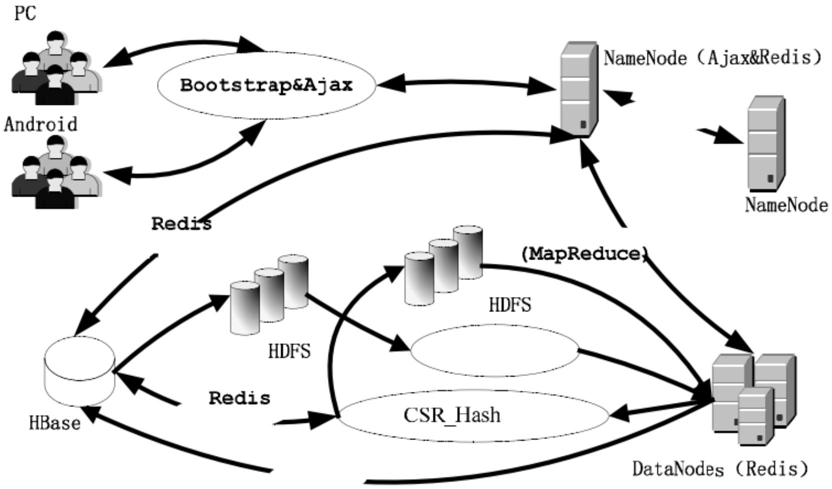
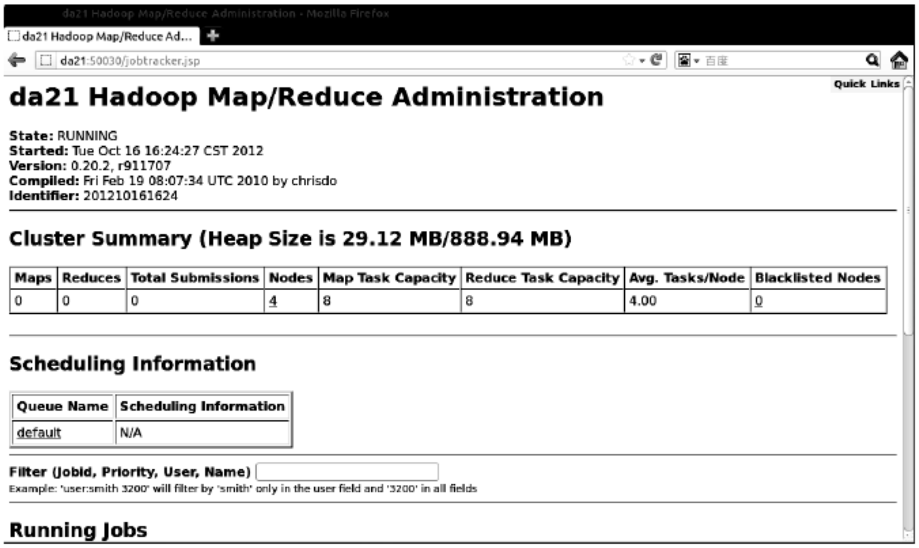


Fig. (7). The frame of searching section.


da21 Hadoop Map/Reduce Administration - Mozilla Firefox
da21 Hadoop Map/Reduce Ad...
da21:50030/jobtracker.jsp
百度
Quick Links

da21 Hadoop Map/Reduce Administration

State: RUNNING
Started: Tue Oct 16 16:24:27 CST 2012
Version: 0.20.2, r911707
Compiled: Fri Feb 19 08:07:34 UTC 2010 by chrisdo
Identifier: 201210161624

Cluster Summary (Heap Size is 29.12 MB/888.94 MB)

Maps	Reduces	Total Submissions	Nodes	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node	Blacklisted Nodes
0	0	0	4	8	8	4.00	0

Scheduling Information

Queue Name	Scheduling Information
default	N/A

Filter (Jobid, Priority, User, Name)
Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields

Running Jobs

Fig. (8). Running status of job tracker.

the memory space. It is very important for large-scale data processing. Simulation results verify the effectiveness of the algorithm.

With the rapid development and popularization of the computer technology, communication technology and network technology, a large number of databases are widely used in various fields of social life. The amount of data accumulation is easy to achieve terabytes, even pet bytes. These data are often noisy, large, heterogeneous and complex, so it is difficult to use them directly. So, how to dig up the valuable knowledge from the huge amounts of data more rapidly, low costly and more efficiently, and help policymakers to make better decisions has become a new topic in the field of data mining technology. The emergence of cloud computing brings new solutions for massive data mining. Hadoop, developed by the Apache foundation, is an open source implementation of cloud computing technology, and its core technology is the Hadoop distributed file system HDFS and parallel programming framework MapReduce. On the basis of in-depth study of traditional data mining algorithms, it is hotspot in the field of data mining to how to use the improvement of traditional data mining algorithm by combining the traditional data mining algorithms with the parallel programming framework MapReduce to deal with huge amounts of data mining as shown in Fig. (**8**).

This thesis researches the cloud computing, the Hadoop distributed file system HDFS and parallel programming framework MapReduce in detail, and expounds the technical architecture of data mining system based on Hadoop. Then, with the further research of traditional association rule mining algorithm Apriori, the thesis gives a parallel processing strategy of Apriori algorithm and puts forward an improved parallel algorithm AprioriMR. Then, with the basis of previous research and the introduction of the concept of power set and matrix, the thesis proposes two improved association rules mining algorithm respectively, which are AprioriPMR based on Hadoop and power set, and AprioriMMR based on Hadoop and matrix. Finally, the thesis sets up the experimental environment combined with Hadoop and HBase, and completes the writing of the improved algorithms with Java, and then uses the different experimental data sets and experimental conditions to test the validity of the improved algorithm. Through the comparative analysis of experimental results, it is concluded that the improved algorithms have higher efficiency in performance.

3. DATA MINING ALGORITHM

Data mining is an important area in KDD, and mining association rules in large databases applies more widely than other methods. Existing algorithms and modules cater to a centralized environment, such as database or data warehouse. With the development of distributed database and network technology, collecting and integrating a large amount of data from Internet sites are not practical ways. To solve the problem, this dissertation researches the mining association rules in distributed databases.

The first aspect deals with the defects in the traditional Apriori algorithm of association rules. Based on the column-oriented database called HBase, this paper presents a novel distributed algorithm of association rules mining, (MCM-Apriori), which associates the Map/Reduce programing model with coding operation. This can quickly find out accurate relations among knowledge models. Further, the two times of Map/Reduce processes greatly reduce the running time of MCM-Apriori, making it accurate and efficient.

Additionally, facing with new requirements for big data management, the paper puts forward a fast lookup algorithm of mix hash. It is based on engine in key-value in-memory database Redis and technology of Cuckoo hash. By building up pubfic an overflow area and using the method of shift keying, the query respond time is reduced and searching efficiency is improved.

Finally, an online bookstore sales system has been designed and implemented under the Hadoop framework of cloud computing. Using the improved MCM-Apriori algorithm and the fast lookup algorithm CSR-Hash, it parses and recommends book data in real-time and high-efficiency. This achieves fast query and analysis, and data-storage reliability, and shows great advantages of NoSQL database combining with Map/Reduce in real-time high-concurrency.

3.1. Purposes

The rapid development of computer and Internet Technology and the maturity of the widely Application of web 2.0 Data has exploded. The traditional data mining algorithms are inefficient when dealing with huge amounts of data, the emergence of cloud computing bringing a new way for its improvement. Through the power of dusting, cloud computing realizes reliable storage and high-speed computing for massive data. Hadoop as a mature open source cloud computing framework, with its highly efficient, scalable, low-cost advantages has been widely used in data mining related areas. On the basis of Laurels and typical Data Collection Systems, and selects Apriori algorithm which is in the algorithm modulo of the new data mining system and used widely to improve, willing enhance its efficiency when dealing massive data.

3.2. Methods

The reach methods used in this paper include: documentary research, structured approach, case study method and comparative analysis. Documentary research can understand the current situation of related research and provide a theoretical reference for this paper' research. Structured approach is a commonly used method for system analysis, which is being of guiding significance to analysis the system architecture of cloud data mining which is based on hadoop. This paper describes the implementation process of traditional Apriori algorithm and the feasibility of the improvement algorithm through instance. Through comparative analysis, this paper analyses the advantages of improvement algorithm. We can see Eq. (1) and (2)

$$\beta_j \approx \alpha_j \times |D| / 16MB \tag{1}$$

$$W_j = \lambda_j^1 \times W_j^{CPU} + \lambda_j^2 \times W_j^{RAM} + \lambda_j^3 \times W_j^{IO} \tag{2}$$

```
Key codes are as follows:
Reduce (itemset, list(sup))
// list(sup): list of support counts
 int result=0;
for each sup in list:
 result+=ParseInt(sup);
Emit (itemset, result);
```

3.3. Results

The previous distributed algorithms communicate over loading and need much more database scanning. For solving those problems, we propose four original association rules mining algorithms which are PDDM, GDS, DFP and MGMF algorithms. The PDDM algorithm improves the expansibility and communication of the previous algorithms effectively with less communication. The GDS and DFP algorithms reduce the scanning database I/O time relative to Apriori algorithm and reduce the communication relative to others distributed algorithm like FDM. The algorithm for mining global maximum frequent item sets (MGMF) is different from other maximum frequent item sets mining algorithms which can conveniently get all global maximum frequent item sets using FP-tree structure by one time mining, and superset checking is very simple and speedy. The MGMF algorithm is more effective than previous maximum frequent item sets mining algorithms, and can mine all maximum frequent item sets thought only two times database scanning as shown in Fig. (**9**).

(1) A system of Data Collection, the combination of proposals based on the data of the introduction of Laurels Laurels, modules of the system structure and functions of debate.

(2) On the basis of this algorithm, the model in the neck of the programming of a large amount of Data Processing, we propose to use the database of the Division of the Best idea. Decipher and details of Design and improve through the Union demonstrates the Improvement and Analysis of Feasibility.

(3) Through Case Studies, the Improvement of Energy Efficiency of the algorithms reduce the complexity of time and Space.

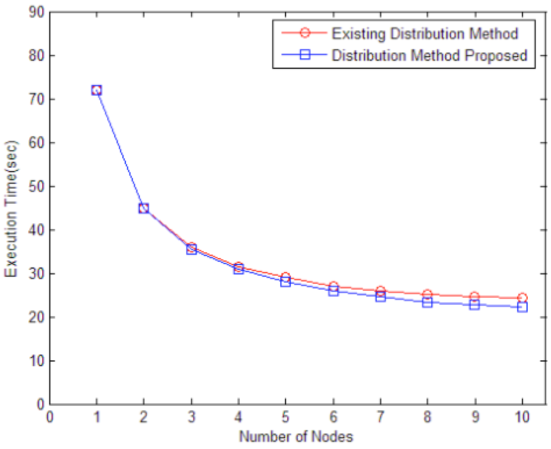


Fig. (9). Experimental results.

4. EXPERIMENTS

In our experiment, we use the Hadoop version of 0.20.0 9 machine running on the cluster (1 weeks, 8 slaves). All machines have a single processor (running on 2.60gh) and 4GB memory. Changes in experimental IBM data generator, transaction form in the database, we use speedup, evaluation the superiority of our algorithm criterion. All nodes, calculation and transmission, and the acceleration of the parallel execution, all data nodes under the condition of accelerated quantitative data under specific conditions increased, the MapReduce can be parallel to the rule mining algorithm Apriori algorithm is of high efficiency of proof. Mainly determines the performance of parallel algorithm of data set. If we prove that small data set on the experiment because its performance is low; the total running times of the communication time additional a higher proportion. It is very easy in the experiment. More information of our predicted nodes not significant proportion of communication time, the opposite effect, here see big data good speedup characteristics. Parallel algorithm is scalable, can accelerate remained almost large data and the data node size to increase the display image. The communication cost of the CD algorithm. The algorithm is O (C, n) each stage - C + N candidate item set and data set the size of a number, respectively. Results are shown in Fig. (**10** and **11**).

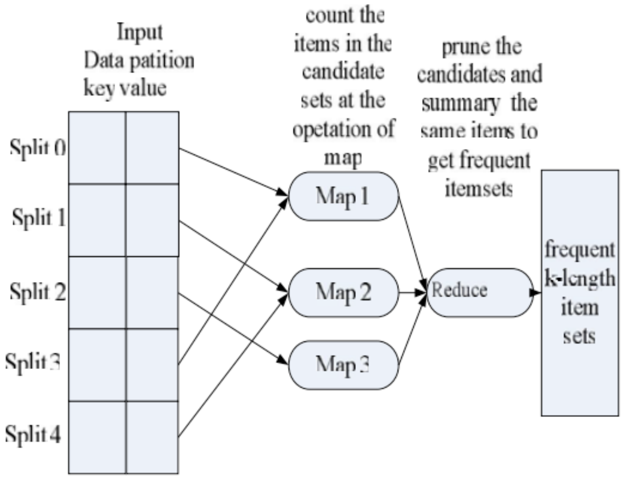


Fig. (10). Dataflow of parallel apriori.

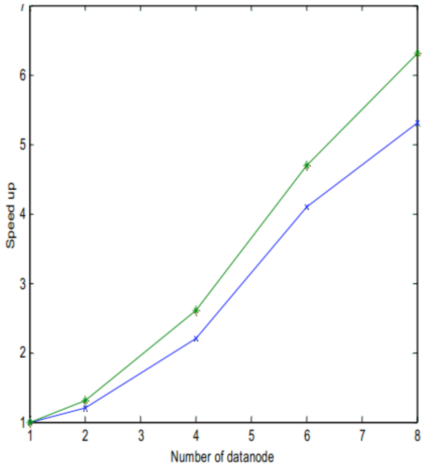


Fig. (11). Apriori not (: 100000: years,, to) = 1%).

CONCLUSIONS

First, this paper analyses and introduces the basic concepts and algorithms of mining association rules and mining association rules in distributed databases. And argues about relation among three kinds of different frequent item sets, proposes mining only the set of maximal frequent item sets instead of every frequent item sets. To make an better improvement, an experiment is performed upon existing algorithms of distributed association rules mining and obtains improving strategy and solution. An efficient distributed algorithm of association rules mining based on constrained sub tree is proposed. The algorithm for mining global maximum frequent item sets is deferent from other algorithms which can conveniently get all global maximum frequent item sets using FP-reed structure by one time mining, and superset checking is very speedy. And can mine all maximum frequent item sets thought only two times database scanning, then a method of adding prior weight among every sites is adopted to obtain easily global maximum frequent item sets. Finally, improved algorithm is applied to mine the data about teaching and scientific research of universities, with the purpose of finding out the potential rules in teaching and scientific research to offer some help to teaching activity and scientific research in the following year.

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] M. Riondato, J. A. DeBrabant, R. Fonseca, and E. Upfal, "PARMA:a parallel randomized algorithm for approximate association rules mining in MapReduce," In:*Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (CIKM), pp. 85-94, 2012.

[2] L. J. Li, and M. Zhang, "The strategy of mining association rule based on cloud computing," In*: Proceedings of the International Conference on Business Computing and Global Informatization*, pp. 475-478, 2011.

[3] L. Yang, and Z. Z. Shi, "An efficient data mining framework on Hadoop using java persistence API," In*: Proceedings of the IEEE* 10th *International Conference on Computer and Information Technology* (CIT), pp. 203-209, 2010.

[4] Z. Wu, J. Cao, and C. Fang, "Data cloud for distributed data mining *via* pipelined MapReduce," In*: Proceedings of the* 7th *International Workshop on Agents and Data Mining Interation* (ADMI), pp. 316-330, 2011.

[5] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "Toolkit-based high-performance data mining of large data on MapReduce-clusters". *In: Proceedings of the IEEE International Conference on Data Mining Workshops* (ICDMW), pp. 296-301, 2009.

[6] C. Khancome, V. Boonjing, and P. Chanvarasuth, "A two-hashing table multiple string pattern matching algorithm," In*: Tenth International Conference on Information Technology: New Generations* (ITNG), pp. 696-701, 2013.

[7] F. Zhao, and Q. Liu, "A string matching algorithm based on efficient hash function," In*: Proceedings of the International Conference on Information Engineering and Computer Science* (ICIECS), pp. 1-4, 2009.

[8] J. Zhang, G. Wu, X. Hu, and X. Wu, "A Distributed Cache for Hadoop Distributed File System in Real-time Cloud Services". In*: Proceedings of the* 2012 *ACM/IEEE 13th International Conference on Grid Computing* (GRID 2012), pp. 12-21, 2012.

Received: January 06, 2015 Revised: April 04, 2015 Accepted: May 15, 2015

© Fenghua Liu; Licensee *Bentham Open.*

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/-licenses/by-nc/3.0/) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.