ISSN 22773061

Design and Development of Apriori Algorithm for Sequential to Concurrent Mining Using MPI

Urmila R. Pol

Department Of Computer Science, Shivaji University Kolhapur

E-Mail: urmilec@gmail.com

ABSTRACT:

Owing to the conception of big data and massive data processing there are increasing owes related to the temporal aspects of the data processing. In order to address these issues a continuous progression in data collection, storage technologies, designing and implementing large-scale parallel algorithm for Data mining is seen to be emerging in a rapid pace. In this regards, the Apriori algorithms have a great impact for finding frequent item sets using candidate generation. This paper presents highlights on parallel algorithm for mining association rules using MPI for passing message base in the Master-Slave based structural model.

KEYWORDS: - Apriori algorithms; Data Mining; Frequent item sets

Academic Discipline And Sub-Disciplines

Computer Science

SUBJECT CLASSIFICATION

Data Mining: Parallel Computing

TYPE (METHOD/APPROACH)

Research Paper

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol 10, No.7

editor@cirworld.com
www.cirworld.com, member.cirworld.com

1. INTRODUCTION

Data mining could be characterized as the methodology of discovering hidden pattern in database. The main objective of the data mining is to manipulate the data into knowledge. Association rule mining is a sort of data mining process [1]. Association rule mining is done to extract interesting correlations, patterns, associations among items in the transaction database or other data repositories [2]. Association rules are widely utilized in various areas such as telecommunication networks, marketing, risk management, inventory control etc. Data Mining directly arranged to the enormous databases which have hundreds of properties and a huge number of records that contain complex relationship between the data sheets, and this will inevitably lead to the dramatic increase of the search space and size in the process of data mining [3]. It is obviously able to improve efficiency when using parallel data mining. Hence, that has become an imperative problem for design the parallel algorithms of association rules for the efficient mining when using the high-performance parallel workstations. In data mining, Apriori is a classic algorithm for studying association rules. Apriori is intended to operate on databases containing transactions for example, collections of items bought by customers, or details of a website frequentation [4].

Association rules mining is a vital zone of research for data mining. A number of potential and interesting relationships will be found in the large amount of data through mining some potential relationship between the item sets of the database [5]. These relationships play an important role in guiding and reference for the market basket analysis, cross-selling of commodities, business decision-making such as advertising mail analysis [6].

In this paper, in order to achieve high-performance parallel computing, there is an algorithm which using Master-Slave structure and communicate by MPI between the hosts, make full use of the resources of the workstation, a unified scheduling, coordination of treatment, under the cluster environment.

2. ASSOCIATION RULE

The association rules problem is as follows. Let I = $\{i1, i2 \ldots. in\}$ be a set of n binary attributes called items. Let D = { t1,t2,….,tm} be a set of transactions called database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$

2.1 Measures Association Rule

Basically, association mining is about discovering a set of rules that is shared among a large percentage of the data [7]. Association rules mining tend to produce a large number of rules. The objective is to find the rules that are suitable to Users. There are two methods of measuring usefulness, being objectively and subjectively. Objective measures involve statistical analysis of the data, such as support and confidence. [8]

Table 1 Example of Support Measure [8]

TID	Items	Support=Occurance/Total Support

1	ABC	Total Support = 5
		Support[AB] = 2/5 = 40%
		Support[BC] = 3/5 = 60%
		Support[ABC] = 1/5 =20%

2	ABD	
3	BC	
4	AC	
5	BCD	

The rule $X \Rightarrow Y$ holds with confidence c if c% of the transaction in D that contain X also contain Y. Rules that have a c grater than a user specified confidence is said to have minimum confidence. [8]

Table 2 Example of Confidence Measure [8]

TID	Items	Given $X \Rightarrow Y$ Confidence=Occurance[Y]/ Occurance[X]
1	ABC	Total Support = 5 **Confidence** [A impliesB] = 2/3 = 60% **Confidence** [B implies C] = 3/4 = 75% **Confidence** [AB implies C] = 1/2 =50%
2	ABD	
3	BC	
4	AC	
5	BCD	

3. IDENTIFYING LARGE KEYWORD SETS

There are numerous algorithms of association rules are available, but out of which Apriori classic algorithm is are available for data mining. The algorithm is used on I / 0 for a lot of time because of the need to repeatedly scan the database and produce a large number of frequent itemsets, therefore, it will resulting very low efficiency for data mining. [9]

Therefore, in order to improve the efficiency of Apriori algorithm, Apriori algorithm have been improved on a large number of extent, would like to find an efficient and reliable algorithm for mining frequent itemsets, but most of them are just confined to the optimize and improvement of the serially algorithm [6]. Optimization of the serial algorithm, to a certain extent, improve efficiency, but it is on a single computer run, in theory, N workstations running Apriori algorithm, then the efficiency will be enhanced, therefore, the algorithm of the association rules mining is attempting to parallel. [10]

Table 3: Notation for sequential algorithm [11]

K– Keyword set	A keyword set of K keywords
Lk	Set of large K keyword sets (those with minimum support) Each member of this set has two fields 1) Keyword Set. 2) Support count.
Ck	Set of candidate K keyword sets (potentially large keyword set) each member of this set has two field 1) Keyword set 2) Support count.

```
Apriori Algorithm:
Pass 1
   1.Generate the candidate itemsets in C1
   2.Save the frequent itemsets in L1
Pass k
 1. Generate the candidate itemsets in Ck
   from the frequent  itemsets in Lk-1
      a.  Join Lk-1 p with Lk-1q, as follows:
           insert into Ck
          select p.item1, p.item2, . . . , p.itemk-1,
          q.itemk-1 from Lk-1 p, Lk-1q
          where p.item1 = q.item1, . . . p.itemk-2
          = q.itemk-2, p.itemk-1 < q.itemk-1

      b.  Generate all (k-1)-subsets from the
           candidate itemsets in Ck

      c.  Prune all candidate itemsets from Ck
          where some (k-1)-subset of the
          candidate itemset is not in the frequent
```

itemset Lk-1

2. **Scan the transaction database to determine the support for each candidate itemset in Ck**

3. **Save the frequent itemsets in Lk**

The key to the parallel association rules is dealing with good communication between the processor and load balancing, so in this paper, there is an algorithm for parallel association rules mining base on MPI [12]. In algorithm design, In order to achieve higher efficiency of load balancing, MPI is used to uniform average distribution of resources; Using centralized architecture: a processor as a control processor dedicated to generate the overall itemsets for frequently, and is responsible for exchange information with other processing, and other processors as a workstation processor, only responsible for generating the local candidates and pruning set and count, there are not existing the exchange of information between the workstations, with the goal that we can reduce the communication time and enhance efficiency. [13]

4. PARALLEL APRIORI ALGORITHM

The algorithm assumes shared-nothing architecture where each of processor has private memory and a private disk. The processors are connected by a communication network and can communicate only by passing messages [14]. The communication primitives used by our algorithms are part of the MPI (Message Passing Interface) communication library supported on the IBM-SP and are keywords set for a message passing communication standard currently under discussion. Data is equally distributed on the disks attached to the processors [11, 14]. Each processor's disk has roughly an equal number of transactions. We do not require transactions to be placed on the disks in any unique way. We can accomplish the parallelism of Apriori Algorithm in different ways; at instructional level or at data level or control level [15]. We are following data level parallelism. Using given database generates the dominant group and also divides the database into N partitions. Each partition will be assigned to a processor. Data level parallelism of Apriori algorithm addresses the problem of finding all frequent keyword sets and the generation of the rules form frequent keyword set [16]. Refer to table 4 for a summary of notations used in the algorithm description. We are using superscripts to indicate processor id or rank and subscripts to indicate the pass number (also the size of keyword set).

Table 4 Notation used in parallel algorithm [11]

K keyword set	**A keyword set having k keyword set**
Pi	Processor with id or rank I
Di	The keyword set local to processor
N	Number of processor
Lk	Set of frequent k keyword set (those with minimum support) each member of this set has two fields 1) keyword set 2) support count.
Ck	Set of candidate k keyword set (potentially frequent keyword set) each member of this set has two fields 1) keyword set 2) support count.

Our proposed data level parallelism approach used irredundant computations in parallel. We have avoided the communication between the child or slave processors.

1. Select one processor to be the master, the other N-1 processors are slaves
2. Master processor devides data equally into n-1 processors.
3. Each processor Pi receives a 1/N part of the database from the parent or master processors, 0<i<N.
4. Processor Pi performs a pass over data partition Di and develops local support count for candidates in Ck.
5. Each processor Pi now computes Lk from Ck.
6. Each processor Pi sends its own local frequent itemsets to the master processor.
7. The master processor gathers the summary to generate global frequent itemsets.
8. The master processor partition the frequent itemsets, send to the local processor together with the global frequent itemsets. Cycle repeat until you find the most frequent itemsets. The master processor finally combine the output of nodes to generate set of global most frequent itemsets overall, delete the redundant information according to the credibility .

5. CONCLUSION

In this paper, I have recommended parallel Apriori Algorithm for Mining of Association Rules. In this paper, Ie have contended that to make data mining practical for ordinary people, data mining algorithms have to be efficient and data mining programs should not require dedicated hardware to run. On these fronts, we can conclude from this that Parallelization is a viable solution to efficient data mining.

REFERENCES

1. Chen, Ming-Syan, Jiawei Han, and Philip S. Yu. "Data mining: an overview from a database perspective." *Knowledge and data Engineering, IEEE Transactions on* 8.6 (1996): 866-883.
2. Zhao, Qiankun, and Sourav S. Bhowmick. "Association rule mining: A survey."*Nanyang Technological University, Singapore* (2003).
3. Kotsiantis, Sotiris, and Dimitris Kanellopoulos. "Association rules mining: A recent overview." *GESTS International Transactions on Computer Science and Engineering* 32.1 (2006): 71-82.
4. **Apriori** algorithm, Retrieved From: http://en.wikipedia.org/wiki/Apriori_algorithm, August 2, 2013.
5. Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.
6. Xue, Xing, Chen Yao, and Wang Yan-en. "Study on Mining Theories of Association Rules and Its Application." Innovative Computing & Communication, 2010 Intl Conf on and Information Technology & Ocean Engineering, 2010 Asia-Pacific Conf on (CICC-ITOE). IEEE, 2010.
7. Sharma, Anubha. "A Survey of Association Rule Mining Using Genetic Algorithm." *IJCAIT* 1.2 (2012): 5-11.
8. Lai, Kenneth, and Narciso Cerpa. "Support vs Confidence in Association Rule Algorithms."
9. Han, Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation." *ACM SIGMOD Record*. Vol. 29. No. 2. ACM, 2000.
10. Censor, Yair. *Parallel optimization: Theory, algorithms, and applications*. Oxford University Press, 1997.
11. Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*. Vol. 1215. 1994.
12. Li, Jianwei, et al. "Parallel data mining algorithms for association rules and clustering." *International Conference on Management of Data*. 2008.
13. Jinguang Sun and Weihao Pan "Parallel association rule mining for image data", Proc. SPIE 7490, PIAGENG 2009: Intelligent Information, Control, and Communication Technology for Agricultural Engineering, 74902A (July 10, 2009); doi:10.1117/12.836853.
14. Agrawal, Rakesh, and John C. Shafer. "Parallel mining of association rules."*Knowledge and Data Engineering, IEEE Transactions on* 8.6 (1996): 962-969.
15. P. Jagadeesh Babu, S. S. V. Apparao, Association Rule Mining using Count Circulation, International Journal of Computer Science and technology, Vol. 3, Issue 2, April - June 2012, 1031-1033
16. Puttegowda, D., Rajesh Shukala, and N. A. Deepak. "Performance Evaluation of Sequential and Parallel Mining of Association Rules using Apriori Algorithms." *Int. J. Advanced Networking and Applications* 2.1 (2010): 458-463.

Author' biography

Mrs. Urmila R. Pol was born in 1969, India. She did her Bachelors in Statistics and Masters in Compueter Application. She is Assistant Professor in Computer Science Department ,Shivaji University Kolhapur . She has also qualified the State Eligibility Test for Lectureship (SET) . She has awarded Ph.D. in Dec. 2010 in the subject of Computer Application and Management of Shivaji University of Kolhapur. She has to her credit 7 research papers in reputed national and international journals. She has completed one minor research project. Her current research interests are Parallel Computing, Cloud computing, Mobile application development .