International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-3, February 2013

A Survey on Parallel Partition Prime Multiple Algorithm

Anjali Singla, Jagpreet Kaur

Abstract- One of the important problems in data mining research is discovering Association rules from databases of transactions, where each transaction contains a set of items. In this dissertation work and improved approach proposed for parallel association rule mining. I proposed a new parallel partition prime multiple algorithms for association rule mining. Partition prime multiple algorithm addresses the shortcoming of previously proposed parallel buddy prime algorithm. New efficient algorithm proposed for load balancing. The proposed algorithm for parallel frequent item set mining and load balancing reduces the time and data complexity and divide transactional database efficiently for good load balancing among the processor.

Keywords- Association Rule, Load Balancing Algorithm

I. INTRODUCTION

Data mining [Chen 1996] [1][11] is the process of extracting interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large information repositories such as: relational database, data warehouses, XML repository, etc. Also data mining is known as one of the core processes of Knowledge Discovery Database (KDD).

Many people take data mining as a synonym for another popular term, Knowledge Discovery in Database (KDD). Alternatively other people treat Data Mining as the core process of KDD. Usually there are three processes. One is called preprocessing, which is executed before data mining techniques are applied to the right data. The pre-processing includes data cleaning, integration, selection and transformation.

The main process of KDD is the data mining process, in this process different algorithm share applied to produce hidden knowledge. After that the next process is post processing, this evaluates the mining result the knowledge can be presented if the result is satisfactory, otherwise we have to run some or all of those processes again until we get the satisfactory result. The actually processes work as follows. First we need to clean and integrate the databases. Since the data source may come from different databases, which may have some inconsistencies and duplications, we must clean the data source by removing those noises or make some compromises.

Manuscript published on 28 February 2013.
***** Correspondence Author (s)
Anjali Singla, Computer Science, Lovely Professional University, Phagwara, Punjab, India
Jagpreet Kaur, Computer Science, Lovely Professional University, Phagwara, Punjab, India

II. APPLICATIONS

We have been collecting a myriad of data, from simple numerical measurements and text documents, to more complex information such as spatial data, multimedia channels, and hypertext documents. Here is a non-exclusive list of a variety of information collected in digital form in databases and in flat files [12]

- Scientific data: Whether in a Swiss nuclear accelerator laboratory counting particles, in the Canadian forest studying readings from a grizzly bear radio collar, on a South Pole iceberg gathering data about oceanic activity, or American university investigating human psychology, our society is amassing colossal amounts of scientific data that need to be analyzed. Unfortunately, we can capture and store more new data faster than we can analyze the old data already accumulated.
- Medical and personal data: From government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups. Governments, companies and organizations such as hospitals, are stockpiling very important quantities of personal data to help them manage human resources, better understand a market, or simply assist clientele. Regardless of the privacy issues this type of data often reveals, this information is collected, used and even shared.
- Surveillance video and pictures: With the amazing collapse of video camera prices, video cameras are becoming ubiquitous. Video tapes from surveillance cameras are usually recycled and thus the content is lost. However, there is a tendency today to store the tapes and even digitize them for future use and analysis.
- Satellite Sensing: There are a countless number of satellites around the globe: some are geo-stationary above a region, and some are orbiting around the Earth, but all are sending a non-stop stream of data to the surface. NASA, which controls a large number of satellites, receives more data every second than what all NASA researchers and engineers can cope with. Many satellite pictures and data are made public as soon as they are received in the hopes that other researchers can analyze them.
- Games: Our society is collecting a tremendous amount of data and statistics about games, players and athletes.

III.RELATED WORK

The current parallel and distributed algorithms are based on the serial algorithm Apriori; an excellent survey given in [Zaki1999] classifies the algorithms by load-balancing. Strategy, architecture and parallelism.

A Parallel Buddy Prima Algorithm

This representation uses Prime numbers to represent the items in the Transaction. Each item is assigned a unique Prime number. Each transaction is represented by the product of the corresponding prime numbers of individual items in the transaction. Since the product of the prime numbers is unique, modulo division of Transaction's prime product by the prime product of the item set can check the presence of item set in the transaction.

- If the remainder is zero, and then the item set is present in the Transaction.
- If the remainder is Non-zero, and then the item set is not present in the Transaction

By checking the present item set in transactions using the above method, Support count can be calculated very quickly. Each transaction in the database can be represented in a single number by using Prime representation. The weakness of this representation is that the product of the prime number is very large number for a transaction with more number of items. However the product of the primes can be stored in floating point representation.

IV.PROPOSED WORK

Parallel buddy prima [8] algorithm use buddy prima approach in which prime number assign to the items is in sequential order. But it will create problem if greater prime number assign items that repeating so many times in the transaction together. Along with that parallel buddy prima algorithm has one more demerit that multiplication are too long and does not able to use proper method for load balancing, which further reduces the performance of association rule mining algorithm. However proposed new efficient parallel approach overcome the shortcoming of parallel buddy prima algorithm for association rule mining, using buddy prima algorithm, which combines bottom up and top down approach. Proposed new approach reduces the size of transaction prime multiple efficiently. Items can be purchased in any order or any combination, however, Sequential allocation of prime number can create serious problem. In proposed methodology first divide the transactions equally between the processors by new load balancing method then less prime number assigned to those items having high frequency of transaction. As example, in Table 5.1 items from computer shop have been taken (as shown in below tables) and arbitrarily number 1, 2…..14 have been assigned. In Table 5.2 transaction ID is showing databases of transaction occur for selling of each item.

Item Name	Selling Frequency	Assign Number
Laptop	6	1
Motherboard	2	2
Hard disk	6	3
DVD	6	4
RAM	6	5
Keyboard	3	6
Pen drive	7	7
Processor	4	8
Mouse	4	9
Speaker	5	10
Printer	4	11
Cabinet	1	12
USB port	3	13
LAN card	3	14

Table 5.1 Items selling from computer shop

Transaction ID	Transactions
T1	1,3,7,13
T2	4,6,10,11
T3	3,9,13
T4	4,5,7,8,14
T5	1,2,3,7,9,13
T6	4,5,7,8,9,10,14
T7	1,2,4,5,6,10,11
T8	1,3,7,9,10
T9	1,3,7,11,13
T10	4,5,6,10,11
T11	1,3,4,5,7,8,14
T12	5,8,12

Table 5.2 transactions ID is showing databases of transactions occur for selling of each item

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Transaction ID	Transaction ID	Transaction multiple
T1	2*5*17*41	6970
T2	7*13*29*31	81809
T3	5*23*41	4715
T4	7*11*17*19*43	1069453
T5	2*5*17*23*41	160310
T6	7*11*17*19*23*29*31	713325151
T7	2*7*11*13*29*31	1799798
T8	2*5*17*23*29	113390
T9	2*5*17*31*41	216070
T10	7*11*13*29*31	899899
T11	2*5*7*11*17*19*43	10694530
T12	11*19	209

Table 5.3 representing the multiplication of prime number that was assigned previously.

Now suppose we want Item set {3, 7} in which transaction set then taking mode 5*17=85.

Transaction ID	Modulo Division	Remainder	Item's Presence
T1	6970mod85	0	Yes
T2	81809 mod85	Non Zero	No
T3	4715 mod85	Non Zero	No
T4	1069453 mod85	Non Zero	No
T5	160310 mod85	0	Yes
T6	713325151 mod85	Non Zero	No
T7	1799798 mod85	Non Zero	No
T8	113390 mod85	0	Yes
T9	216070 mod85	0	Yes
T10	899899 mod85	Non Zero	No
T11	10694530 mod85	0	Yes
T12	209 mod85	Non Zero	No

V. Weakness

The weakness of this representation is that the product of the prime number is a very large number for a transaction with more number of items. In this algorithm no proper load balancing method was developed. Hence we are proposing a new method to overcome demerits.

VI.New Proposed Algorithm

STEP 1: Find the infrequent item of length 1 and store in memory IF1 by putting constrained, minimum support count of 3.

STEP 2: Remove the Infrequent 1 item as denoted by IF1.

STEP 3: Now take the transaction count for each Ti

STEP 4: Find the size Maxlen of maximal size transaction in Database and stored value.

STEP 5: Divide the Transaction equally based on number of node for that we apply static load balancing.

STEP 6: Assign prime number Pi to each unique item Ti to each processor separately, for it Arrange the item in basis of frequency descending order and allot prime number which has higher frequency allot low prime number corresponding.

STEP 7: Represent each transaction Ti of m by the multiple Mi of all prime number representation Pi of the items in the transaction (P1* p2….*pm) and store corresponding processor memory.

STEP 8: find the support count of Item set S Mi mod k and counting its presence using remainder store it in support separately.

VII Conclusion and Future Works

Association rule mining [1][5] has a wide range of applicability such market basket analysis, medical diagnosis research, Website navigation analysis, homeland security and so on. Sequential association rule mining algorithm [6] [14] cannot provide scalability, in terms of the data dimension, size or runtime performance for large databases. Therefore high performance parallel and distributed computing [51] can be employed for enhancing the performance, scalability. Several algorithms [21] [52] have been proposed so for to mine all the frequent item set in a transaction database parallels. These algorithms differ from one another in the method of handling the candidate sets, parallel design space and the method of reducing the number of database passes. The main idea behind most of the algorithm is to divide transaction equally to each processor and then apply bottom up approach for generating frequent item set. If the maximum item set is longer, top down search is suitable. For transaction with a medium sized maximal frequent set, combination of both these approaches performs well.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Proposed a new parallel Partition Prime Multiple algorithm for association rule mining. The new algorithm integrates both the bottom up search and well as top down search. This algorithm is suitable for item set of any size. The parallel Partition Prime Multiple algorithm efficiently removes the shortcoming of previously proposed parallel buddy prima algorithm [8]. Prime number use to represent each transaction, it uses less memory as it stores only one number for each transaction. The computational complexity is reduced as the product of their equivalent prime number represents each candidate item set. A new algorithm reduce the size of transaction prime multiplies efficiently. The support count of any set is found without any additional scan of the database. The pruning of the infrequent item in first scan reduces the size of the dataset in the main memory. A new parallel Partition Prime Multiple algorithm can be improved to provide dynamic load balancing. With the increasing amount of data, it is important to develop more efficient algorithms to extract knowledge from the data. However, the volume of data size is increasing much faster than CPU execution speeds, which has a strong influence on the performance of software algorithms. The performance of Parallel Computing, however cannot improve linearly as the number of the parallel nodes grows. Some researchers have tired to use hardware devices to accomplish data mining tasks.

References:

1] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, San Francisco, USA, 2001, ISBN 1558604898.

[2] Arun K. Pujari." Data Mining Techniques", Universities Press (India) Private Limited.

[3] IBM Research Report "Parallel Mining of Association Rules: Design", Implementation and Experience and Data Eng., vol. 8, no. 6, pp. 962± 969, Dec. 1996.

[4] Mohammed J.Zaki. "Parallel & Distributed Association Mining: A Survey." IEEE Trans. Knowledge & Data Engg. 1999.

[5] Sotiris Kotsiantis, Dimitris Kanelloponles, "Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol 32(1), 2006.

[6] Yanbo Wang, "Categorization of Association Rule Mining Algorithms", Department of computer sciences, University of Liverpool, U.K.

[7] R. Agrawal and J.C. Shafer, ªParallel Mining of Association Rules,º IEEE Trans. Knowledge

[8] Dr.S.N.Sivanandam, Dr.S.Sumathi, Ms.T.Hamsapriya, Mr.K.Babu "Parallel Buddy Prima – A Hybrid Parallel Frequent item set mining algorithm for very large databases"P.S.G College of Technology Coimbatore, Tamilnadu, India.

[9] Pieter Adriaans and Dolf Zantinge,Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, Third Edition (Potomac, MD: Two Crows Corporation, 1999); Data Mining (New York: Addison Wesley, 1996).

[10] Jeffrey W. Seifert, Data Mining: An Overview: Updated December 16, 2004 Analyst information Science and Technology Policy Resources, Science, and Industry Division.

[11] Sourav S. Bhowmick Association Rule Mining: A Survey: Qiankun Zhao Nanyang Technological University, Singapore Nanyang Technological University, Singapore

[12] Osmar R. Zaïane, 1999 CMPUT690 Principles of Knowledge Discovery in Databases.

[13] Agrawal, R.; Imielinski, T.; and Swami, A. Database Mining: A performance Perspective IEEE Transactions on Knowledge and Data Engineering, Special issure on Learning and Discovery in Knowledge-Based Datadabases, 5(6):914-925, December 1993.

[14] M. Hegland, "Algorithms for Association Rules", Lecture Notes in Computer Science, Vol. 2600, Jan 2003.

[15] Brain, S.; Motwani, R.; and Silverstein, C. Beyond Market Baskets: Generalizing Association Rules to Correlations. In Proc. Of the 1997 SIGMOD Conf. on the Management of Data, pp. 265-276

[16] Michael Goebel "A survey of Data Mining and knowledge Discovery software Tools" Department of Computer Science University of Auckland Le Grunewald School of Computer Science University of Oklahoma.

[17] Data Mining Lecture Notes http:// ww.db.stanford.edu/~ullman/mining/ Data Mining Lecture Notes

[18] C. Győrödi, R. Győrödi. "Mining Association Rules in Large Databases". Proc. of Oradea EMES'02: 45-50, Oradea, Romania, 2002

[19] R. Győrödi, C. Győrödi. "Architectures of Data Mining Systems". Proc. Of Oradea EMES'02: 141-146, Oradea, Romania, 2002.

[20] Cornelia Győrödi, Robert Győrödi, prof. dr. Stefan Holban,"A Comparative Study of Association Rules Mining Algorithms"

[21] Eui-Hong (Sam) Han, George Karypis, Member, IEEE, and Vipin Kumar, Fellow, IEEE "Scalable Parallel Data Mining for Association Rules".

[22] C. Győrödi, R. Győrödi, S. Holban, M. Pater. "Mining Knowledge in Relational Databases". Proc. of CONTI 2002, 5th International Conference on Technical Informatics: 1-6, Timisoara, Romania, 2002.

AUTHOR'S PROFILE

Anjali Singla , computer science,Lovely Professional University,Phagwara,Punjab,India,(email: smily_anjali@yahoo.com)

Jagpreet Kaur, computer science,Lovely Professional University,Phagwara,Punjab,India,(email**:** jagpreet.16934@lpu.co.in)