A STUDY ON EFFECTIVE MINING OF ASSOCIATION RULES FROM HUGE DATABASES

V.Umarani,
Asst Professor, Dept of Computer Science
Sri Ramakrishna College of Arts and Science for Women
Coimbatore, India

Dr.M.Punithavalli,
Director and Head, Dept of Computer Science,
Sri Ramakrishna College of Arts and Science for Women,
Coimbatore, India.

Abstract— **Association rule discovery is one of the most important technique in the field of data mining. It aims at finding interesting patterns among the databases. However it becomes much tedious to mine the association rules as the data are growing more and more like mountain. Hence it is important in developing techniques in such a way that interesting rules are mined effectively from huge databases. This paper provides an overview of techniques that are used to improvise the efficiency of Association Rule Mining (ARM) from huge databases.**

Keywords: Association rules, sampling, apriori, fptree, partitioning Clustering.

I. INTRODUCTION

Owing to the current explosion of information and the accessibility of cheap storage, collecting enormous data has been achievable during the last decades. The ultimate intent of this massive data collection is the utilization of this information to achieve competitive benefits, by determining formerly unidentified patterns in data that can direct the process of decision making. Lately it has been demonstrated that analysis of data with the aid of Online Analytical Processing(OLAP) tools alone is highly tedious; illustrating the necessity of an automated process to ascertain interesting and concealed patterns in data. Data mining techniques have increasingly been studied especially in their applications in real-world databases [1], [2], [3]. A formal definition of data mining is given as "A process of non-trivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database" [4].

Data mining is a major step in the Knowledge Discovery in Databases (KDD) process, consisting of applying computational techniques that, under acceptable computational efficiency limitations produce a particular enumeration of patterns (or models) over the data [2]. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods [7].The progress in data mining research has made it possible to implement several data mining operations efficiently on large databases [4]. The 'mined' information is typically represented as a model of the semantic structure of the dataset, where the model may be used on new data for prediction or classification. Data mining techniques have been successfully applied in many different fields including marketing, manufacturing, process control, fraud detection, and network management [6].

In general, data mining tasks can be classified into two categories: Descriptive mining and Predictive mining. Descriptive mining is the process of drawing the essential characteristics or general properties of the data in the database. Clustering, Association and Sequential mining are some of the descriptive mining techniques. Predictive mining is the process of inferring patterns form data to make predictions. The predictive mining techniques involve tasks like Classification, Regression and Deviation detection [8].Mining frequent itemsets from transaction databases is a fundamental task for several forms of knowledge discovery such as association rules, sequential patterns, and classification [9]. One of the popular descriptive data mining techniques is Association Rule Mining (ARM) [11], owing to its extensive use in marketing and retail communities in addition to many other diverse fields. Since its introduction in 1993 by Agrawal *et al.* [18], the task of association rule mining has received a great deal of attention. Today the mining of such rules is still one of the most popular pattern-discovery methods in Knowledge Discovery in Databases [18]. Data mining is a major step in the Knowledge Discovery in Databases (KDD) process, consisting of applying computational techniques that, under acceptable computational efficiency limitations produce a particular enumeration of patterns (or models) over the data [2]. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can includes statistical models, mathematical algorithms, and machine learning methods [7].The progress in data mining research has made it possible to implement several data mining operations efficiently on large databases [4]. The 'mined' information is typically represented as a model of the semantic structure of the dataset, where the model may be used on new data for prediction or classification. Data mining techniques have been successfully applied in many different fields including marketing, manufacturing, process control, fraud detection, and network management [6].

II Association Rules

Mining association rules is particularly useful for discovering relationships among items from large databases [10]. A standard association rule is a rule of the form X→ Y which says that if X is true of an instance in a database, so is Y true of the same instance, with a certain level of significance as measured by two indicators, support and confidence. The goal of standard association rule mining is to output all rules whose support and confidence are respectively above some given support and coverage thresholds. These rules encapsulate the relational associations between selected attributes in the database, for instance, coke → potato chips: 0.02 support; 0.70 coverage denotes that in the database, 70% of the people who buy coke also buy potato chips, and these buyers constitute 2% of the database. This rule signifies a positive (directional) relationship between buyers of coke and potato chips [19]. The mining process of association rules can be divided into two steps.

1. **Frequent Itemset Generation:** generate all sets of items that have support greater than a certain threshold, called minsupport.
2. **Association Rule Generation:** from the frequent itemsets, generate all association rules that have confidence greater than a certain threshold called minconfidence [33]. Apriori is a renowned algorithm for association rule mining primarily because of its effectiveness in knowledge discovery [34]. However, there are two bottlenecks in the Apriori algorithm. One is the complex frequent itemset generation process that uses most of the time, space and memory. Another bottleneck is the multiple scan of the database [35].

III. Improving the efficiency of Association Rule Mining

A. Sampling

Sampling is a powerful data reduction technique that has been applied to a variety of data mining algorithms for reducing computational overhead. In the context of association rules, sampling can be utilized to gather quick preliminary rules. This may help the user to direct the data mining process by refining the criterion for "interesting" rules. Sampling can speed up the mining process by more than an order of magnitude by reducing I/O costs and drastically shrinking the number of transaction to be considered. The validity of the sample is determined by two characteristics the size of the sample and the quality of the sample. The quality, in the context of statistical sampling techniques, refers to whether the sample captures the characteristics of the database. The highest quality sample would be an exact miniature of the database; it would preserve the distributions of individual variables and the relationships among variables [20].The quality of the sample for association rule mining can be improved by considering factors like transaction length and transaction frequency [12].

A number of studies were conducted to propose efficient methods for mining association rules by reducing either the CPU computation time or the disk access overhead .Some studies considered the usage of sampling techniques for reducing the processing overhead [27, 28,30,31]. Most of the prior works on sampling have concentrated on speeding up the phase by running a frequent itemset mining algorithm only on a small sample of the database [14]. Chiefly, researchers have evaluated the viability of using sampling [6] to reduce the dataset size. While such methods have shown quite a lot of promise it has been observed by several researchers [13,14,20] that it is often very difficult to quantify, apriori, the quality of the results obtained for a given sample size [29], necessitating novel and more effective sampling-based association rule mining algorithms to foster better mining results.

B. Reducing the number of passes

The disadvantage of Apriori algorithm made the researchers to think about new techniques to mine frequent patterns. The 2 main negative sides are the possible need of generating a huge number of candidates if the number of frequent 1-itemsets is high or if the size of the frequent pattern is big, the database has to be scanned twice repeatedly to match the candidates and determine the support What if we find a way to mine the frequent patterns without candidate generation? This would be a big improvement over Apriori. That is what the frequent pattern growth (FP-growth) algorithm does [15].

Wang et al [16] presented PRICES, an efficient algorithm for mining association rules. Their approach reduces large itemset generation time, which dictates most of the time in generating candidates by scanning the database only once and using logical operations in the process.

Another algorithm [17] called Matrix Algorithm generates a matrix which entries 1 or 0 by passing over the cruel database only once, and then the frequent candidate sets are obtained from the resulting matrix. Association rules are then mined from the frequent candidate sets.

C. Hash-based itemset counting

A hash technique is very efficient in generating the candidate item sets, in particular for the large two-itemsets, thus greatly improving the performance bottleneck of the entire process.

Soo et al [21] proposed Direct Hashing and Pruning [DHP] algorithm, an effective hash based technique for mining the association rules. This algorithm employs effective pruning techniques to progressively reduce the transaction database size. DHP utilizes a hashing technique to filter the ineffective candidate frequent 2 itemsets. DHP also avoids database scans in some passes as to reduce the disk I/O cost involved.

Another novel hash-based approach [22] for mining frequent item-sets over data streams was developed by En et al. The algorithm compresses the information of all itemsets into a structure with a fixed hash-based technique. This approach skillfully summarizes the information of the whole data stream by using a hash table to estimate the support counts of the non-frequent itemsets, and keeps only the frequent itemsets for speeding up the mining process.

Another algorithm Inverted Hashing and Pruning (IHP) [39] proposed by John et al. It for is developed for mining association rules between words in text databases. The characteristics of text databases are quite different from those of retail transaction databases, and existing mining algorithms cannot handle text databases efficiently, because of the large number of itemsets (i.e., words) that need to be counted. Two well-known mining algorithms, the Apriori algorithm [*1*] and Direct Hashing and Pruning (DHP) algorithm [5], are evaluated in the context of mining text databases, and are compared with the proposed IHP algorithm. It has been shown that the IHP algorithm has better performance for large text databases.

D. Transaction Reduction

Transaction reduction is another way that helps in mining association rules effectively. It relies on a concept that a transaction that does not contain any frequent k-itemset is useless in subsequent scans.

AprioriTid algorithm [26] is another way of improving the performance of Association Rules. This algorithm is used to construct the frequent itemset. The main idea of all these algorithm is according to the theory that the subset of a frequent itemset is a frequent itemset and the superset of an infrequent itemset is an infrequent itemset. They scan the database repeatedly to mining the association rules. There is another feature for algorithm AprioriTID, the support of the candidate frequent itemsets are calculated only at the first time it scanned the database D and also generated candidate transaction database D’ which only includes the candidate frequent itemsets. Then the latter mining are based on the database D’, It reduce the time of I/O operation because D’ is smaller than D, so, it enhance the efficiency of the algorithm. Another approach called MTR-FMA (modified transaction reduction based frequent itemset mining algorithm) developed by Thevar et al [27] maintains its performance even at relative low supports.

E. Partitioning

Various approaches to generate large item sets have been proposed based on portioning of the set of transactions. In this case, D is divided into *p* partitions D^1, D2….D^p. Partitioning may improve the performance of finding large item sets in several ways. By using partitioning, parallel and/or distributed algorithms can be easily created, where each partition could be handled by a separate machine.

Cheung et al [28] presented an algorithm called FDM. FDM is a parallelization of Apriori to shared nothing machines, each with its own partition of the database. At every level and on each machine, the database scan is performed independently on the local partition. Distributed Pruning is then done.

FPM (Fast Parallel Mining) for Association rule mining has been proposed [32]. It adopts Count Distribution approach and has incorporated two powerful candidate pruning techniques. It has a simple communication scheme which performs only one round of message exchange in each iteration.

Parthasarathy et al [36] have presented an excellent survey on parallel association rule mining with shared memory architecture covering most of challenges and approaches adopted for parallel data mining.

F. Adding extra constraints

Another type of association rule mining involves in retrieving patterns by adding extra constraints on the structure of patterns. Techniques applicable to constraint-driven pattern discovery can be classified into the following groups:

- Post-processing(filtering out patterns that do not satisfy user-specified constraints after the actual discovery process;
- Pattern filtering (integration of pattern constraints into the actual mining process in order to generate only patterns satisfy pattern constraints.
- Dataset filtering(restricting the source data set to objects that can possibly contain patterns that satisfy pattern constraints)

Wojciechowski et al [23] proposed a constraint based algorithm that improves the efficiency of constraint based frequent pattern mining by using dataset filtering techniques. Dataset filtering conceptually transforms a given data mining task into an equivalent one operating on a smaller dataset.

Rapid Association Rule mining (RARM) [25] is another method that uses a tree structure to represent original database and avoids candidate generation process. Constraints were applied during the mining process to generate only those association rules that are interesting to the users which guarantees the improvement of the efficiency of the existing mining algorithm.

Tien et al [24] presented a category based algorithm as well as the associated algorithm for constraint rule mining based on Apriori. This approach reduces computational complexity of mining process by passing most of the subsets of final itemsets.

G. Association Rule Clustering System.

Association Rule Clustering is useful when the user desires to segment the data. Lent et al [38] proposed a Clustering Association rule in which they measure the quality of the segmentation generated by ARCS (Association Rule Clustering System) using the minimum description length (MDL) principle of encoding the clusters on several databases including noise and errors. Scale-up experiments show that ARCS, using the BitOp algorithm scales linearly with the amount of data.

Pi et al [41] proposed a new Fuzzy Clustering Algorithm on Association Rules for Knowledge Management. A fuzzy simulation degree and simulated matrix for association rules are put forward and a new algorithm based on dynamic tree is used for implementing the fuzzy clustering. The experimental results show that this algorithm clusters the Association rules efficiently.

Gupta et al [40] recently proposed a cluster based algorithm that uses a novel approach to the insignificant transactions dynamically. During a particular pass only those clusters that seem to be statistically useful are scanned and as a consequence all insignificant tuples are filtered dynamically. The results of the algorithm show that removing false frequent items and insignificant transactions dynamically improves the performance of association rule mining.

H. Advanced Association Rule Techniques

Some of the other advanced Association Rule techniques [42] includes the following:

a. Generalised Association Rules
b. Multiple-Level Association Rule.
c. Quantitative Association Rule.
d. Using Multiple Minimum Support
e. Correlation Rules.
f. Temporal Association Rule.

IV. Conclusion

Association rule mining is one of the most important procedures in data mining. Mining association rules is a prototypical problem as the data are being generated and stored every day in corporate computer database systems. To manage this knowledge, rules have to be pruned and grouped, so that only reasonable numbers of rules have to be inspected and analyzed. Thus an appropriate technique has to be employed to mine the association rule efficiently.

References

[1] J. P. Bigus., “Data Mining with Neural Networks”, McGraw-Hill, 1996

[2] T. M. Mitchell., “Machine Learning”, McGraw-Hill, 1997.

[3] Sousa, M.S. Mattoso, M.L.Q. Ebecken, N.F.F. "Data Mining on Parallel Database Systems" Proc. Int. Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA'98), Special Session on Parallel Data Warehousing, CSREA Press, Las Vegas, E.U.A., Pp.1147-1154, July 1998.

[4] Fayyad U, “Data Mining and Knowledge Discovery in Databases: Implications from scientific databases,” In Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, pp. 2-11, 1997.

[5] Tsau Young Lin, "Sampling in association rule mining", Conference on Data mining and knowledge discovery: Theory, Tools, and Technology VI, vol. 5433, pp.: 161-167, 2004.

[6] Klaus Julisch," Data Mining for Intrusion Detection -A Critical Review" in proc. of IBM Research on application of Data Mining in Computer security, Chapter 1 , 2002.

[7] Jeffrey W. Seifert, "Data Mining: An Overview", in proceedings of CRS Report for Congress, 2004.

[8] Coenen F, Leng P, Goulbourne, G., “Tree Structures for Mining Association Rules,” In Journal of Data Mining and Knowledge Discovery, Vol. 15, pp. 391-398, 2004.

[9] Marek Wojciechowski, Krzysztof Galecki, Krzysztof Gawronek: ‘Concurrent Processing of Frequent Itemset Queries Using FP-Growth Algorithm’, Proc. of the 1st ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD'05), Tallinn, Estonia, 2005.

[10] Yu-Chiang Li, Jieh-Shan Yeh, Chin-Chen Chang, “Efficient Algorithms for Mining Shared-Frequent Itemsets”, In Proceedings of the 11th World Congress of Intl. Fuzzy Systems Association, 2005.

[11] F. Bodon, “A Fast Apriori Implementation”, In B. Goethals and M. J. Zaki, editors, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Vol. 90 of CEUR Workshop Proceedings, 2003.

[12] Basel A. Mahafzah, Amer F. Al-Badarneh and Mohammed Z. Zakaria "A new sampling technique for association rule mining," in Journal Of Information Science, Vol.35, pp. 358-376, 2009.

[13] Venkatesan T. Chakaravarthy, Vinayaka Pandit and Yogish Sabharwal, "Analysis of sampling techniques for association rule mining," In Proceedings of the 12th International Conference on Database Theory, Vol. 361, pp. 276-283, 2009.

[14] Y. Zhao, C. Zhang and S. Zhang, “Efficient frequent itemsets mining by sampling,” Proceedings of the fourth International Conference on Active Media Technology (AMT), pp. 112-117, 2006.

[15] Han, j. and Pei, J. 2000. Mining frequent patterns by pattern-growth: methodology and implications. ACM SIGKDD Explorations Newsletter2, 2, 14-20.

[16] Wang, C., Tjortjis, C., Prices: An Efficient Algorithm for Mining Association Rules, Lecture Notes in Computer Science, Volume 2447, 2002. pp. 77-83.

[17] Yuan, Y., Huang, T., A Matrix Algorithm for Mining Association Rules, Lecture Notes in Computer Science, Volume 3664, Sep2005.pp 370-379.

[18] R.Agrawal, T.Imielinski, and A.Swami, “Mining association rules between sets of Items in large databases”, in proceedings of the ACM SIGMOD Int'l Conf. on Management of data, pp. 207-216, 1993.

[19] Choh Man Teng, "A Comparison of Standard and Interval Association Rules", In Proceedings of the Sixteenth International FLAIRS Conference, pp.: 371-375, 2003.

[20] Suzuki Kaoru, “Data Mining and the Case for Sampling,” SAS Institute Best Practices Paper, SAS Institute, 1998.

[21] Soo, J., Chen, M.S., and Yu, P.S., 1997, “Using a Hash-Based Method with Transaction Trimming and Database Scan Reduction for Mining Association Rules” IEEE Transactions On Knowledge and Data Engineering, Vol.No.5. pp. 813-825.

[22] En Tzu Wang and Arbee L.P. ChenData,“ A Novel Hash-based approach for mining frequent itemsets over data streams requiring less Memory space” Data Mining and Knowledge Discovery, Volume 19, Number 1, pp 132-172.

[23] Wojciechowski, M., Zakrzewiez, M., Dataset filtering Techniques in Constraint based Frequent pattern Mining, Lecture Notes in Computer Science, Volume 2447, 2002, pp77-83.

[24] Tien Dung Do, Siu Cheng Hui,Alvis Fong, Mining frequent itemsets with category Based Constraints. Lecture Notes in Computer Science, Volume 2843, 2003, pp226-234.

[25] Das, A., Ng, W.K., and Woon, Y, K. 2001. Rapid association rule mining. In the proceedings of the tenth international conference on Information and knowledge management.. ACM press, 474-481.

[26] Rakesh Agarwal, Ramakrishnan Srikant,” Fast Algorithms for Mining Association Rules” 20th Intl Conference on VLDB, Santigo, Chile, Set.1994.

[27] Thevar., R.E; Krishnamoorthy, R,” A new approach of modified transaction reduction algorithm for mining frequent itemset”, ICCIT 2008.11th conference on Computer and Information Technology.

[28] Cheung, D., Han, J.Ng, V., Fu, A and Fu, Y. (1996), “A fast distributed algorithm for mining association rules” in Proc of 1996 Int’l Conference on Parallel and Distributed Information Systems’. Miami Beach, Florida, pp.31-44.

[29] Parthasarathy, S., "Efficient progressive sampling for association rules", IEEE International Conference on Data Mining, pp.: 354- 361, 2002.

[30] V.Umarani and M.Punithavalli,” Developing a Novel and Effective Approach for Association Rule Mining Using Progressive Sampling” In the proc of 2nd Int’l Conference on Computer and Electrical Engineering (ICCEE 2009), vol.1, pp610-614.

[31] V.Umarani and M.Punithavalli,” On Developing an Effectual Progressive Sampling Based Approach for Association Rule Discovery” In the proc of 2nd IEEE Int’l Conference on Information and data Engineering (2nd IEEE ICIME 2010), Chengdu ,China April 2010.

[32] Cheung, D., Xaio, Y., Effect of data skewness in parallel mining of association rules, Lecture Notes in Computer Science, Volume 1394,Aug 1998,pages 48-60.

[33] Raymond Chi-Wing Wong, Ada Wai-Chee Fu, "Association Rule Mining and its Application to MPIS", 2003.

[34] Agrawal, R. and Srikant, R., Fast algorithms for mining association rules. In Proc.20th Int. Conf. Very Large Data Bases, 487-499, 1994.

[35] Sotiris Kotsiantis, Dimitris Kanellopoulos,” Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, Vol.32, No: 1, pp. 71-82, 2006.

[36] Parthasarathy, S., Zaki, M.J.J., Ogihara, M., Parallel data mining for association rules on shared-memory systems, Knowledge and Information Systems: An International Journal,3(1):1-29,February 2001.

[37] Basel A. Mahafzah, Amer F. Al-Badarneh and Mohammed Z. Zakaria "A new sampling technique for association rule mining," in Journal of Information Science, Vol. 35, pp. 358-376, 2009.

[38] B.Lent, A.Swami,J.Wisdom, “Clustering association rules”, In the proc of 13th Int’l Conference on Data Engineering,pp.220.

[39] John D. Holt and Soon M. Chung,” Mining of Association Rules in Text Databases Using Inverted Hashing and Pruning” Lecture Notes in Computer Science, 2000, Volume 1874/2000, 290-300.

[40] Rajendra K.Gupta and Dev Prakash Agarwal,”Improving the performance of Association Rule Mining Algorithms by Filtering Insignificant Transactions dynamically”, Asian Journal of Information Management, pp.7-17. 2009 Academic Journals Inc.

[41] Pi Dechang and Qin Xiaolin,” A New Fuzzy Clustering Algorithm on Association Rules for Knowledge Management”, Information Technology Journal. Pp. 119-124, 2008. Asian Network for Scientific Information.

[42] Margaret H.Dunham,”Data mining Introductory and Advanced Topics”, Pearson Education 2008.