Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Enhanced Scaling Apriori Algorithm to Minimize the number of candidate sets while Generating Association Rules

K.Shyam Prasad*
M.Tech 2nd year, Dept of CSE,
ASCET, GUDUR, India.
Shyampsd999@gmail.com

C. Rajendra
Professor & HOD ,Dept of CSE,
ASCET, GUDUR, India.
hod.cse@audisankara.com

Abstract— This research work proposes an improved Apriori algorithm to minimize the number of candidate sets while generating association rules by evaluating quantitative information associated with each item that occurs in a transaction, which was usually, discarded as traditional association rules focus just on qualitative correlations. The proposed approach reduces not only the number of item sets generated but also the overall execution time of the algorithm. Any valued attribute will be treated as quantitative and will be used to derive the quantitative association rules which usually increases the rules 'information content. Transaction reduction is achieved by discarding the transactions that does not contain any frequent item set in subsequent scans which in turn reduces overall execution time. Dynamic item set counting is done by adding new candidate item sets only when all of their subsets are estimated to be frequent. The frequent item ranges are the basis for generating higher order item ranges using Apriori algorithm. During each iteration of the algorithm, use the frequent sets from the previous iteration to generate the candidate sets and check whether their support is above the threshold. The set of candidate sets found is pruned by a strategy that discards sets which contain infrequent subsets. This work evaluates the scalability of the algorithm by considering transaction time, number of item sets used in the transaction and memory utilization. Quantitative association rules can be used in several domains where the traditional approach is employed. The unique requirement for such use is to have a semantic connection between the components of the item-value pairs..

Keywords— Include at least 5 keywords or phrases

I. INTRODUCTION

. Data mining, also known as knowledge discovery in databases, has been recognized as a new area for database research. The problem of discovering association rules was introduced in latter stages. Given a set of transactions, where each transaction is a set of items, an association rule is an expression of the form X + Y, where X and Y are sets of items. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints. Conceptually, this problem can be viewed as finding associations between the "l" values in a relational table where all the attributes are Boolean. The table has an attribute corresponding to each item and a record corresponding to each transaction. The value of an attribute for a given record is "1" if the item corresponding to the attribute is present in the transaction corresponding to the record, "0" else. Relational tables in most business and scientific domains have richer attribute types. Attributes can be quantitative (e.g. age, income) or categorical (e.g. zip code, make of car). Boolean attributes can be considered a special case of categorical attributes. This research work defines the problem of mining association rules over quantitative attribute in large relational tables and techniques for discovering such rules. This is referred as the Quantitative Association Rules problem. The problem of mining association rules in categorical data presented in customer transactions was introduced by Agrawal, Imielinski and Swami [3]. This research work provided basic idea to several investigation efforts resulting in descriptions of how to extend the original concepts and how to increase the performance of the related algorithms [10]. The original problem of mining association rules was formulated as how to find rules of the form set1 => set2. This rule is supposed to denote affinity or correlation among the two sets containing nominal or ordinal data items. More specifically, such an association rule should translate the following meaning: customers that buy the products in *set1* also buy the products in *set2*. Statistical basis is represented in the form of minimum support and confidence measures of these rules with respect to the set of customer transactions. The original problem as proposed by Agrawal [3] was extended in several directions such as adding or replacing the confidence and support by other measures, or filtering the rules during or after generation, or including quantitative attributes. Srikant and Agrawal [1] describe a new approach where quantitative data can be treated as categorical. This is very important since otherwise part of the customer transaction information is discarded. Whenever an extension is proposed it must be checked in terms of its performance. The algorithm efficiency is linked to the size of the database that is amenable to be treated. Therefore it is crucial to have efficient algorithms that enable us to examine and extract valuable decision-making information in the ever larger databases [8]. This work present an algorithm that can be used in the context of several of the extensions provided in the literature but at the same time preserves its performance. The approach in our algorithm is to explore multidimensional properties of the data (provided such properties are present), allowing to combine this additional information in a very efficient pruning phase. This results in a very flexible and efficient

algorithm that was used with success in several experiments using quantitative databases with performance measure done on the memory utilization during the transactional pruning of the record sets.

II. PREVIOUS RESEARCH

Various proposals for mining association rules from transaction data were presented on different context. Some of these proposals are constraint-based in the sense that all rules must fulfill a predefined set of conditions, such as support and confidence [12],[6]. The other proposal identifies just the most interesting rules (or optimal) in accordance to some interestingness metric, including confidence, support, gain, chi-squared value, gini, entropy gain, laplace, lift, and conviction [9]. However, the main goal common to all of these algorithms is to reduce the number of generated rules . The research work extend the first earlier techniques since it do not relax any set of conditions nor employ a interestingness criteria to sort the generated rules. In this context, many algorithms for efficient generation of frequent item sets have been proposed in the literature since the problem was first introduced[5],[24]. The Direct Hashing and Pruning (DHP) algorithm [16] uses a hash table in pass *k* to perform efficient pruning of (*k+1*)-item sets. The Partition algorithm minimizes Input / output (I/O) by scanning the database only twice. In the first pass it generates the set of all potentially frequent item sets, and in the second pass the support for all these is measured. The above algorithm are all specialized techniques which do not use any database operations. Algorithms using only general purpose Data Base Management System (DBMS) systems and relational algebra operations have also been proposed [7]. Few other works tried to solve this mining problem for quantitative attributes. The authors proposed an algorithm which is an adaptation of the Apriori algorithm for quantitative attributes[18]. It partitions each quantitative attribute into consecutive intervals using *equi-depth* bins. Then adjacent intervals may be combined to form new intervals in a controlled manner. From these intervals, *frequent item sets* (c.f. *large item sets* in Apriori Algorithm) will then be identified. Association rules will be generated accordingly. The problems with this approach is that the number of possible interval combinations grows exponentially as the number of quantitative attributes increases, so it is not easy to extend the algorithm to higher dimensional cases. Besides, the set of rules generated may consist of redundant rules for which they present a "greater-than-expected-value" interest measure to identify the interesting ones. Some other efforts exploited quantitative information present in transactions for generating association rules. The quantitative rules were generated by discrediting the occurrence values of an attribute in fixed-length intervals and applying the standard Apriori algorithm for generating association rules[13],[2]. However, although simple, the rules generated by this approach may not be intuitive, mainly when there are semantic intervals that do not match the partition employed.Other authors [4],[15],[14] proposed novel solutions that minimize this problem by considering the distance among item quantities for delimiting the intervals, that is, their "physical"' placement, but not the frequency of occurrence as a relevance metric.

III. PROPOSED WORK

The proposed work comprises of two phases. The first phase concerns about the quantitative association rule mining with the enhancement on Apriori algorithm. The second phase deals with the reduction of memory utilization during the pruning phase of the transactional execution. The algorithm for generating quantitative association rules starts by counting the item ranges in the database, in order to determine the frequent ones. These frequent item ranges are the basis for generating higher order item ranges using an algorithm similar to Apriori, taking into account the size of a transaction as the number of items that it comprises.

a) Define an item set 'm' as a set of items of size 'm'
b) Specify frequent (large) item sets by 'Fm'
c) Specify candidate item sets (possibly frequent) by 'Lm'.

A 'n' range set is a set of n- item ranges, and each m-item set has a n-range set that stores the quantitative rules of the item set. During each iteration of the algorithm, the system uses the frequent sets from the previous iteration to generate the candidate sets and check whether their support is above the threshold. The set of candidate sets found is pruned by a strategy that discards sets which contain infrequent subsets. The algorithm ends when there are no more candidates' sets to be verified. The enhancement of Apriori is done by increasing the efficiency of candidate pruning phase by reducing the number of candidates that are generated for further verification. The proposed algorithm use quantitative information to estimate more precisely the overlap in terms of transactions. The basic elements considered in the development of the algorithm are number of transactions, average size of transaction, average size of the maximal large item sets, number of items, and distribution of occurrences of large item sets. The second phase of this work claimed improvement over Apriori by considering memory consumption for data transaction. This part of the algorithm generate all candidates based on 2-frequent item sets on sorted database, and all frequent item sets that can no longer be supported by transactions that still have to be processed. Thus the new algorithm no longer has to maintain the covers of all past item sets into main memory. In this way, the proposed level-wise algorithm accesses a database less often than Apriori and requires less memory because of the utilization of additional upward closure properties.

IV. RULE MINING ON QUALITATIVE ASSOCIATION OF ITEM

IV.1. Algorithm - Quantitative Association rule mining

a. Find all frequent item sets (i.e., satisfy minimum support)
b. Generate strong association rules from the frequent item sets (each rule must satisfy minimum support and minimum confidence).

c. Identify the quantitative elements
d. Sorting the item sets based on the frequency and quantitative elements
e. Merge the more associated rules of item pairs
f. Discard the infrequent item value pairs
g. Iterate the steps c to f till the required mining results are achieved

Let I = { i1, i2 … im} be a set of items, and T a set of transactions, each a subset of I. An association rule is an implication of the form A=>B, where A and B are non-intersecting The support of A=>B is the percentage of the transactions that contain both A and B. The confidence of A=>B is the percentage of transactions containing A that also contain B (interpret as P(B|A)). The occurrence frequency of an item set is the number of transactions that contain the item set.

V. IMPLÉMENTATION OF QUANTITATIVE APRIORI

5.1 Pseudo Code

```
{ * Outer Sequential Loop *}
While() {
{* Reduction Loop*}
Foreach(element e ) {
( i, val) = process (e);
Reduc(i) = Reduc(i) op val;
}
```

The function op is an associative and commutative function. Thus, the iterations of the for each loop can be performed in any order. The data-structure Reduc is referred to as the reduction object. The attribute selection method, which selects the attribute based on the confidence and support value (Fig1). The main correctness challenge in parallelizing a attribute like this on a shared memory machine arises because of possible race conditions when multiple processors update the same element of the reduction object. The element of the reduction object that is updated in a loop iteration (i) is determined only as a result of the processing. In the Apriori association mining algorithm, the data item read needs to be matched against all candidates to determine the set of candidates whose counts will be incremented.

The major factors that make these loops challenging to execute efficiently and correctly are as follows:
It is not possible to statically partition the reduction object so that different processors update disjoint portions of the collection. Thus, race conditions must be avoided at runtime. The execution time of the function process can be a significant part of the execution time of an iteration of the loop. Thus, runtime pre processing or scheduling techniques cannot be applied.The updates to the reduction object are fine grained. The reduction object comprises a large number of elements that take only a few bytes, and the for each loop comprises a large number of iterations, each of which may take only a small number of cycles.

VI. EXPERIMENTAL RESULTS FROM QUANTITATIVE APRIORI

The experiment focused on evaluating all quantitative Apriori techniques. Since we were interested in seeing the best performance, we used banking data set samples. We used a 1 GB dataset. The total number of distinct items was 1000 and the average number of items in a transaction was 15. A confidence of 90% and support of 0.5 is used. Execution times using 1, 2, 3, and 4 threads are presented on the processor. With 1 thread, Apriori does not have any significant overheads as compared to the sequential version. Therefore, this version is used for reporting all speedups. Though the performance of quantitative Apriori is considerably lower than Apriori, they are promising for the cases when sufficient memory for supporting full replication may not be available. The second experiment demonstrates that each of these Apriori techniques can be the winner, depending upon the problem and the dataset. We use a dataset with 20 distinct items, where the average number of items per transaction is 6. The total size of the dataset is 500 MB and a confidence level of 90% is used. We consider four support levels, 10%, 5%, 3%, and 2%. The execution time efficiency is improved for the quantitative Apriori on frequent item set evaluation with the support count (fig.1)

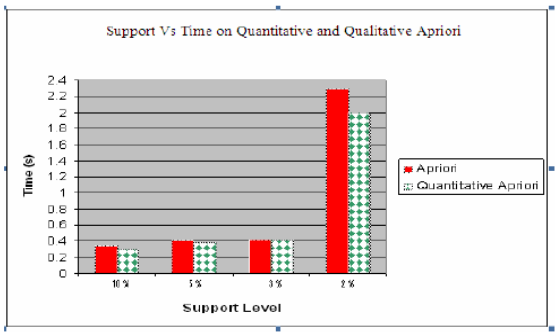

Figure 1: Support Vs Time on Quantitative and Qualitative Apriori

The thread execution on the quantitative Apriori and qualitative Apriori are evaluated for the same data set (Fig. 4). Here the initial thread requires more time, however consequent threads shows better scalable performance of quantitative Apriori.

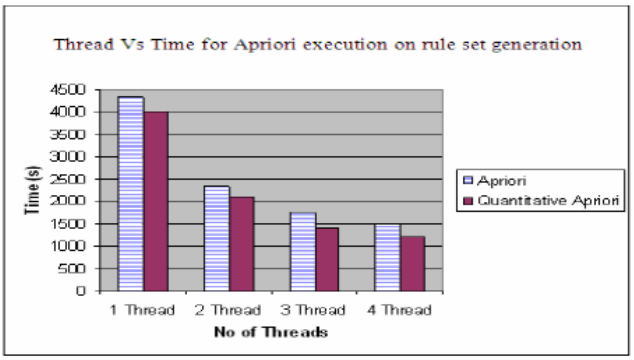

Figure 2: Thread Vs Time for Apriori execution on rule set generation

VII. Conclusion and Future Work

The research work has defined a new rule set namely the informative rule set that presents prediction sequences equal to those presented by the association rule set using the confidence priority. The informative rule set is significantly smaller than the association rule set, especially when the minimum support is small. The proposed work has characterized the relationships between the informative rule set and the non-redundant association rule set, and revealed that the informative rule set is a subset of the nonredundant association rule set. The work considers the upward closure properties of informative rule set for omission of uninformative association rules, and presented a direct algorithm to efficiently generate the informative rule set without generating all frequent item sets. The informative rule set generated in our work is significantly smaller than both the association rule set and the non-redundant association rule set for a given database that can be generated more efficiently than the association rule set. The efficiency improvement results from that the generation of the informative rule set needs fewer candidates and database accesses than that of the association rule set rather than large memory usage like some other algorithms. The number of database accesses of the proposed algorithm is significantly fewer than other direct methods for generating association rules on all items. So far we have identified that the performance of quantitative Apriori is considerably lower than Apriori. To improve the performance of quantitative Apriori the fuzzy logic will be applied in future.

REFERENCES

[1] Agarwal R. and V. Prasad, "A Tree Projection Algorithm for Generation of Frequent Itemsets," *Parallel and Distributed Computing*, 2000.

[2] Agrawal R, H. Mannila, R. Srikant, H. Toivonen, and A.Verkamo. "Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining, San Jose, CA*, pages 307-328, 1996.

[3] Agrawal R, T. Imielinski, and A. Swami. "Mining association rules between sets of items in large databases". *Proc.oOf the ACM SIGMOD Washington, D.C*, pages 207-216, May 1993.

[4] Alok Sharma, and Kuldip K. Paliwal, "Rotational Linear Discriminant Analysis Technique For Dimensionality Reduction*", IEEE Transactions on Knowledge and Data Engineering* Vol. 20,No. 10, October 2008.

[5] Anthony K.H. Tung, Hongjun Lu, Jiawei Han, Member, and Ling Feng, "Efficient Mining of Inter transaction Association Rules" *IEEE Transactions on Knowledge and Data Engineering,*Vol. 15, No.1, January / February'03.

[6] Bodon.F, " A Survey on Frequent Itemset Mining", Technical report, Budapest Univ. Of Technology and Economics,2006.

[7] Cheung D, V.T Ng, A. Fu, and Y.Fu. "Efficient mining of association rules in distributed databases". *IEEE Trans. Knowledge and Data Engineering*, pp 1-23, 1996.

[8] Elena Baralis, Tania Cerquitelli, and Silvia Chiusano, "IMine: Index Support for Item Set Mining", *IEEE Transactions on Knowledge and Data Engineering,* Vol. 21, No. 4, April 2009.

[9] Ghosh, A. and S. Dehuri, "Evolutionary algorithms for multi-criterion optimization: A survey" . *International Journal on Computers and Infromation Science,* 2: 38–57, 2004.

[10] Han J, Y Cai, and N Cercone, "Data Driven Discovery of Quantitative Rules in Relational Databases," *IEEE Trans Knowledge and Data Eng* ,Vol 5, pp 29 40,1993.

[11] Jen-Wei Huang, Chi-Yao Tseng, Jian-Chih Ou, and Ming-Syan Chen, "A General Model For Sequential Pattern Mining with a Progressive Database", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 9, September 2008.

[12] Ken Sun and Fengshan Bai "Mining Weighted Association Rules without Pre assigned Weights ". *IEEE Transactions on Knowledge and Data Engineering*, Vol 20, No 4, April 2008.

[13] Kurland.O and L.Lee, " Respect My Authority! HITS with out Hyperlinks, Utilizing Cluster-Based Language Models", *Proc. ACM SIGIR*,2006.

[14] Liu B, W.Hsu, and Y.Ma, "Integrating Classification and Association Rule Mining", *Proceedings ACM SIGKDD 1998* pp 80-86, 1998.

[15] Masseglia F., P. Poncelet, and M. Teisseire, "Incremental Mining of Sequential Patterns in Large Databases," *Data and Knowledge Eng*., vol. 46, pp. 97-121, July 2003.

[16] Park J.S, M.-S. Chen, and P.S. Yu, "Using a Hash-Based Method with Transaction Trimming for Mining Association Rules," *IEEE Trans Knowledge and Data Eng*, Vol. 9, No. 5, pp. 813- 825, Sept./ Oct. 1997.

[17] Park J.S, M.-S. Chen, and P.S. Yu, "Mining Association Rules with Adjustable Accuracy," *IBM Research Report*, 1995.

[18] Pasquier N. Bastide Y. Taouil R. and L.Lakhal, " Efficient mining of association rules using closed itemset lattices", *Information Systems*, Vol 24, No.1, 1999, pp.25-46.

