An Efficient Approach for Association Rule Mining

Mrs. Rashmi K.Thakur , Dr. Ketan Shah

Abstract - A great research work has been done in last decade in association rules mining (ARM) algorithms . Therefore, various algorithms were proposed to discover frequent item sets and then mine association rules. Apriori algorithm is the most frequently used algorithm for generating association rules. Apriori algorithm has some abuses, such as too many scans of the database, large load of system's I/O and vast unrelated middle item sets. In this research, we propose a novel association rule mining scheme for discovering frequent itemsets which uses clustering and graph-based approach. This approach scans database only once, and then clusters the transactions according to their length. This approach reduces main memory requirement since it considers only a small cluster at a time and hence it is scalable for any large size of the database.

Index Terms - Data mining, Apriori , Frequent Itemset, CGAR

1 INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Among the various data mining applications, mining association rules is an important one. The strategies for mining frequent itemset, which is the essential part of discovering association rules, have been widely studied over the last decade such as the Apriori, and FPgrowth.[1]

1.1 Association Rule Mining

Association rule mining, one of the most important and well researched techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc.

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second sub problem is quite straight forward, most of the researches focus on the first sub problem.

The first sub-problem can be further divided into two sub-problems: candidate large item sets generation process and frequent item sets generation process. We call those item sets whose support exceed the support threshold as large or frequent item- sets, those item sets that are expected or have the hope to be large or frequent are called candidate item sets.

In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only "interesting" rules, generating only "non redundant" rules, or generating only those rules satisfying

- *Mrs.Rashmi K.Thakur is currently pursuing masters degree in Computer Engineering in MPSTME,NMIMS University,Mumbai, India, E-mail: rashmi.thakur@thakureducation.org.com*
- *Dr.Ketan Shah is Associate Professor in Information Technology dept. of MPSTME,NMIMS University, Mumbai, India, E-mail: ketanatnmims@gmail.com*

certain other criteria such as coverage, leverage, lift or strength.[6][7]

1.2 Applications of Association Rule Mining

An association rule is an *implication* or *if-then-rule* which is supported by data. The motivation given in for the development of association rules is *market basket analysis* which deals with the contents of point-of sale transactions of large retailers[2]. A typical association rule resulting from such a study could be 90 percent of all customers who buy bread and butter also buy milk". Insights into customer behavior may also be obtained through customer surveys, but the analysis of the transactional data has the advantage of being much cheaper and covering all current customers. Compared to customer surveys, the analysis of transactional data does have some severe limitations, however. For example, point-of-sale data typically does not contain any information about personal interests, age and occupation of customers. Nonetheless, market basket analysis can provide new insights into customer behavior and has led to higher profits through better customer relations, customer retention, better product placements, product development and fraud detection.[3]

Market basket analysis is not limited to retail shopping but has also been applied in other business areas including
_ credit card transactions,
_ telecommunication service purchases,
_ banking services,
_ insurance claims, and
_ medical patient histories.[5]

2 Related Work

At present there are many data mining algorithms present. These methods have good performances and practical perspectives for static data discovering association rules, decision tree classification algorithms etc.

Many Techniques exist for frequent itemset mining. In 1994 Agrawal etc. put forward famous Apriori algorithm according to the property of association rule: the sub sets of the frequent itemset is also frequent itemset, the supersets of non-frequent itemset is also non- frequent item set. The algorithm each time makes use of k-frequent itemset carrying on conjunction to get k+1 candidate itemset.Then get k+1 frequent itemset through cutting. So keep on, until there is not frequent itemset.Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers). Apriori algorithm everytime rescans the database to enumerate the support count of each itemset and finds frequent itemsets. Han et al. [8] proposed the FP-growth algorithm using FP-tree data structure. FP-tree takes the advantage of common paths to record the sorted frequent items in transactions. In addition, FP-growth uses a header table to record the frequent items and point to the node in the FP-tree to improve the performance of mining frequent itemsets.[4]

3 Existing Techniques For Association Rule Mining

3.1 Association Rule Mining Method using CGAR

Clustering and graph-based association rule (CGAR)[2] is used for efficient association rules mining which overcome the drawbacks of the Apriori algorithms i.e rescanning the database every time . The CGAR method is accurate and effective. In today's world, the advent of the Internet has made cluster computing a powerful and cost-effective way to share and process data. CGAR can take advantage of this computing paradigm to speed up its execution time. CGAR don't scan the entire database again and again. It just scans the cluster.

3.2 CGAR – Clustering and Graph Based Association Rule [2]

Here we will see how CGAR works for association rule generation. Consider there are 18 transactions with 5 items as shown.[2]

Table 1
The original database *ODB*

TID	Items
T1	A,B,C
T2	B,C
T3	A,E
T4	A,C,D,E
T5	A,C
T6	A,C,E
T7	C,E
T8	B,C,E
T9	A,B,C,D

T10	A,D
T11	A,B,D
T12	C,E
T13	A,B,C,E
T14	C,D
T15	B,C,D
T16	A,D,E
T17	B,D,E
T18	A,C,D

In First Step Scan the database to determine the length of each transaction. n above example the maximum transaction length is 4, and so, there will be at most four clusters. These bit vectors are used in building the graph and determining the frequent 1-itemsets.

Table 2
Formation Of Clusters

	1	2	3	4	5
T2	0	1	1	0	0
T3	1	0	0	0	1
T5	1	0	1	0	0
T7	0	0	1	0	1
T10	1	0	0	1	0
T12	0	0	1	0	1
T14	0	0	1	1	0
T1	1	1	1	0	0
T6	1	0	1	0	1
T8	0	1	1	0	1
T11	1	1	0	1	0
T15	0	1	1	1	0
T16	1	0	0	1	1
T17	0	1	0	1	1
T18	1	0	1	1	0
T4	1	0	1	1	1
T9	1	1	1	1	0
T13	1	1	1	0	1

The bit vectors for the items are:[2]

BV1 = 011010011010101111

5.1.1 Reduction in Step of finding 1 frequent itemset

Instead of counting all 1's in bit vector just count number of 1's.While counting number of 1's if the number of 1's exceed minimum support stop further counting 1's.This will reduce time in generation of 1 frequent itemsets.

BV2 = 100000010111010011
BV3 = 101101111101001111
BV4 = 000010100011111110
BV5 = 010101001100110101

By counting the number of *1*s in each bit vector, we determine the support for each candidate itemset of length 1, as the following: support ({1}) = 55%, support ({2}) = 40%, support ({3}) = 45%, support ({4}) = 65%, and support ({5}) = 0.45.Thus the frequent 1-itemsets are: {{1}, {3}, {4}, {5}} as their supports are not less than 45%.

The second step is started by making logical and between each pair of frequent 1 itemsets, as we mentioned earlier in this paper, and by assigning 30% as a new value to the minimum support threshold, we found that the frequent 2- itemsets will be: {{1, 3}, {1, 4}, {3, 5}}, and the graph is constructed by drawing an edge between each pair of frequent items.

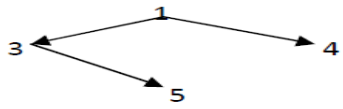

Fig 1: The Grph constructed as a result of CGAR

To determine frequent 3-itemsets, we traverse the graph as if there is a path among three nodes {i , j} and {j, k} then the set {i, j, k} will be frequent 3-itemset. Here, in this example, {{1, 3, 5}} is the only frequent 3-itemsets. As there are no extra edges, the algorithm terminates.

4 Problem Definition

In the some situation when the database contains hundreds of thousands of transactions and different items, constructing only one graph is not practical, hence there is requirement of different graphs for each cluster and finding from this graph all frequent item sets. Then combine the subsets of frequent item sets together to get the whole set of frequent item sets which is time consuming process.

5 Proposed System

To overcome the problem of graph construction, an improvement in CGAR is suggested in this paper which uses clustering approach for finding frequent itemsets. The main objective of this research is to group transactions into only 2 groups .

5.1.2 Reduction in step of finding 2 frequent itemset

When we are finding 2 frequent itemset if any of the items support is 100%,there is no need of combing that item with other items and finding the support. simply combine item with 100% support with others. This step will reduce time for generation of 2 frequent itemsets.

5.1.3 Alternative for graph Structure

When we find 2 frequent itemsets, instead of constructing graph, just examine 2 frequent itemsets support. Group into 2 categories. First category (A) is frequent itemsets having same support and second category (B) is having different support. Then local self join category A with Category B to produce 3 frequent itemsets. Find supports of 3 frequent item sets. Again examine support and insert into category A and Category B. Continue process till Single Category Exists.

6 CONCLUSION

In this paper we have discussed the traditional data mining techniques such as Apriori.This methods has the disadvantage of multiple scans which wastes time.

Hence we have discussed CGAR technique to avoid multiple scan problem. This method uses graph structure for 3 and more itemset generation. We need to construct multiple graphs as number of clusters increases.

So we have proposed a novel method in which clustering is done on the basis of support.Only 2 clusters are formed.Also we donot require any graph structure to be constructed. Intuitively we can say that this method will be more efficient than existing methods such as Apriori.

REFERENCES

[1] Chen, M. S., Han, J., & Yu, P. S. "Data mining: An overview from a database perspective", IEEE Transactions on Knowledge and Data Engineering," 8(6), 866–883,1996

[2] Wael A. AlZoubi, Azuraliza Abu Bakar, Khairuddin Omar, " Scalable and Efficient Method for Mining Association Rules ", Proc. Of IEEE, vol. 1,no.8, 2009, pp. 36-41.

[3] Lijuan Zhou, Shuang Li, Mingsheng Xu "Research on Algorithm of Association Rules in Distributed Database System", Proc. Of IEEE, vol. 3,no.9, 2010, pp. 216-219.

[4] Yu Shaoqian, "A kind of improved algorithm for weighted Apriori and application to Data Mining", Proc. Of ACM, 2010, pp. 507-510.

[5] Rui Chang, Zhiyi Liu "An Improved Apriori Algorithm", ACM Transactions on Database Systems, Vol. 9, No. 4,2011,pp.403-408

[6] Jianwei Li, Ying Liu, Wei-keng Liao, Alok Choudhary, "Parallel Data Mining Algorithms for Association Rule and Clustering," 2006 by CRC Press, LLC

[7] Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," Data Mining and Knowledge Discovery, 8, 53–87, 2004

[8] Jian Pei, Jiawei Han, Hongjun Lu, Shojiro Nishio, Shiwei Tang, Dongqing Yang, "H-Mine: Fast and space-preserving frequent pattern mining in large databases," Data Mining and Knowledge Discovery, 8, 53-87, 2004