RESEARCH ARTICLE OPEN ACCESS

Adaptive and Fast Predictions by Minimal Itemsets Creation

Ms. Sonali Vaidya*, Mr. Struan Souto**, Mr. Keegan Pereira**, Mr. Vincent Soares**, Mr. Neil Rego**
*,**(Department of Information Technology, Mumbai University, St. Francis Institute of Technology, Mumbai)

ABSTRACT
Association rules in data mining are useful for the analysis and prediction of an individual user's behavior which facilitates the data analysis on a regular basis for market basket data, clustering of products, designing catalogs and playing an immense role for store layout setting. This paper presents a successor of the Apriori and variation of the CHARM algorithm, which is the MG-CHARM algorithm that is used for finding relationships between certain attributes instead of the whole dataset. The MG-CHARM algorithm is an Association Rule Mining (ARM) algorithm that is used to select the target database which in turn takes less time to find the desired association rules. In the proposed implementation, the actions performed by a user will be effectively recorded in the target database. The datasets that are generated will have to be fed to the semi - automated, adaptive software which will fetch the output based on ARM algorithm. The algorithm will determine the relations for different datasets by mining Minimal Generators (mGs) from Frequently Closed Itemsets (FCI's) to carry out decision making and pattern analysis of the Itemsets.
Keywords–Apriori, CHARM, MG-CHARM, ARM, mG's, FCI's

I. INTRODUCTION

Programmers use association rules to build programs capable of machine learning. This Association Rule Mining (ARM) algorithm is used to select the target database which in turn takes less time to find the desired association rules. The incorporating user's preference in selection of target attribute also helps to search the association rules efficiently both in terms of space and time.

II. NATURE OF PROBLEM

The key element that makes association rule mining practical is the minimal support and threshold which is used to prune the search space and to limit the number of frequent itemsets and rules generated. Some of the major drawbacks of association rule algorithms are huge number of considered rules, generation of non-interesting rules and low algorithm performance. At present, almost all algorithms for mining Minimal Generators of FCIs are based on Apriori algorithm. The first method to find mGs extended from Apriori algorithm found candidates that are mGs, and then defined their closures to find out frequent closed itemsets. Firstly, all frequent closed itemsets were found using CHARM algorithm. Then using level-wise method all mGs that correspond to each closed itemset was found. Both of these methods have disadvantage in large size of frequent itemsets since the number of considered candidates is large. Also the time and space involved is too large since all the itemsets are taken into consideration. Also association rule mining of different datasets taking into consideration a dynamic environment is troublesome.

III. PREVIOUS WORK

The past few years have seen a tremendous interest in the area of data mining. Data mining is generally thought of as the process of finding the information associated or included in a large collection of data which is thought to be hidden, non-trivial and previously unknown. Finding regularities data patterns can be done by a trivial data mining component and an important class of methods referred to as the association rules. Association mining has been used in many application domains and therefore it paves its way as the most important model invented and extensively studied by databases and data mining community. Discovering of purchase patterns or association between products is very useful for decision making and effective marketing which is an immense part of the business field. Biological databases pattern exploration, software engineering metrics has inclusive knowledge that can be extracted or retrieved periodically, personalization of web content, and mining of textual data are some of the applications recently developed taking into consideration the field of data mining. Most of the research efforts in the scope of association rules have been oriented to simplify the rule set and improve the performance of the algorithm. But these are not the only problems that can be found when rules are generated and applied in different domains.

Troubleshooting for them should also be taken into consideration. The purpose of association model and data they come from also need to be understood.

IV. PURPOSE

The proposed solution focuses on overcoming the drawbacks of the Apriori and CHARM algorithms. In this technique the number of frequent closed itemsets (FCIs) generated is usually fewer than the number of frequent itemsets and therefore it is necessary to find Minimal Generators (mGs) for mining association rule from them. Exploring of one or more mG's depends on the approaches based on generating candidate lose timeliness when the number of frequent closed itemsets is large. Finding all mGs of frequent closed itemsets is done efficiently and effectively by the MG-CHARM in a closed hierarchical manner. Using the ARM algorithm, an adaptive system which does not generate candidates, by mining directly the mGs of frequent closed itemsets during the mining of FCI's can be developed. Thus, the time for finding mGs of frequent closed itemsets is insignificant. Also mining (minimal) non-redundant association rules is possible. The system can effectively handle different datasets in order to provide an output of association rules, thereby facilitating data visualization in order to manage the data in an organized manner. It will also work to provide greater accuracy and will perform reliably at a given point of time.

V. CONTRIBUTION OF PAPER

This paper aims to use and detect the frequency patterns i.e. the minimally frequent itemsets taking into consideration a large number of datasets. However to effectively provide suggestions in real time, the datasets need to fed to the software separately at regular intervals of time to generate the necessary output. Association Rule Mining and Data Visualization will be carried out for the same and yield the necessary output. Mining of any dataset is possible. Also the software is semi-automated when it comes to finding out association rules for a given support and confidence and thus time and work effort is also reduced.

VI. FIGURES AND TABLES

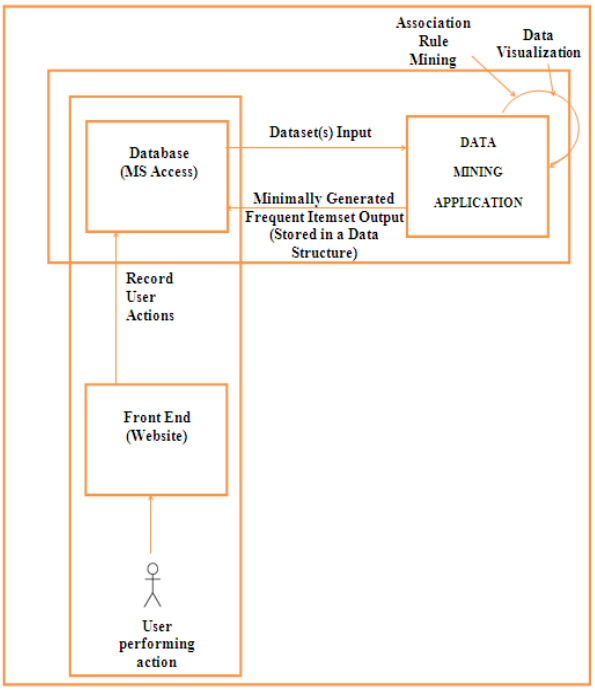

Fig. 1

Fig.1 speaks about the consideration and the mining of association rules from datasets mined or fetched from the target database.

The system implementation can be carried out in two parts:

1. User to database: To record user action(s).
2. Database to software: To analyze, visualize, generate and return an association rule output.

The ARM algorithm will determine effectively the relations for different datasets by mining Minimal Generators (MGs) from Frequently Closed Itemsets (FCI's).

Data Visualization can be effectively carried out to study and visualize the data. This in turn will aid in effective decision making. The output of the ARM algorithm will be returned back to the database.

In order to prevent system bottlenecks and allow the system to continue its tasks without having to stop or suspend its working, the mining will have to be carried out periodically depending on the urge and need of the organization to analyze the data. The idea of mining the FCI's periodically is to enable the system to yield suggestions with ease and guide the user, thereby avoiding any difficulty while accessing the system at any point in time.

The system will hold true for any dataset. This will therefore give an organization a clear cut idea of how to use the data provided as well as the data generated in an efficient manner.

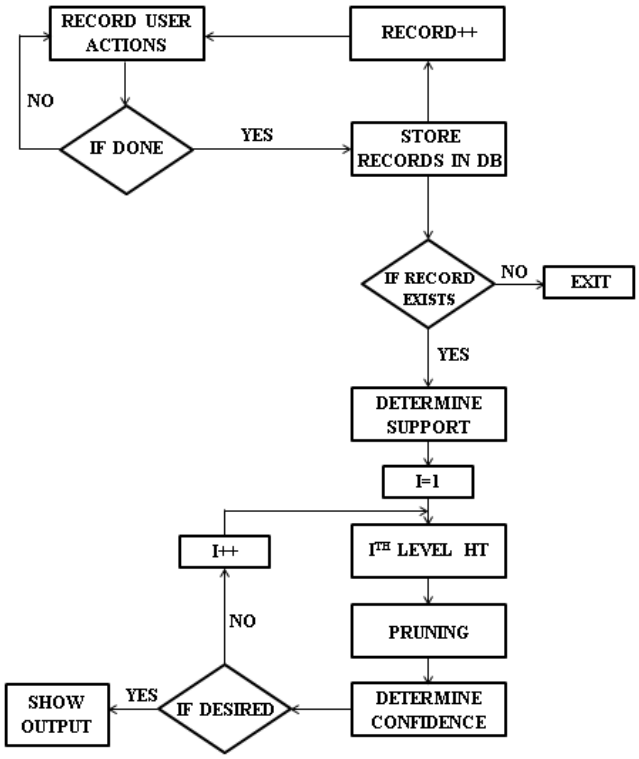

Fig. 2

Fig.2 represents a flow chart of the system in which a determined support and confidence will give the required output for a recorded user action provided the details of the user activity for example: A user transaction for certain items at a store, is recorded in the database. The itemset relation is obtained by pruning which is immensely important and saves time and memory.

VII. THE ALGORITHM

Input: The database D and support threshold minSup
Output: all FCI satisfy minSup and their mG
Method:
MG-CHARM(D, minSup)
[Φ]={lj× t(li) ,(Ii): Ii ∈I/\ σ (li) ≥minSup}
MG-CHARM-EXTEND([Φ], C = Φ)
Return C
MG-CHARM-EXTEND([P], C)
for each li × t(lj), mG(li) in [P] do
Pi= Pi ∪ li and [Pi]=Φ
for each lj× t(lj), mG(lj) in [P], with j > i do
X = lj and Y=t(li) ∩ t(lj)
MG-CHARM-PROPERTY(X ×Y,li,lj,Pi, [Pi],[Pj)
SUBSUMPTION-CHECK(C, Pi)
MG-CIIARM-EXTEND([Pi],C)
MG-CHARM-PROPERTY(X ×Y,li,lj,Pi, [Pi],[Pj)
ifσ (X) ≥minSup then
if t(li) = t(lj) then // property I
Remove lj from P
Pi=Pi ∪lj
mG(Pi) = mG(Pi) + mG(lj)
else if t(li) ⊂ t(lj) then // property 2
Pi=Pi ∪lj
else if t(li) ⊂ t(lj) then // property 3
Remove Ij from [P]
Add X ×Y,mG(Ij) to [Pi]
else if t(li) ≠ t(1j) then // property 4
Add X ×Y ,∪ [mG(li), mG(lj)]to [Pi]
The MG-CHARM(D, minSup) function takes in input Database D and the minimum support necessary for the item sets . [Φ]- denotes the set of all Item-Transaction pair having minimum support
The MG-CHARM-EXTEND([P], C) function is is Recursive function until all possible mG's are generated. This function calls the MG-CHARM-PROPERTY(XXY,li,lj,Pi, [Pi],[Pj])
Function where it states that
If t(Xi) = t(Xj) then c(Xi) = c(Xj) = c(Xi∪Xj).
If t(Xi) ⊂ t(Xj) then c(Xi) ≠ c(Xj) but c(Xi) = c(Xi∪Xj).
If t(Xi) ⊃ t(Xj) then c(Xi) ≠ c(Xj) but c(Xj) = c(Xi∪Xj).
lf t(Xi) ⊄ t(Xj) and t(Xj) ⊄ t(Xi) then c(Xi) ≠ c(Xj) ≠ c(Xi∪Xj).
The SUBSUMPTION-CHECK function checks whether frequent itemset Pi is closed or not. If it is, add it into C, otherwise it will be removed and its mGs ,will be added to its parent closed. It uses hash-table to store C, therefore the time of checking is insignificant.
∪ [mG(li), mG(lj)] are minimal generators of Xi ∪Xj.
The algorithm stated in [11] has been used for the effective implementation of the adaptive system in data mining to predict and analyze the output as per requirement

VIII. EXPERIMENTAL RESULTS AND COMPARISONS

The experiment has been performed on a JAVA Swing application coded on a Windows 7 OS, CPU: 2.93 GHz, 2 GB RAM
Let I = {i1, i2, …, in} be a set of items, T = {t1, t2, …, tm} be a set of transaction identifiers (tids or tidset) in a database D. The input database is a binary relation δ⊆ I × T. If an item i occurs in a transaction t, we write it as (i,t) ∈δ or iδt.
Consider the following transaction database

Transaction	Items Bought
1	1 3 4
2	2 3 5
3	1 2 3 5
4	2 5
5	1 2 3 5

Table 1

For the given transaction database the assumed support is 50 % and confidence is 50 %. The association rules generated for the assumption is shown in Table 2:

Algorithm	Output (Association Rules)	Time (ms)	Maximum Memory Usage (Mb)	No. of association rules generated

APRIORI	3 ==> 1 #SUP: 3 #CONF: 0.75
	1 ==> 3 #SUP: 3 #CONF: 1
	3 ==> 2 #SUP: 3 #CONF: 0.75
	2 ==> 3 #SUP: 3 #CONF: 0.75
	5 ==> 2 #SUP: 4 #CONF: 1
	2 ==> 5 #SUP: 4 #CONF: 1
	5 ==> 3 #SUP: 3 #CONF: 0.75
	3 ==> 5 #SUP: 3 #CONF: 0.75
	3 5 ==> 2 #SUP: 3 #CONF: 1
	2 5 ==> 3 #SUP: 3 #CONF: 0.75
	2 3 ==> 5 #SUP: 3 #CONF: 1
	5 ==> 2 3 #SUP: 3 #CONF: 0.75
	3 ==> 2 5 #SUP: 3 #CONF: 0.75
	2 ==> 3 5 #SUP: 3 #CONF: 0.75
		~ 2	~ 22	14

CHARM	3 ==> 1 #SUP: 3 #CONF: 0.75
	1 ==> 3 #SUP: 3 #CONF: 1
	5 ==> 2 #SUP: 4 #CONF: 1
	2 ==> 5 #SUP: 4 #CONF: 1
	3 5 ==> 2 #SUP: 3 #CONF: 1
	2 5 ==> 3 #SUP: 3 #CONF: 0.75
	2 3 ==> 5 #SUP: 3 #CONF: 1
	5 ==> 2 3 #SUP: 3 #CONF: 0.75
	3 ==> 2 5 #SUP: 3 #CONF: 0.75
	2 ==> 3 5 #SUP: 3 #CONF: 0.75
		~ 1	~ 25	10

MG-CHARM	2 3 ==> 5 #SUP: 3 #CONF: 1
	3 5 ==> 2 #SUP: 3 #CONF: 1
	3 ==> 2 5 #SUP: 3 #CONF: 0.75
	3 ==> 1 #SUP: 3 #CONF: 0.75
	1 ==> 3 #SUP: 3 #CONF: 1
	2 ==> 3 5 #SUP: 3 #CONF: 0.75
	2 ==> 5 #SUP: 4 #CONF: 1
	5 ==> 2 3 #SUP: 3 #CONF: 0.75
	5 ==> 2 #SUP: 4 #CONF: 1
		~ 0	~ 4	9

Table 2

The results in Table 2 show a mere comparison of the association rule mining algorithms in terms of total time required for mining rules, total memory required and the number of association rules generated. MG-CHARM is faster than CHARM which in turn is faster than the traditional Apriori algorithm. The time required for mining FCI's and mining association from the FCI's is sufficiently less in case of MG-CHARM which makes it faster than the prior algorithms. Also the space required is less for the same datasets considered and for the assumed support and confidence. Mining of non-redundant association rules is also possible using the MG-CHARM algorithm which facilitates generation of non-redundant queries.

IX. RELATED WORK

3.1 In [10] the proposed method for mining mG's of FCI's needs not generate candidates. Experiments showed that the time of updating mG's of frequent closed itemsets is insignificant. Especially, in case of the large size of closed itemsets, the time of updating is very fewer. Some applications of mGs in mining non-redundant association rules (NARs) based on methods presented in [3,5].

3.2 Method of Bastide et al

In this section, we present a method for mining minimal NARs (MINARs) based on FCI's (Bastide et al [3]). Assume we had all FCI's from database D satisfy minSup. Now, we want to mine MINARs. As we mentioned above, MINARs only generate from X →Y, where X is a minimal generator and Y is a FCI. Bastide et al divided mining MINARs problem into two sub-problems. Mining MINARs with the confidence = 100% and mining MINARs with the confidence < 100%.

Phase 1: Mining MINARs with the confidence = 100% for all g belongs to E mG's (in ascending length), if g →y(g) then generate the rule { g →y(g) \ g} (y(g) is closure of g).

Phase 2: Mining MINARs with the confidence < 100% for k = 1 to μ-1 do // μ is longest FCI's for all g belong to mG's with |g| = k do generate rules from g →Y \ g, where g is a subset of Y and Y is a FCI (Y is not equal to (g)).

3.3 METHOD OF ZAKI

Zaki [5] mined the two kinds of rules. i) Rules with the confidence = 100%: Self-rules (the rules that generating from mG's(X) to X, where X is a FCI) and Down-rules (the rules that generating from mG's(Y) to mG's(X), where X, Y are FCI's, X is a subset of Y). ii) Rules with the confidence < 100%: From mG's(X) to mG's(Y), where X, Y belongs to FCI's, X is a subset of Y. Number of rules was generated by this approach is linear with FCI's.

X. CONCLUSION

In this paper, the presented MG-CHARM algorithm is used for mining minimal generators of frequent closed itemsets. By mere comparisons, experimental study, and paper research, the bottleneck identified with the CHARM algorithm is that the number of frequent items is large and it takes more time. To solve this problem the numbers of items were decreased the iterations and new comparison methodologies were used by enhancing CHARM. The implementation proposed defines a generic basis for exact association rules and transitive reduction of the informative basis for approximate association rules, non-redundant in nature and does not represent any loss of information from the user's point of view. Visualization and analysis of the association rules is possible. Altogether the intended system can be made to work and function in a dynamic environment. The project has future scope when very large numbers of datasets need to be taken into consideration. The software will be fruitful and show effective performance when used with a Big Data Database. Providing suggestions to the user on the basis of a dynamic scenario will be put into effect in the future.

Acknowledgements

We hereby take the privilege to present our project report on "**Adaptive and Fast Predictions by Minimal Itemsets Creation**". We are very grateful

to our Project Supervisor **Ms. Sonali Vaidya** for contributing her valuable moments in the Project from her busy and hectic schedule right from the Project's inception. Being after us like a true mentor and a great academic parent.
We are very thankful to Ms. Sonali Vaidya whose guidance and support was an immense motivation for us to carry on with our Project. She has been a constant source of inspiration to us. Her suggestions have greatly contributed for the betterment of our project.
Our special thanks to the Head of Department Mr. **Pramod Shanbhag**, staff members and lab assistants for their co-operation.

REFERENCES

[1] R. Agrawal, R. Srikant: *Fast algorithms for mining association rules*. In: VLDB'94, pp(1994).
[2] R. Agrawal, T. Imielinski, A. Swami: *Mining association rules between sets of items in large databases*. In: SIGMOD'93, pp. 207 - 216 (1993).
[3] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, L.Lakhal: *Mining Minimal Non-Redundant Association Rules using Closed Frequent Itemssets*. In: 1st International Conference on Computational Logic, pp. 972 - 986 (2000).
[4] M. 1. Zaki, C.J. Hsiao: *Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structur*e. In: IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No 4, April 2005, pp. 462-478 (2005).
[5] M. 1. Zaki: *Mining Non-Redundant Association Rules, Data Mining and Knowledge Discovery*, 9, 223-248, 2004 Kluwer Academic Publishers. Manufactured in The Netherlands, pp. 223-248 (2004).
[6] M. J. Zaki, K. Gouda: *Fast Vertical Mining Using Diffsets*. In: Proc. Ninth ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining, pp. 326 335 (2003).
[7] M. J. Zaki, B. Phoophakdee: MIRAGE: *A Framework for Mining, Exploring, and Visualizing Minimal Association Rules*. In: Technical Report 03-4, Computer Science Dept., Rensselaer Polytechnic Inst., July (2003).
[8] M. J. Zaki: *Generating Non-Redundant Association Rules.* In: 6th ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining, (2000).
[9] S. B. Yahia, T. Hamrouni, E. M. Nguifo: *Frequent Closed Itemset based Algorithms: A thorough Structural and Analytical Survey*. In: ACM SIGKDD Explorations Newsletter 8 (1), pp. 93 - 104 (2006).
[10] Bay Vo, Bac Le: *Fast Algorithm for Mining Minimal Generators of Frequent Closed Itemsets and Their Applications: A thorough Structural and Analytical Survey.* year-2009.