Volume 2, Issue 2, February 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Eliminating Redundant Frequent Pattern's In Non – Taxonomy Data Sets

R. Vijaya Prakash*
Department of Informatics
Kakatiya University

Prof. A. Govardhan
Dept of Comp. Sci. Eng. &
JNT University

Prof. S.S.V.N. Sarma
Department of Computer Sci, & Eng.
Vaagdevi College of Engineering

***Abstract* - Frequent pattern mining is an important area of data mining used to generate the association rules. The extracted Frequent Patterns quality is a big concern, as it generates huge sets of rules and many of them are redundant. Mining Non Redundant frequent patterns in Non –Taxonomy Data sets is a big concern in the area of association rule mining. In this paper we proposed a method to eliminate the redundant frequent patterns, to generate the quality association rules.**

***Keywords*— Frequent Pattern, Closed Pattern, Association Rule Mining (ARM), Non Redundant Itemset, Itemset.**

I. INTRODUCTION

Frequent pattern mining is an important area of Data mining research. The frequent patterns are patterns (such as itemsets, subsequences, or substructures) that appear in a data set frequently. For example, a set of items, such as milk and bread that appear frequently together in a transaction data set is a *frequent itemset*. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a *frequent sequential pattern*. A *substructure* can refer to different structural forms, such as subgraphs, subtrees, or sublattices, which may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a *frequent structured pattern*.[1] Finding such frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data classification, clustering, and other data mining tasks as well.

The process of discovering interesting and unexpected rules from large data sets is known as association rule mining. This refers to a very general model that allows relationships to be found between items of a database. An association rule is an implication or if-then-rule which is supported by data. The association rules problem was first formulated in [2] [3] and was called the market-basket problem. The initial problem was the following: given a set of items and a large collection of sales records, which consist in a transaction date and the items bought in the transaction, the task is to find relationships between the items contained in the different transactions. A typical association rule resulting from such a study could be "90 percent of all customers who buy bread and butter also buy milk" – which reveals a very important information. Therefore this analysis can provide new insights into customer behavior and can lead to higher profits through better customer relations, customer retention and better product placements.

Mining of association rules is a field of data mining that has received a lot of attention in recent years. The main association rule mining algorithm, Apriori, not only influenced the association rule mining community, but it affected other data mining fields as well. Apriori and all its variants like Partition, Pincer-Search, Incremental, Border algorithm etc. take too much computer time to compute all the frequent itemsets. The papers [4], [5] contributed a lot in the field of Association Rule Mining (ARM). In this paper, an attempt has been made to compute frequent itemsets by using closed frequent itemsets to remove the redundant itemsets.

II. ASSOCIATION RULE MINING (ARM)

Association Rule Mining aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [6]. The major aim of ARM is to find the set of all subsets of items or attributes that frequently occur in many database records or transactions, and additionally, to extract rules on how a subset of items influences the presence of another subset. ARM algorithms discover high-level prediction rules in the form: IF the conditions of the values of the predicting attributes are true, THEN predict values for some goal attributes

In general, the association rule is an expression of the form X=>Y, where X is antecedent and Y is consequent. Association rule shows how many times Y has occurred if X

has already occurred depending on the support and confidence value.

Support: It is the probability of item or item sets in the given transactional data base: support(X) = n(X) / n where n is the total number of transactions in the database and n(X) is the number of transactions that contains the item set X. Therefore,

$$supp(X \Rightarrow Y) = p(X \cup Y) \text{ or } supp(X \cup Y) \quad (1)$$

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)},$$

$$p(Y / X) = \frac{p(X \cup Y)}{p(X)} \quad (2)$$

Frequent itemset: Let A be a set of items, T be the transaction database and *minsup* be the user specified minimum support. An itemset X in A (i.e., X is a subset of A) is said to be a *frequent itemset* in T with respect to *minsup* if $support(X)_T > minsup$

The problem of mining association rules can be decomposed into two sub-problems:

- Find all itemset whose support is greater than the user-specified minimum support, *minsup*. Such itemsets are called *frequent itemsets*.
- Use the frequent itemsets to generate the desired rules. The general idea is that if, say ABCD and AB are frequent itemsets, then we can determine if the rule AB=>CD holds by checking the following inequality support({A,B,C,D}) / support({A,B}) > *minconf*, where the rule holds with confidence *minconf.*

To demonstrate the use of the support-confidence framework, we illustrate the process of mining association rules by the following example

Example 1. Assume that we have a transaction database in a supermarket, as shown in Table I. There are six transactions in the database with their transaction identifiers (TID's) ranging from 100 to 600. The universal itemset I ={A, B, C, D, E}, where A, B, C, D and E can be any items in the supermarket. For instance, A = ''bread", B = ''milk", C = ''sugar", D = ''coffee", and E = ''biscuit".

TABLE I :
EXAMPLE TRANSACTION DATABASE

TID	Items Bought
100	ABDE
200	BCE
300	ABDE
400	ABCE
500	ABCDE
600	BCD

There are totally 25(=32) itemsets. {A}, {B}, {C}, {D}, and {E} are all 1-itemsets, {AC} is a 2-itemset, and so on. All Frequent Itemset with min support =50% is

TABLE II:
GENERATED FREQUENT ITEMSETS

	Itemsets	Support
1	B	100%
2	E, BE	83%
3	A, C, D, AB, AE, BC, BD, ABE	66%
4	AD, CE, DE, ABD, ADE, BDE, BCE, ABDE	50%

ABDE, BCE are maximal-by-inclusion frequent itemsets i.e., they are not a subset of any other frequent itemset.

A. Generating confident rules

This step is relatively straightforward; rules of the form X=>Y, Where X, Y are generated frequent itemset with *minconf.* The following Table III shows the generated confidence rules with minconf ≥ 1.0.

TABLE III:
ASSOCIATION RULES

	Association Rule	Confidence
1	A→B, A→E, A→BE, C→B, D→B, E→B	100%
2	AB→E, AD→B, AD→E, AE→B,	100%
3	AD→BE,DE→AD, ABD→E,	100%
4	ADE→B, BDE→A	100%
5	CE→B, DE→A, DE→B	100%

Looking at Table III, the rules contained within are considered to be useful based on the fact their support and confidence values meet or exceed a predefined minimum support and minimum confidence. However, some of these rules do not contain or present new information to a user. In particular, the consequent concluded by some rules can be obtained from other rules with the same or higher confidence level but without requiring more conditions to be satisfied. For example, we can obtain the rule A⇒B by transitivity from the two rules A⇒E, and E⇒B. The rule CE⇒B can be obtained by augmentation of the two rules E⇒B and C⇒B, etc. We can see that the redundant rules have an antecedent of equal or greater length and a consequent of equal or shorter length respectively, while the confidence of the redundant is not greater than the corresponding non-redundant rules. From this the following definition defines this kind of redundant rules.

***Definition: (Redundant Rules)*:** If we let $X \Rightarrow Y$ and $X' \Rightarrow Y'$ e two association rules with confidences *cf* and *cf'* respectively, then $X \Rightarrow Y$ *is* said to be a redundant rule to $X' \Rightarrow Y'$ if $X' \subseteq X$, $Y' \subseteq Y$ and $cf \leq cf'$.

From definition 1 if we have an association rule $X \Rightarrow Y$, if there is no other rule $X' \Rightarrow Y'$ in existence such that the confidence of $X' \Rightarrow Y'$ is equal to or larger than the confidence of $X \Rightarrow Y$ and $X' \subseteq X$, $Y' \subseteq Y$, then the association rule $X \Rightarrow Y$ is said to be non-redundant.

Definition 1 is similar to the definition of MinMax association rules as defined in [7] in terms of requiring a shorter antecedent and a longer consequent. The definition of MinMax association rules however requires that a redundant rule and the corresponding non - redundant must have identical confidence and support. Definition 3.5 only requires that the confidence of the redundant rule is not larger or greater than that of the non-redundant rule.

Eliminating redundant association rules safely without damaging the capacity of the remaining rules is essential and it is crucial to successfully define a boundary between redundant and no redundant in order to ensure safe

redundancy removal. Several different approaches to achieve this have been proposed [8][9][10]. However none have specifically discussed the boundary. This approach presented here proposed to use the Certainty Factor (CF) to determine the boundary. If deleting an association rule does not reduce the CF value of the remaining association rules then the deletion of that rule is considered to be safe.

The concept of the certainty factor was first proposed in [11] in order to express the level of accuracy and truth behind an association rule and also determine how reliable the antecedent of the given rule is. Certainty factor is founded on two functions; the measure of belief $\delta(X,Y)$ and the measure of disbelief $\gamma(X,Y)$ for a rule of the form $X \Rightarrow Y$. The functions of δ and γ are given as follows:

$$\delta(X,Y) = \begin{cases} 1 & P(Y) = 1 \\ 0 & P(Y/X) \le P(Y) \\ \dfrac{P(Y/X) - P(Y)}{P(Y)} & otherwise \end{cases} \quad (3)$$

$$\gamma(X,Y) = \begin{cases} 1 & P(Y) = 0 \\ 0 & P(Y/X) \ge P(Y) \\ \dfrac{P(Y) - P(Y/X)}{P(Y)} & otherwise \end{cases} \quad (4)$$

Where P*(Y/X)* and P*(Y)* in the context of association rules represent the confidence of the rule and the support of the consequent respectively. For both δ and γ the values range between 0 and 1 and measure the strength of the belief or disbelief in the consequent *Y* given the antecedent *X*. Thus, $\delta(X,Y)$ weighs how much the antecedent *X* increases the possibility of consequent *Y* occurring, while $\gamma(X,Y)$ weighs how much the antecedent *X* decreases the possibility of consequent *Y* occurring. If *P(Y/X)* equals 1, then the antecedent completely supports the consequent and thus $\delta(X,Y)$ will be 1. On the other hand, if *P(Y/X)* is equal to 0, then this indicates that the antecedent completely denies the consequent and thus $\gamma(X,Y)$ will be 1. The total strength of the belief or disbelief captured by the association rule is measured by the certainty factor, which is defined as follows:

$$CF(X,Y) = \delta(X,Y) - \gamma(X,Y) \quad (5)$$

The value of the certainty factor will be between 1 and -1, where negative values represent the cases where the antecedent is against (denying) the consequent. Positive values indicate that the antecedent supports the consequent. A certainty factor value of 0 means that the antecedent does not influence (neither supporting nor denying) the consequent. Association rules with a high certainty factor value are the most useful as they represent strong positive associations between the rule's antecedent and consequent. The aim of association rule mining is to discover these rules that have strong positive associations. It is therefore proposed that the certainty factor can be used to measure the strength of discovered association rules

The following theorem (3.1) states that the certainty factor value of a redundant rule, as defined by Definition 3.5 will never be greater than the certainty factor of the corresponding no redundant rules. It thus means that the association between the antecedent and consequent of the non-redundant rule is stronger than any corresponding redundant rule.

We applied the Definition 1 and δ, γ *and CF* measures on Table I transaction database, we may get the following non – redundant association Rules.

TABLE IV :
NON – REDUNDANT ASSOCIATION RULES

	Association Rule (X→Y)	***δ(X→Y)***	***γ(X→Y)***	***CF(X→Y)***	**Confidence**
1	AD→BE	0.204	0.796	-0.596	100%
2	DE→AD	1.00	0	1.00	100%
3	ABD→E	1.00	0	1.00	100%
4	ADE→B	0.00	1	-1.00	100%
5	BDE→A	0.50	0	0.5	100%

From the above Table IV we may say that the association rules 2 and 3 are strong belief non redundant association riles for the given antecedent and consequent. Where as the association rule 4 is a strong disbelief non redundant association rules.

III. EXPERIMENTS

The Datasets used in these experiments were obtained from UCI KDD Machine Learning Repository (http://kdd.ics.uci.edu/). The Mushroom dataset contains 8,124 transactions each of which describes the characteristics of one mushroom object. Each mushroom object has 23 attributes. The other datasets included in these experiments is Census income. The Census income dataset contain 32,561 transactions each of which describes the characteristics of a person's income, the Census income data set object has 15 attributes. They produce large numbers of frequent itemsets and thus a huge number of association rules even for very high values of support. Redundancy elimination is particularly important to these dense datasets. The no. of exact rules and redundancy elimination is given in Table V.

TABLE V:
THE NON REDUNDANT RULES For CENSUS INCOME And MUSHROOM DATA SETS

	Census Income Data Set		Mushroom Data Set	
min_conf	FCI	Non Redundant	FCI	Non Redundant
0.1	191	38	1709	427
0.5	32	12	331	89
0.9	04	03	56	24

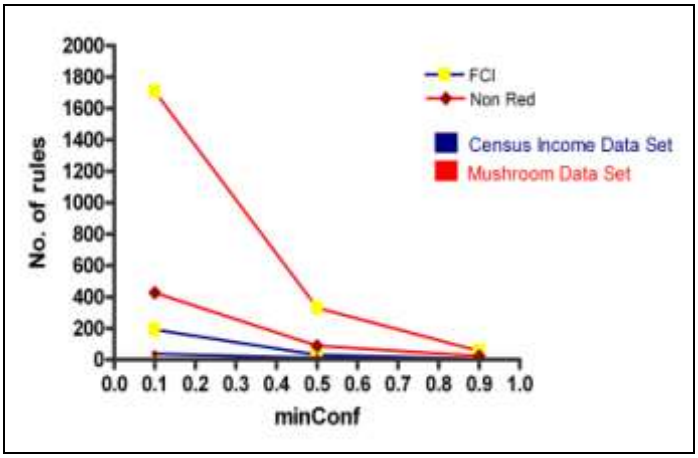

Fig:1 Non Redundant Rules for Census income and Mushroom Data sets

We conducted the experiments on census income and mushroom data sets with minimum support 0.5, and we have taken different min confidence values 0.1, 0.5 and 0.9. At 0.1 min confidence 0.1, min support = 0.5, there are large no. of association rules are generated with redundancy, this are removed with our given proposal.

IV. CONCLUSIONS:

The challenging problem of generating the association rule mining is redundancy exists in the extracted association rules, which may affect the quality of the association rules and frequent pattern generation in Non – Taxonomy Data sets.. With the above approach we can remove the redundant frequent patterns, to get quality frequent patterns and then association rule mining.

REFERENCES

[1]. Jiawei Han and Micheline Kamber, *Data Mining Concepts and Techniques*, Second Edition, Morgan Kaufmann Publishers,

[2]. R. J. Bayardo and R. Agrawal, *Mining the most interesting rules*. In 5th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining, August 1999.

[3]. S. Brin., R. Motwani, J. Ulmann and S. Tsur., *Dynamic Itemset counting and implication rules for market basket data*. In ACM SIGMOD Conf. Management of Data, May 1997.

[4]. A. Savasere .E, Omiecinski and S. Navathe., *An efficient Algorithms for mining association rules in large database* in 21st VLDB conf. 1995.

[5]. H. Toivonen, M. Klememminen., P. Ronkaiven, K. Hatonen and H. Manika, *Pruning and grouping discovered association rules*. In ML Net wkshp on statistics, Machine Learning and discovery in databases, April 1995.

[6]. D.I. Lin. And Z.M. Kedam – pincer search, *A new algorithm for discovering maximum frequent sets*. In 6th Int. Conf. Extending database technology, March 1998.

[7]. N. Pasiquir, Bastide .Y, Stemme G & Lakhal, *Generating a condensed Representation for association rule*, Journal of Intellegent System 24(1), 29-60, 2005.

[8]. Calders T., & Geothals B, *Mining all non – derivalble Frequent itemsets*, In Proceedings of the 6th European conference on principles of Data mining and Knowledge Discovery (vol 2431 pp 74-85): springer – verlag 2002.

[9]. Zaki. M,.J, *Generating non redundant association rules*, Paper presented at the proceedings of the KDD conference.

[10]. N. Pasiquir, Bastide .Y, Touil R & Lakhal, *Efficient Mining of Association rules using closed itemset lattices*., Information System (24), 25-46.

[11]. Shortliffe. E.H., & Buchanan B.G., *A Model of inexact reasoning in medicine*, Mathematics Bioscience 23(3/4) 351-379.

About Authors:

R. Vijaya Prakash is presently working as Associate Professor in Department of Computer science & Engineering, SR Engineering College., Pursuing Ph.D from Kakatiya University, Warangal. He completed his MCA from Kakatiya University and M.Tech (IT) form Punjabi University, Patiala.

Prof. A. Govardhan is working as professor in Department of Computer Science & Engineering, JNT University, Hyderabad. He Received his B.Tech in 1992 from Osmain University, M.Tech in 1994 from Jawahar Lal Nehru University, Delhi and Ph.D from JNT University in 2003. He is very eminent person in the field of teaching he guided 4 Ph.D scholars.

Prof. SSVN. Sarma is an eminent professor in Mathematics, He Worked in Department of Mathematics and Department of Informatics, Kakatiya University over period of 30 years. He is well known person in the field of Mathematics and Computer Science. He guided many Ph.D Scholars