*International Journal on Emerging Technologies* **2**(2): 18-20(2011)

**ISSN No. (Print) : 0975-8364**
**ISSN No. (Online) : 2249-3255**

# An Efficient Association Rule Mining Without Pre-assign Weight

***Manish Kumar Singhal* and Vineet Richariya***

*Department of Computer Science Engineering, LNCT Bhopal, (MP)*

*(Received 12 July, 2011, Accepted 10 September, 2011)*

**ABSTRACT : The association rule mining on Market "Basket Data" is Boolean Association Rule Mining in which only Boolean attributes are considered. In order to do association rule mining on quantitative data, Association rule mining is one of the important problems of data mining. The goal of the Association rule mining is to detect relationships or associations between specific values of categorical variables in large data sets. This is a common task in many data mining projects. In this paper we are using Association Rule mining without pre-assign weights and its implementation in matlab. In the past, several authors have proposed various association rule mining algorithms with and without pre-assigned weights. The paper consists of two parts. Part two consists of association rule mining for frequent itemset. Complete implementation has been done in Matlab on various real life datasets. Part one consisted of association rule mining without pre-assigned weights using binary form.**

## I. INTRODUCTION

Association rule mining (Aggarwal *et. al* [1]., 1993) is one of the important problemsof data mining. The goal of the Association rule mining is to detect relationships orassociations between specific values of categorical variables in large data sets. This is acommon task in many data mining projects. Suppose I is a set of items, *D* is a set oftransactions, an association rule is an implication of the form *X* => *Y*, where *X*, *Y* aresubsets of *I*, and *X*, *Y* do not intersect. Each rule has two measures, support andconfidence. Association rule mining was originally proposed in the domain of marketbasket data. The association rule mining on Market "Basket Data" is Boolean Association Rule.

Mining in which only Boolean attributes are considered. In order to do association rule mining on quantitative data, such as Remotely Sensing Image data, some mapping should be done from quantitative data to Boolean data. The main idea here is to partition the attribute values into *Transaction Patterns*.

Basically, this technique enables analysts and researchers to uncover hidden patterns in large data sets.

### Notation and basic concepts

Let ? = {*i*1, *i*2 … *im*} be a universe of items. Also, let *T* = {*t*1, *t*2 … *tn*} be a set of all transactions collected over a given period of time. To simplify a problem, we will assume that every item *i* can be purchased only once in any given transaction *t*. Thus *t* ? ? ("*t* is a subset of omega"). In reality, each transaction t is assigned a number, for example a transaction *id* (TID).

#### *A. Support*

The support of an itemset is the fraction of the rows of the database that contain all of the items in the itemset. Support indicates the frequencies of the occurring patterns. Sometimes it is called frequency. Support is simply a probability that a randomly chosen transaction t contains both itemsets *A* and *B*.

#### *B. Confidence*

Confidence denotes the strength of implication in the rule. Sometimes it is called accuracy. Confidence is simply a probability that an itemset B is purchased in a randomly chosen transaction t given that the itemset A is purchased In general, a set of items (such as the antecedent or the consequent of a rule) is called an itemset. The number of items in an itemset is called the length of an itemset. Itemsets of some length k are referred to as k-itemsets. Generally, an association rules mining algorithm contains the following steps:

- The set of candidate *k*-itemsets is generated by 1-extensions of the large (*k* – 1)-itemsets generated in the previous iteration.
- Supports for the candidate k-itemsets are generated by a pass over the database.
- Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large k-itemsets.

## II. PROPOSED WORK

All, there is not weights to items and transactions. Whereas, using weighted association rule mining, we have to assign weights to items and/or transactions. In Weighted association rule mining, we have to assign weight to items and/or transactions at the beginning in the database. But according to Ke.

Sun [1], we can find weights of items and weights of transactions. Method to assign weights to items on the basis of items belongs to transactions, and similarly we can find weight of transaction on the basis of items, that

are available in transactions. And also, if an item belongs in more transactions, then weight or importance of that item is high. Similarly, if a transaction that contains many items, then weight or importance of that transaction is also high. i.e., a good transaction, which is highly weighted, should contain many good items; at the same time, a good item should be contained by many good transactions. The reinforcing relationship of transactions and items is just like the relationship between hubs and transaction. So, we can find weights of items and weights of transactions by Transaction. We can assume that every transaction as a link/hub (which contain many items) and items belong to the transaction as an authority (item belongs to many link). Wang and Su [9] proposed a novel approach on item ranking. A directed graph is created where nodes denote items and links represent association rules. A generalized version of Transaction is applied to the graph to rank the items, where all nodes and links are allowed tohave weights [10].

Transactions are:

1. = Bread, Milk

2. = Bread, Beer, Diaper, Eggs

3. = Milk, Beer, Diaper, Coke

4. = Bread, Milk, Beer, Diaper

5. = Bread, Milk, Diaper, Coke

In this data set we are calculating the weights before this we are normalized the dataset and then calculating the final hub weights and after taking and calculating the weight then we apply apriori with this calculating weights. This weight helps us for finding the frequent item set. For calculating frequent item set we are taking 50% threshold value. The item whose support is less then 50% they are neglecting and then we apply 2-frequent item set and again the sets whose support is less than 50% they are neglected. and this process is repeated until all items which are less than 50%. Implementation and Result For all these implementation have been compared on different datasets. Datasets that has taken which is real life datasets as well as computer generated datasets (IBM Synthetic data generator). There are so many real life datasets which were taken, these are

**Kosarak.** The kosarak dataset comes from the click-stream data of a Hungarian online news portal, Number of Instances = 990,002, Number of Attributes = 41,270.

**Chess.** A game datasets. Attribute Information: Classes (2): – White-can-win ("won") and White-cannot-win ("nowin"). It believes that White is deemed to be unable to win if the Black pawn can safely advance. Number of Instances = 3196, Number of Attributes =36.

**Retail.** This is retail datasets, Number of Instances = 16470, Number of Attributes = 88162

**Mushroom.** This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom. Number of Instances = 8124, Number of Attributes = 22.

**Connect.** This database contains all legal 8-ply positions in the game of connect-4 in which neither player has won yet, and in which the next move is not forced. Number of Instances = 67557, Number of Attributes = 42.

**Plumbs Star.** Number of Instances = 49046, Number of Attributes = 49046.43. Some implementations have been done on IBM Synthetic data generator datasets, with different sizes, with different transaction, with different number of items in each transaction.

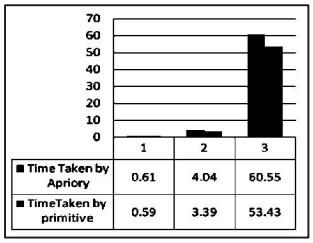
***Database name Mushroom***

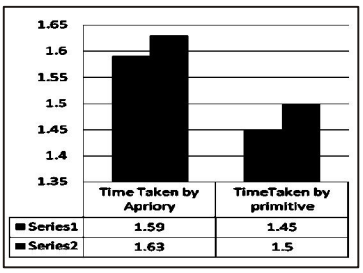
| **Large item sets** | **Time by Apriori** | **Time by Primitive** |
|---|---|---|
| 3 | 0.46 | 0.29 |
| 4 | 2.166 | 1.86 |
| 5 | 12.04 | 11.11 |
| 4 | 6.12 | 7.12 |

***Database name Retail***

| **Large item sets** | **Time by Apriori** | **Time by primitive** |
|---|---|---|
| 3 | 1.59 | 1.45 |
| 4 | 1.63 | 1.5 |

**Graphs**




## III. CONCLUSION

An efficient way for discovering the frequent set can be very useful in various data mining problems, such as discovery of association rules. In this Thesis, new

approaches to association rule mining has been explored in depth. In part one of the Thesis, weighted association rule mining without preassigned weights was discussed and implementation was done on real life datasets. Comparison of the algorithms, Apriori and Primitive Association Rule Mining was done in this section and there we found many advantages of Primitive Association Rule Mining over Apriori. Some interesting patterns may be discovered when the hub weights of transactions are taken into account. Moreover, the transaction ranking approach is precious for estimating customer potential when only binary attributes are available, such as in Web log analysis or recommendation systems

## IV. FUTURE SCOPE

For our approach, the related information may not fit in the main memory when the size of the database is very large. In the future, we shall consider this problem by reducing the memory space requirement. Also, we shall apply our approach on different applications, such as document retrieval and resource discovery in the World Wide Web environment. Best part of previously known algorithms can be combined with to develop hybrid approaches which perform best for all cases. Number of solutions has been presented, but still a lot of research is possible in this particular area. Descriptive data mining techniques were discussed in the thesis which can be further extended to explore various other approaches. Besides that, the work can be extended to perform predictive data mining task. And last but not the least; here also we are dealing with the time-space tradeoff problem. As the size of frequent itemset increases, computational time for the initial phases increases exponentially with increase in the requirement in memory space. So, a better way to consider only the relevant transaction or items can be possible field of research. If data cannot fit in the memory than more page faults may occur resulting in the decrease in the performance of the system.

## REFERENCES

[1] Rakesh Aggarwal, Tomasz Imielinski, Arun Swami, " Mining Association Rules between Sets of Items in Large Databases" ACM Sigmod Conference Washington DC, May (1993).

[2] Rakesh Aggarwal , Ramakrishanan Srikant, "Fast Algorithm for mining Association Rules", IBM Almaden Research Centre, Proceedings of 20th VLDB Conference, Santiago, Chile, (1994).

[3] Lin D., Z.M. Kedem, "Pincer-Search: An Efficient Algorithm for Discovering the Maximum Frequent Set", *IEEE Tran. Know. and Data Engg.*, Vol. **14**, No. 3, May/ June (2002), pp. 553-556.

[4] J. Han, J. Pei, and Y Yin, "Mining Frequent Patterns Without Candidate Generation", Proc. ACM SIGMOD (2000).

[5] C.H. Cai, A.W.C. Fu, C.H. Cheng, and W.W. Kwong, "Mining Association Rules with Weighted Items", Proc. IEEE Int'l Database Engg. And Applications Symp. (IDEAS '98), pp. 68-77, (1998).

[6] K.Sun and F.Bai,"Mining Weighted Association Rules Without Preassigned Weights", *IEEE Transactions on Knowledge and Data Engineering*, Vol. **20**, No. 4, April (2008), pages 489-495.

[7] Show-Jane Yen, Arbee L. P. Chen: "A Graph-Based Approach for Discovering Various Types of Association Rules". *IEEE Trans. Knowl. Data Eng.* **13**(5): 839-845 (2001).

[8] R. Srikant and R. Agrawal In, "Mining Generalized Association Rules", Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, September (1995).

[9] K. Wang and M.-Y. Su, "Item Selection by "Hub-Authority" Profit Ranking, Proc. ACM SIGKDD, (2002).

[10] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, "Link Analysis Ranking : Algorithms, Theory, and Experiments", Panayiotis Tsaparas, *ACM Transactions on Internet Technology*, Vol. **5**, No. 1, February (2005), Pages 231-297.

[11] T.Y. Lin, Xiaohua Ho, Eric Louie "A Fast Association Rule Algorithm Based on Bitmap and Granular Computing", *IEEE International Conference on Fuzzy Systems*, (2003).

[12] The IBM Synthetic Data Generator, http://www.almaden.ibm.com/software/projects/iis/hdb/Projects/data_mining/datasets/syndata.html.

[13] Sotiris Kotsiantis, Dimitris Kanellopoulos, "Association Rules Mining: A Recent Overview", *GESTS International conference on computer science and engineering*, Vol. **32**(1) pp- 71-82, (2006).