FREQUENT DATA GENERATION USING RELATIVE DATA ANALYSIS

R.ARCHANA

M.S (Software Engineering)
VIT University
Vellore-632014
India

N.MANIKANDAN

Assistant Professor (Senior)
School of Information Technology & Engineering
VIT University
Vellore-632014
India

ABSTRACT : Traditional association rule mining method mines association rules only for the items bought by the customer. However an actual transaction consists of the items bought by the customer along with the quantity of items bought. This paper reconsiders the traditional database by taking into account both items as well as its quantity. This new transaction database is named as bag database and each transaction consists of item along with its quantity (called itembag). This paper proposes algorithms for mining frequent items as well as rare items from the bag database. The method for mining frequent items from the database makes use of fuzzy functions to avoid sharp boundaries between itemsets and the method for mining rare items makes use of relative support to discover rare data that appear infrequently in the database but are highly associated with specific data.

Keywords: BAG DATABASE, MINING, ITEMBAG, ASSOCIATION RULE, FUZZY FUNCTION

1. Introduction

Data mining is a technique to analyze the stored data in large databases to discover potential information and knowledge. Association rule mining is an important technique in data mining which is used to discover useful associations among items in a database. Association rules can be discovered for both frequent data that occur repeatedly as well as rare data that appear infrequently but highly associated with specific data.

The project focuses on mining association rules from a database to identify associations among the frequent item sets and associations among the rare item sets. This project reconsiders the traditional transaction database by assuming that each transaction consists of a set of items as well as their quantities. The problem is concerned with mining frequent as well as rare data from the database using fuzzy functions and relative support.

Association Rule
It is a technique to investigate the possibility of simultaneous occurrence of the data. A => B [confidence=0.5]. The above rule states that – If a customer buys A then the customer also buys B with a probability of 50%. Each association rule consists of two parts namely one is Antecedent (A) and the other is Consequent (B). For example, if 70% of customers who buy soap also buy shampoo.

Association Rule Terms

SUPPORT(X) = Number of transactions containing itemset X.

CONFIDENCE(X=>Y) =
Number of transactions containing itemset (X U Y) / Number of transactions containing itemset X

2. Literature Review

[1] In this paper, Ramakrishnan Srikant and Rakesh Agarwal propose two new algorithms, Cumulate and EstMerge. Empirical evaluation showed that these two algorithms run 2 to 5 times faster than Basic; for one real-life dataset, the performance gap was more than 100 times. To solve the problem of generation of many uninteresting or redundant rules a new interest measure that uses the taxonomy information to prune redundant rules was developed. The measure is based on the intuition that if the support and confidence of a rule are close to their expected values based on an ancestor of the rule, the rule can be considered redundant.

[2] In this paper Yuh-Jiuan Tsay and Jiunn-Yann Chiang provide the characteristics of CBAR (Cluster Based Association Rule). It only requires a single scan of the database, followed by contrasts with the partial cluster tables. This not only prunes considerable amounts of data reducing the time needed to perform data scans and requiring less contrast, but also ensures the correctness of the mined results. By using the CBAR algorithm to

create cluster tables in advance, each CPU can be utilized to process a cluster table. Thus large itemsets can be immediately mined even when the database is very large.

[3] In this paper, Ling Zhou and Stephen Yau devise two new algorithms to generate association rules among rare items. New interesting measures are used to capture rules of strong interest. Matrix-based Scheme (MBS) and Hash-based Scheme (HBS) are proposed to explore interesting associations among infrequent items with memory-resident data structure and frequent items with bounded length. In both our schemes, only two passes over the database are needed.

[4] In this paper, Show-Jane Yen and Yue-Shi Lee introduce a data mining language. From the data mining language, users can specify the interested items or the sequences, and the minimum support and the minimum confidence threshold to discover association rules and sequential patterns. The authors propose the efficient data mining algorithms MIAR (Mining Interesting Association Rules) and MISP (Mining Interesting Sequential Pattern) to process the user requirements which can reduce the number of the combinations of itemsets or sequences in each customer sequence for counting the supports of the candidates, and reduce the number of the candidates according to the user's requests.

[5] In this paper, Yi-Chung Hua and Gwo-Hshiung Tzeng proposes a new fuzzy data mining technique consisting of two phases to find fuzzy if–then rules for classification problems: one to find frequent fuzzy grids by using a pre-specified simple fuzzy partition method to divide each quantitative attribute, and the other to generate fuzzy classification rules from frequent fuzzy grids. To improve the classification performance of the proposed method, they specially incorporate adaptive rules proposed by Nozaki *et al.*

3. Materials and Methods

a. Architecture Diagram:

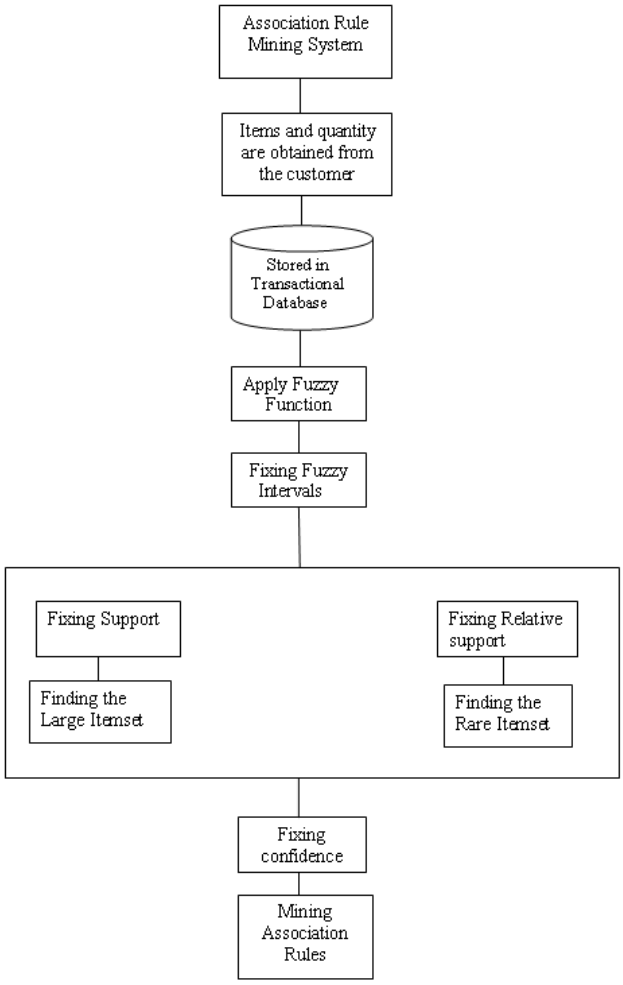

Fig 1. Architectural Design

b. Modular Description:

(i) Input Module:

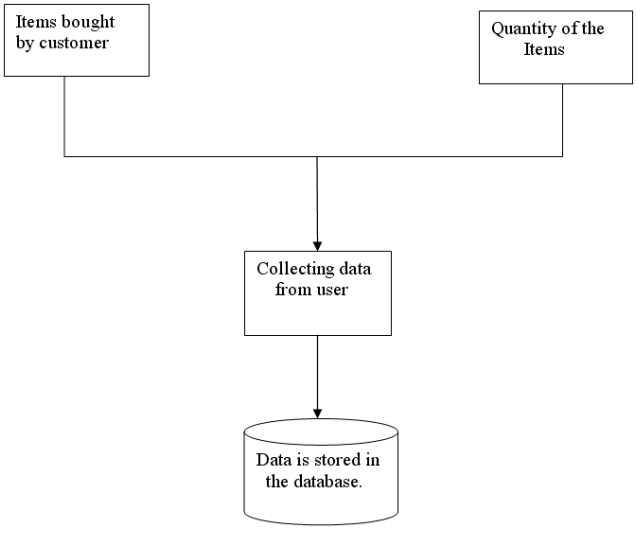

Fig 2. Input module

In this phase input is obtained from the user. This phase consists of the following inputs:
1. Customer ID
2. Items bought by the customer
3. Quantity of the items.

The customer id helps in identifying the customer with the items bought and the quantity of the items. The items bought by the customer are obtained along with the quantity of the items. This project takes into account the quantity of the item also which helps in identifying the relationship between items based on quantity.

(ii) Large itemset generation module

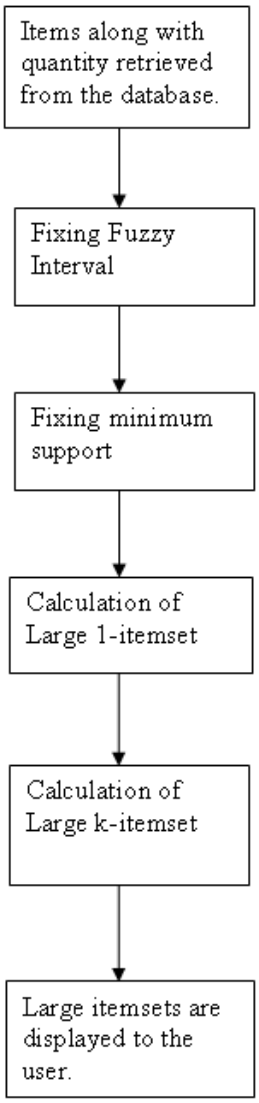

Fig 3. Large itemset generation module

In this module the large itemsets are calculated from the items bought by the customer based on the support. Initially the items bought by the customer are retrieved from the database. The minimum support is fixed. The support of the retrieved items is calculated based on the number of times the customers have bought the item. The support of an item can be calculated by counting the number of times the item has occurred in the transaction database.

The fuzzy function will return a value between 0 and 1, which denotes the fuzzy interval. Large 1-itemset is calculated by identifying the items having support greater than the minimum support. The next step is identifying combination of items having combined support more than the minimum support. This step is repeated until Large k-itemset (k>=1) is obtained such that there are no more combinations of items having support greater than minimum support. The obtained large itemsets are displayed to the user.

(iii) Rare itemset generation module:

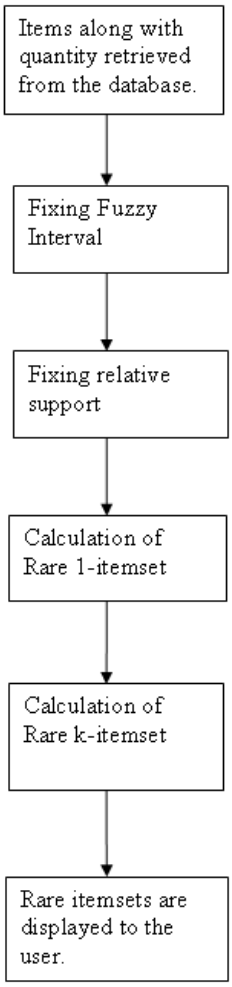

Fig 4. Rare itemset generation module

In this module the rare itemsets are calculated from the items bought by the customer based on the

relative support. Initially the items bought by the customer are retrieved from the database. The support of an item can be calculated by counting the number of times the item has occurred in the transaction database. The relative support for the rare items is fixed based on the support of the frequent items. Those items whose support is greater than the relative support but lesser than the support for frequent items is taken as the set of rare items. Rare 2-itemset is determined by combining the rare 1-itemset and pruning them based on the support. This process is repeated until rare k-itemset is obtained.

(iv) Mining association rule module:

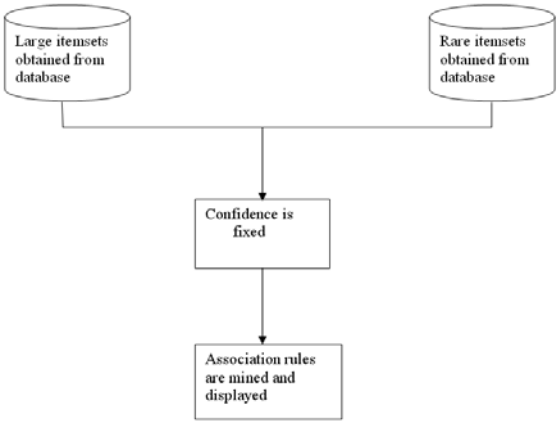

Fig 5 Mining association rule module

In this module, the association rules are mined. The large itemsets and the rare itemsets are retrieved from the database. A minimum confidence is fixed. If the confidence of the mined association rule is more than the fixed confidence, the mined association rule is displayed

4. Results and Discussion

The paper has some possible extensions. How to express the rules in higher level concepts, i.e., generalization of rules, is very important for practical purposes. Without generalization, the generated rules may be too detailed and are not fit for decision makers. Therefore, by including the concept hierarchies we may produce rules that are more abstract and concrete. Another possible extension is to prune less interesting rules that are trivial or implied by other rules. Without removing the uninteresting rules, we may be overwhelmed by enormous rules. Besides, we may try to generate only those rules that comply with some given meta-forms. This will help us to avoid lots of rules that do not fit our expectations or applications.

MQA-F Algorithm uses the fuzzy functions to eliminate sharp boundaries between the itemsets. Association of items along with their quantities is an excellent advantage of the above described system. It reduces the execution time which is very vital for its edge over other models. Moreover it is performance efficient in all the aspects. Mining association rules among items in a large database of sales transactions. It helps to identify associations among items with their quantities bought by customer. It also helps to increase sales by identifying which items are bought of which quantity by customer in association with each other. It increases the sales of associated items.

5. Conclusions:

Data mining is a challenging research area that aims to extract implicit, previously unknown and potentially useful information from databases. Mining association rules is one of the most important approaches proposed to extract the data from customer databases.

In this paper a new technique has been used to mine association rules from transaction databases. Two important studies have been undertaken – First is until now, only the items bought by the customer have been taken into consideration. In this paper, the items along with their quantities are taken. Second is large itemsets are generated for both frequent and rare items using fuzzy based algorithm. A new attribute termed relative support is used as the criterion for identifying the rare items in the database. Thus this project focuses on a new data mining technique to mine frequent and rare items from a transaction database containing both items bought by the customers as well as their quantities.

Acknowledgements

We would like to thank the management of VIT University, Vellore, India for allowing us to carry out our work on their premises. Our sincere thanks go to our school dean, fellow faculty and students who directly and indirectly contributed for the betterment of our work.

References:

[1] Mining generalized association rules Ramakrishnan Srikant , Rakesh Agrawal *IBM Almaden Research Centec 650 Harry Road, San Jose, CA 95120, USA* Received 1 October 1996; accepted 1 April 1997.

[2] CBAR: an efficient method for mining association rules Yuh-Jiuan Tsay, Jiunn-Yann Chiang Department of Management Information Systems, National Ping-Tung University of Science and Technology, Ping-Tung 912, Taiwan, ROC Received 20 October 2002; accepted 6 April 2004 Available online 13 November 2004

[3] Efficient association rule mining among both frequent and infrequent items Ling Zhou, Stephen Yau_Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, IL, United States Received 12 December 2006; received in revised form 26 January 2007; accepted 20 February 2007.

[4] An efficient data mining approach for discovering interesting knowledge from customer transactions Show-Jane Yen, Yue-Shi Lee Department of Computer Science and Information Engineering, Ming Chuan University, Taipei, Taiwan, ROC.

[5] Elicitation of classification rules by fuzzy data mining Yi-Chung Hua, Gwo-Hshiung Tzengb Department of Business Administration, Chung Yuan Christian University, Chung-Li 320, Taiwan, ROC, b Institute of Management of Technology, National Chiao Tung University, Hsinchu 300, Taiwan, ROC.