Global Journal of Computer Science and Technology
Volume 11 Issue 11 Version 1.0 July 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Survey on Data Mining Algorithm for Market Basket Analysis

By Dr. M. Dhanabhakyam , Dr. M. Punithavalli

Dr. SNS College of Arts and Science

Abstracts - Association rule mining identifies the remarkable association or relationship between a large set of data items. With huge quantity of data constantly being obtained and stored in databases, several industries are becoming concerned in mining association rules from their databases. For example, the detection of interesting association relationships between large quantities of business transaction data can assist in catalog design, cross-marketing, lossleader analysis, and various business decision making processes. A typical example of association rule mining is market basket analysis. This method examines customer buying patterns by identifying associations among various items that customers place in their shopping baskets. The identification of such associations can assist retailers expand marketing strategies by gaining insight into which items are frequently purchased jointly by customers. It is helpful to examine the customer purchasing behavior and assists in increasing the sales and conserve inventory by focusing on the point of sale transaction data. This work acts as a broad area for the researchers to develop a better data mining algorithm. This paper presents a survey about the existing data mining algorithm for market basket analysis.

Keyewords : *Association Rule Mining, Apriori Algorithm, Market Basket Analysis*

GJCST Classification : *H.2.8*

A Survey on Data Mining Algorithm for Market Basket Analysis

Dr. M. Dhanabhakyam[α], Dr. M. Punithavalli[Ω]

Abstract - Association rule mining identifies the remarkable association or relationship between a large set of data items. With huge quantity of data constantly being obtained and stored in databases, several industries are becoming concerned in mining association rules from their databases. For example, the detection of interesting association relationships between large quantities of business transaction data can assist in catalog design, cross-marketing, lossleader analysis, and various business decision making processes. A typical example of association rule mining is market basket analysis. This method examines customer buying patterns by identifying associations among various items that customers place in their shopping baskets. The identification of such associations can assist retailers expand marketing strategies by gaining insight into which items are frequently purchased jointly by customers. It is helpful to examine the customer purchasing behavior and assists in increasing the sales and conserve inventory by focusing on the point of sale transaction data. This work acts as a broad area for the researchers to develop a better data mining algorithm. This paper presents a survey about the existing data mining algorithm for market basket analysis.

Keywords : Association Rule Mining, Apriori Algorithm, Market Basket Analysis

I. Introduction

The majority of the recognized organizations have accumulated masses of information from their customers for decades. With the e-commerce applications growing quickly, the organizations will have a vast quantity of data in months not in years. Data Mining, also called as Knowledge Discovery in Databases (KDD), is to determine trends, patterns, correlations, anomalies in these databases that can assist to create precise future decisions.

Mining Association Rules is one of the most important application fields of Data Mining. Provided a set of customer transactions on items, the main intention is to determine correlations among the sales of items. Mining association rules, also known as market basket analysis, is one of the application fields of Data Mining. Think a market with a gathering of large amount of customer transactions. An association rule is $X \Rightarrow Y$, where X is referred as the antecedent and Y is referred as the consequent. X and Y are sets of items and the rule represents that customers who purchase X are probable to purchase Y with probability %c where c is known as the confidence. Such a rule may be: “Eighty percent of people who purchase cigarettes also purchase matches". Such rules assists to respond questions of the variety “What is Coca Cola sold with?" or if the users are intended in checking the dependency among two items A and B it is required to determine rules that have A in the consequent and B in the antecedent.

Figure1 shows a typical Market basket analysis. This is a perfect example for illustrating association rule mining. It is a fact that all the managers in any kind of shop or departmental stores would like to gain knowledge about the buying behavior of every customers. This market basket analysis system will help the managers to understand about the sets of items are customers likely to purchase. This analysis may be carried out on all the retail stores data of customer transactions. These results will guide them to plan marketing or advertising approach. For example, market basket analysis will also help managers to propose new way of arrangement in store layouts. Based on this analysis, items that are regularly purchased together can be placed in close proximity with the purpose of further promote the sale of such items together. If consumers who purchase computers also likely to purchase anti-virus software at the same time, then placing the hardware display close to the software display will help to enhance the sales of both of these items.

Classification rule mining intends to determine a small set of regulations in the database that forms a perfect classifier. Association rule mining discovers all the rules offered in the database that assures some minimum support and minimum confidence constraint. In the case of association rule mining, the goal of discovery is not pre-determined, while for classification rule mining there is only one predetermined goal. This paper provides various existing data mining algorithms for market basket analysis.

Author [α] : Research Scholar
Author [Ω] : Director and Guide , MCA Department in Dr. SNS College of Arts and Science ,Coimbatore.

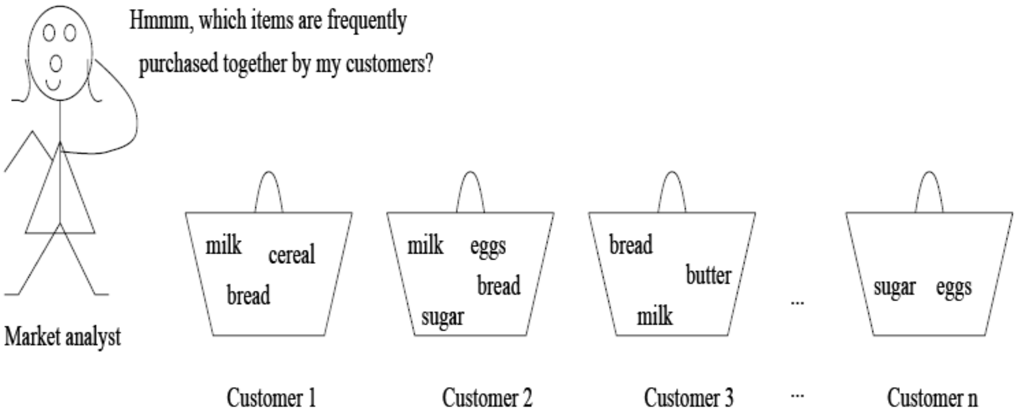

Figure1: Market Basket Analysis

II. Literature Survey

Zhixin *et al.,* [1] recommended an improved classification technique based on predictive association rules. Classification dependent predictive association rules (CPAR) is a type of association classification methods which unites the benefits of associative classification and conventional rule-based classification. For generation of the rule, CPAR is highly effective when compared to the conventional rule-based classification because most of the repeated calculation is ignored and multiple literals can be chosen to produce multiple rules at the same time. Even though the benefit mentioned above avoids the repeated calculation in rule generation, the prediction processes have the disadvantage in class rule distribution inconsistency and interruption of inaccurate class rules. Further, it is ineffective to instances satisfying no rules. To ignore these difficulties, the author recommends Class Weighting Adjustment, Center Vector-based Pre-classification and Post-processing with Support Vector Machine (SVM). Experimental observations on Chinese text categorization corpus TanCorp proves that this approach gains an average enhancement of 5.91% on F1 score compared with CPAR.

Qiang *et al.,* [2] presented association classification based method on compactness of rules. Associative classification provides maximum classification accurateness and strong flexibility. On the other hand, this associative classification suffers from a difficulty of over fitting because the classification rules satisfied least support and lowest confidence are returned as strong association rules return to the classifier. In this paper, proposed an innovative association classification technique based on neatness of rules, it extends Apriori Algorithm which considers the interestingness, importance, overlapping relationships among rules. Experimental observation proves that the proposed approach has better classification accuracy in comparison with CBA and CMAR are highly intelligible.

Wang *et al.,* [3] suggested a novel rule weighting approach in classification association rule mining. Classification association rule mining (CARM) is a latest classification rule mining technique that constructs an association rule mining based classifier by utilizing classification association rules (CARs). The specific CARM algorithm which is used is not considered, a similar set of CARs is constantly produced from data, and a classifier is generally presented as a structured CAR list, depending on a selected rule ordering approach. Many numbers of rules ordering approaches have been established in the recent past, which can be classified as rule weighting, support-confidence and hybrid. In this approach an alternative rule-weighting method, called CISRW (Class-Item Score based Rule Weighting) and build up a rule-weighting based rule ordering mechanism depending on CISRW. Later, two hybrid techniques are additional developed by merging (1) and CISRW. The simulation results indicates that the three proposed CISRW based/related rule ordering techniques do well by means of accuracy of classification.

Bartik [4] presented association based classification for relational data and its use in web mining. Classification according to the mining association rules is a technique with better accuracy and human understandable classification scheme. The intention of the author is to put forward a alteration of the fundamental association based classification technique that can be helpful in data gathering from Web pages. In this paper, the alteration of the technique and required discretization of numeric characteristics are provided.

Sumithra *et al.,* [5] proposed a distributed Apriori association rule and classical Apriori mining algorithms for grid based knowledge discovery. The intention of this paper is to obtain knowledge with the help of predictive apriori and distributed grid dependent apriori algorithms for association rule mining. The author provides the implementation of an association rules discovery data mining task with the help of Grid technologies. A consequence of implementation with a contrast of existing apriori and distributed apriori is also provided by the author. Distributed data mining systems offers an effective utilization of multiple processors and databases to accelerate the execution of data mining and facilitate data distribution. For evaluating the effectiveness of the described technique, performance investigation of apriori and predictive apriori techniques on a standard database have been provided using weka tool. The key intention of grid computing is to offer the organizations and application builders the capability to generate distributed computing environments that can make use of computing resources on demand. Hence, it can assist amplify the effectiveness and decrease the cost of computing networks by reducing the time for data processing and optimizing resources and distributing workloads, thus permitting users to attain much faster outcome on large operations and at lesser costs.

Trnka [6] uses Data Mining Methods for Market Basket Analysis. This paper provides the technique for Market Basket Analysis implementation to Six Sigma technique. Data mining techniques offers more prospects in the market sector. Basket Market Analysis is one among them. Six Sigma technique utilizes various statistical techniques. With execution of Market Basket Analysis to Six Sigma, the results can be enhanced and Sigma performance level of the method can also be modified. The author used GRI (General Rule Induction) technique to construct association rules among products in the market basket. These associations provide a variety among the products. Web plot is utilized here for representing the dependence among the products.

Mining Interesting Rules by Association and Classification Algorithm is put forth by Yanthy *et al.,* [7]. The main intention in data mining is to disclose hidden knowledge from data and several techniques have been suggested so far. But the demerit is that characteristically not all rules are interesting - only little portions of the created rules would be of interest to several provided user. Therefore, numerous measures like confidence, support, lift, information gain, etc., have been suggested to find the best or highly interesting rules. On the other hand, some techniques are good at creating rules high in one interestingness measure but not good in other interestingness measures. The relationship among the techniques and interestingness measures of the created rules is not clear until now. The author studied the relationship among the techniques and interesting measures. The author used synthetic data so that the outcome result is not restricted to particular situations.

Market Basket Analysis Based on Text Segmentation and Association Rule Mining is suggested by Xie *et al.,* [8]. Market basket analysis is very useful in offering scientific decision support for trade market by mining association rules between items people purchased collectively. The author provides an innovative market basket analysis technique by mining association rules on the items' internal features that are obtained with the help of automatic words segmentation technique. This technique has been used for dynamic dishes recommend system and results better in the experimental results.

Chiu *et al.*, [9] proposed a market-basket analysis with principal component. Market-basket analysis is a well-known business crisis that can be solved computationally with the help of association rules, mined from transaction data to reduce the cross-selling results. The author model the market-basket analysis as a finite mixture density of human consumption activities based on social and cultural activities. This results in the usage of principle component analysis and perhaps mixture density analysis of transaction data that was not obvious previously. The author contrast PCA and association rules mined from a set of benchmark transaction data, to discover common and differences among these two data exploration tools.

Market Basket Analysis of Library Circulation Data is provided by Cunningham *et al.,* [10]. Market basket analysis technique have lately seen extensive usage in evaluating consumer purchasing patterns - particularly, in identifying products that are often purchased. The author utilized the a-priori market basket tool to the process of detecting subject classification grouping that co-occur in transaction records of books borrowed from a university library. This data can be utilized in directing users to extra portions of the gathered that may consists of documents that are related to their information requirement, and in finding a library's physical layout. These results can also offer insight into the amount of scatter that the classification method provokes in a particular gathering of documents.

Zongyao *et al.,* [11] proposed a mining local association patterns from spatial dataset. The author provides a model and algorithm to mine local association rules from available spatial dataset while completely considering the reality that spatial heterogeneity may extensively available in realism. The important element of the model is the computation Localized Measure of Association Strength (LMAS) which is utilized to measure local association patterns. Spatial association relations are exclusively defined as spatial relations that are modeled by DE-9IM model. The author presents mining technique for determining local association patterns from spatial dataset. The proposed technique mines reference and target objects that have possible association patterns and processes LMAS for every object in the reference objects for some interested spatial relation. Hence, the result of the algorithm is a LMAS distribution map that replicates association potential variations over the examination area. Spatial interpolation for LMAS is recommended to generate a continuous LMAS distribution that can be utilized to investigate hot spots that reveal strong association patterns. This proposed technique was applied in an ecological system research.

Mining association rules based on apriori algorithm and application is given by Pei *et al.,* [12]. In the data mining research, mining association rules is an significant subject. Intended at two difficulties of discovering frequent itemsets in a large database and mining association rules from frequent itemsets, the author carries some analysis on mining frequent itemsets algorithm with the help of apriori algorithm and mining association rules algorithm with the help of enhanced measure system. Mining association rules technuque with the help of support, confidence and interestingness is enhanced, aiming at generating interestingness ineffective rules and losing helpful rules. Useless rules are cancelled, creating more reasonable association rules including negative items. The suggested technique is utilized to mine association rules to the 2002 student score list of computer dedicated field in Inner Mongolia university of science and technology.

Yong *et al.,* [13] proposed a mining association rules with new measure criteria. In recent days, association rules mining from bulk databases is an active research field of data mining motivated by many application areas. But, there are some difficulties in the strong association rules mining depending on support-confidence framework. Initially, there are a huge number of redundant association rules are created, then it is complicated for user to discover the interesting ones. Then, the correlation among the features of specified application areas is avoided. Therefore innovative measure criteria called Chi-Squared test and cover should be introduced to association rules mining, and the more important aspect is the use of Chi-Squared test to reduce the amount of rules. The Chi-Squared test and cover of measures are utilized by author for association rules mining for the purpose of eliminating the itemsets that are statistic free, while frequent itemsets or rules are created. Therefore the number of patterns itemsets reduced and it is effortless for user to gather the highly noticeable association rules. The simulation results suggest that the Chi-Squared test is efficient on decreasing the quantity of patterns through merging support and cover constrain. Pattern choosing according to Chi-Squared test can remove few irrelevant attributes and the efficiency and veracity of mining association rules are enhanced.

Mining traditional association rules using frequent itemsets lattice is given by Vo *et al.*, [14]. Numerous methods have been formulated for the enhancement of time in mining frequent itemsets. However, the methods which deal with the time of mining association rules were not put in deep research. In reality, in case of database which contains many frequent itemsets (from ten thousands up to millions), the time of mining association rules is much larger than that needed for mining frequent itemsets. In this paper, developed an application of lattice in mining conventional association rules which will significantly decrease the time for mining rules. This technique comprises of two stages: (a) construction of frequent itemsets lattice and (b) mining association rules from lattice. Fort the quick determination of association rules, the parent-child relationships in lattice is used. The experiments observation proves that the mining rule from lattice is more efficient than the straight mining from frequent itemsets by means of hash table.

Rastogi *et al.,* [15] presents mining optimized association rules with categorical and numeric attributes. Mining association rules on bulky data sets has gained significant attention recently. Association rules are helpful for predicting correlations among the features of a relation and contain applications in marketing and many retail sectors. Moreover, optimized association rules are an efficient approach to focus on the most interesting features linking certain attributes. Optimized association rules are allowed to include uninstantiated attributes and the difficulty is to find out instantiations such a way that either the support or confidence of the rule is maximized. In this approach, the optimized association rules difficulty is simplified in three ways: (a) association rules are permitted to include disjunctions in excess of uninstantiated features, (b) association rules are allowed to contain a random number of uninstantiated features, and (c) uninstantiated features can be either categorical or numeric. This generalized association rules permits to mine more helpful information about seasonal and local patterns linking multiple features. This paper also suggests an efficient method for pruning the search space when calculating optimized association rules for both categorical and numeric features. Experimental result shows that pruning techniques are effective for a huge number of uninstantiated features, disjunctions, and values in the domain of the features.

Wang et al., [16] performs an investigation on association rules mining based-on ontology in e-commerce. Commercial actions carried out with the use of Internet turn out to be more and more popular. And plenty of transaction logs are created, which helps to gather useful information by data mining. In this manner, association rule mining is very important in e-commerce. However there are various problems occur in the existing association rules mining systems. The existing conventional techniques can't solve these problems very well. With the intention of solving these difficulties better, this paper proposes association rules mining depending on ontology. Generally researches the specified three parts during data mining: (1) methods of ontology creation and principles of commodity categorization; (2) simplifying R-interesting based on actual situations; (3) implementing association rules mining depending on ontology by improved Apriori. Additionally, this paper tests the enhanced algorithm using FoodMart2000, Java as the development language and Jena as the ontology engine, completes the entire process of mining, and confirms the validity of the algorithm by the example of the database.

III. Problems and Directions

The various existing data mining algorithm for market basket analysis are discussed in this paper. All the techniques have its own advantages and disadvantages. This section provides some of the drawbacks of the existing algorithms and the techniques to overcome those difficulties.

Among the methods discussed for data mining, apriori algorithm is found to be better for association rule mining. Still there are various difficulties faced by apriori algorithm. The various difficulties faced by apriori algorithm are

- It scans the database lot of times. Every time the additional choices will be created during the scan process. This creates the additional work for the database to search. Therefore database must store huge number of data services. This results in lack of memory to store those additional data. Also, the I/O load is not sufficient and it takes very long time for processing. This results in very low efficiency.
- Frequent item in the larger set length of the circumstances, leads to significant increase in computing time.
- Algorithm to narrow face. At this situation, the algorithm will not result in better result. Therefore it is required to improve, or even need to re-design of algorithms.

Those drawbacks can be overcome by modifying the apriori algorithm effectively. The time complexity for the execution of apriori algorithm can be solved by using the fast apriori algorithm. This has the possibility of leading to lack of accuracy in determining the association rule. To overcome this, the fuzzy logic can be combined with the apriori algorithm. This will help in better selection of association rules for market basket analysis.

IV. Conclusion

Information is collected almost everywhere in our everyday lives. This leads to the huge increase in the amount of data available. Physical analysis of these huge amount of information stored in modern databases is very difficult. Data mining provides tools to reveal unknown information in large databases which are stored already. A well-known data mining technique is association rule mining. Association rule mining and classification technique to find the related information in large databases is becoming very important in the current scenario. Association rules are very efficient in revealing all the interesting relationships in a relatively large database with huge amount of data. The large quantity of information collected through the set of association rules can be used not only for illustrating the relationships in the database, but also used for differentiating between different kinds of classes in a database. This paper provides some of the existing data mining algorithms for market basket analysis. The analysis of existing algorithms suggests that the usage of association rule mining algorithms for market basket analysis will help in better classification of the huge amount of data. The apriori algorithm can be modified effectively to reduce the time complexity and enhance the accuracy.

References Références Referencias

1. Zhixin Hao, Xuan Wang, Lin Yao, Yaoyun Zhang, "Improved Classification based on Predictive Association Rules," SMC 2009, IEEE International Conference on Systems, Man and Cybernetics, Pp. 1165 – 1170, 2009.
2. Qiang Niu, Shi-Xiong Xia, Lei Zhang, "Association Classification Based on Compactness of Rules," WKDD 2009, Second International Workshop on Knowledge Discovery and Data Mining, Pp. 245 – 247, 2009.
3. Y.J. Wang, Qin Xin, F. Coenen, "A Novel Rule Weighting Approach in Classification Association Rule Mining," ICDM Workshops 2007, Seventh IEEE International Conference on Data Mining Workshops, Pp. 271 – 276, 2007.
4. V. Bartik, "Association based Classification for Relational Data and its Use in Web Mining," CIDM '09, IEEE Symposium on Computational Intelligence and Data Mining, Pp. 252 – 258, 2009.
5. R. Sumithra, S. Paul, "Using Distributed Apriori Association Rule and Classical Apriori Mining Algorithms for Grid Based Knowledge Discovery," International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1 – 5, 2010.

6. Trnka, A.,"Market Basket Analysis with Data Mining Methods", International Conference on Networking and Information Technology (ICNIT), Pp. 446 - 450, 2010.
7. W. Yanthy, T. Sekiya, K. Yamaguchi, "Mining Interesting Rules by Association and Classification Algorithms," FCST '09. Fourth International Conference on Frontier of Computer Science and Technology, Pp. 177 – 182, 2009.
8. Xie Wen-xiu, Qi Heng-nian and Huang Mei-li, "Market Basket Analysis Based on Text Segmentation and Association Rule Mining", First International Conference on Networking and Distributed Computing (ICNDC), Pp. 309 – 313, 2010.
9. Chiu, K.S.Y., Luk, R.W.P., Chan, K.C.C. and Chung, K.F.L., "Market-basket Analysis with Principal Component Analysis: An Exploration", IEEE International Conference on Systems, Man and Cybernetics, Vol. 3, 2002.
10. Cunningham, S.J. and Frank, E., "Market Basket Analysis of Library Circulation Data", International Conference on Neural Information Processing, Vol. 2, Pp. 825-830, 1999.
11. Zongyao Sha, Xiaolei Li, "Mining local association patterns from spatial dataset," Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Vol. 3, pp. 1455 – 1460, 2010.
12. Pei-ji Wang, Lin Shi, Jin-niu Bai and Yu-lin Zhao, "Mining Association Rules Based on Apriori Algorithm and Application", International Forum on Computer Science-Technology and Applications, Vol. 1, Pp. 141-143, 2009.
13. Yong Xu, Sen-Xin Zhou and Jin-Hua Gong, "Mining Association Rules with New Measure Criteria", International Conference on Machine Learning and Cybernetics, Vol. 4, Pp. 2257-2260, 2005.
14. Vo, B. and Le, B., "Mining traditional association rules using frequent itemsets lattice", International Conference on Computers & Industrial Engineering, Pp. 1401 – 1406, 2009.
15. Rastogi, R. and Kyuseok Shim, "Mining optimized association rules with categorical and numeric attributes", IEEE Transactions on Knowledge and Data Engineering, Vol. 14 , No. 1, 2002.
16. Wang Xuping, Ni Zijian and Cao Haiyan, "Research on Association Rules Mining Based-On Ontology in E-Commerce", International Conference on Wireless Communications, Networking and Mobile Computing, Pp. 3549-3552, 2007.