International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 7– No.10, October 2014 – www.ijais.org

Data Mining in Market Basket Transaction: An Association Rule Mining Approach

S.O. Abdulsalam
Department of Computer, Library and Information Science,
Kwara State University,
Malete, Nigeria

K.S. Adewole
Department of Computer Science University of Ilorin,
Ilorin, Nigeria

A.G. Akintola
Department of Computer Science University of Ilorin,
Ilorin, Nigeria

M.A. Hambali
Department of Computer Science,
Federal University,
Wukari, Nigeria

ABSTRACT
Data is one of the valuable resources for organization, and database management systems are gradually becoming ubiquitous in many small and medium scale companies. Although, some of the benefits of database management systems have been explored, however, many companies have not been able to exploit the advantages of gaining business intelligence from their databases. This has led to inadequate business decision making based on the data contained in the databases.

In this paper, association rules mining also known as market basket analysis using Apriori algorithm is presented for extracting valuable knowledge embedded in the database of a supermarket. Data representing six (6) distinct products across thirty (30) unique transactions were generated from a well-structured transactional database representing the sales pattern of a supermarket store. The frequencies of purchasing these products were extracted for the above data and different association rules were deduced. It was established from these rules that purchase of one product would invariably lead to the purchase of another product as evident in the association between Apple and Chocolate. The discovered relationship will guide companies in planning marketing and advertising strategies that will help them outshine their competitors.

General Terms
Data Mining, Database, Association Rule Mining

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

1. INTRODUCTION
It has been observed over time that customer buying pattern in a shopping trip can lead to multiple products usually purchase from multiple categories. Efforts have been made by researchers to analyze the multi-category purchase information in order to plan marketing activities accordingly so as to maximize profit. For instance, a retailer can reduce the price of cake mix to sell more cake mix and frostings [1], as a result of which the overall profit may improve. Despite the fact that companies generate huge amount of data on daily basis, decision makers in organization rarely utilize data mining techniques to uncover interesting patterns from the vast amount of data in their data repositories.

Research efforts using market basket data focused on identifying what products are purchased together frequently. Given the number of products carried in a typical retailing store, this trivial question can be computationally challenging. For example, in a store carrying a thousand products, to identify what two products have been purchased together frequently, one needs to check the frequency of roughly half a million pairs of products. There is need to carry out this analysis within a reasonable short period of time since the retailer needs to adjust promotion and pricing decisions accordingly. A retailer can then decide as to which products to promote and at what levels based on these rules. Early research in association rule mining was based on frequency measures, but more recent research has started examining associations using other measures such as Chi-square, lift, etc. Marketing researchers have long realized that products or categories can be complements or substitutes [2]. Promotions in one category can improve not only its own sales but also sales of its complements, while suppressing sales of its substitutes. Sometimes, products may also appear in a basket together coincidentally without necessarily being complements, since a consumer may decide to purchase several products in one trip just in order to minimize the shopping cost. During the late 1990s, multivariate models have been developed to address the question of cross-category effects. These utility theory based models usually decompose the utility of a category into its own effects and cross-category effects. Some models go further to isolate cross category marketing mix effects from cross-category co-incidence effect. An understanding of these effects can help retailers improve their profits. For example, if the co-occurrence is due to complementarities rather than just co-incidence, a retailer will be able to co-ordinate the marketing mix in order to maximize profits. Compared to the data mining approach, these multivariate models provide more precise measures of cross-category effects, even though at a high computational cost since they typically involve Markov Chain Monte Carlo (MCMC) models [2, 3].

According to [4], Data Mining refers to mining knowledge from huge amounts of data. Data mining techniques are used to operate on huge amount of data to discover hidden patterns and relationships helpful for decision making. Data mining is

defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [5]. Data mining techniques include association, clustering, classification and prediction. Association rule mining was widely used as an exploratory tool in market basket data analysis. Association rule is one of the most popular and well researched methods used in data mining, especially as it regards supermarket store. Implementing association rule using standard data mining tools can reveal very important pattern which can be a competitive advantage to retail markets both in large and small scale. This paper focus on practical application of association rule mining using Apriori algorithm so as to determine which items are purchased together frequently by customers in a supermarket.

2. RELATED WORK

Data mining has provided a lot of opportunities to mine customer purchasing patterns and uncover hidden knowledge from data. Researchers have explored rule extraction using association analysis. Oladipupo and Oyelade [6] identify student's failure patterns using association mining. In their research, 30 courses for 100 level and 200 level students were considered to discover patterns of failed courses. The discovered patterns can assist academic planners in making constructive recommendation, curriculum structure and modification in order to improve students' performances. Agarwal et al., [7] explored application of Apriori algorithm in grocery store. Tissera et al., [8] presented a real-world experiment conducted in an ICT educational institute in Sri Lanka for analyzing students' performance. In their research, they applied a series of data mining task to find relationships between subjects in the undergraduate syllabi. They used association rules to identify possible related two subjects' combination in the syllabi, and apply correlation coefficient to determine the strength of the relationships of subject combinations identified by association rules. Researchers have also carried out survey of data mining algorithms in market basket analysis [9, 10]. The use of association based classification for relational data in web environment was presented in [11]. 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. Sumithra and Paul [12] presented a distributed Apriori association rule mining and classical Apriori mining algorithms for grid-based knowledge discovery. Qiang et al., [13] proposed association classification based method on compactness of rules. The proposed approach 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.

3. METHODOLOGY

3.1 Association Rule Analysis

Association rule mining associates one or more attributes of a dataset with another attributes, to discover hidden and significant relationship between the attributes, producing an if-then statement concerning attribute values in the form of rules [14, 15]. An association rule states that if we pick a customer at random and find out that he/she selected some item (that is, bought some products), we can be assured by indicate the quantity by a percentage that he/she also bought some other products. Based on the concept of strong rules, [16] introduced association rules for discovering regularities between products in large scale transaction data recorded by Point-of-Sale (POS) systems in supermarkets. For example, the rule {Onions, Potatoes}= {Burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as promotional pricing or product placements. In addition to the above example from market basket analysis, association rules are employed today in many application areas including Web usage mining, intrusion detection and bio-informatics [17].

According to [6], an association rule is an implication expression of the form X => Y, where $X \subset I$, and $Y \subset I$, and X and Y are disjoint itemsets, i.e $X \cap Y = \phi$. The strength of an association rule can be measured in terms of its support and confidence. The rule X => Y holds in the transaction set D with confidence *c* and support *s*, if c% of the transactions in D that contains X also contains Y, and s% of transactions in D contains $X \cup Y$. Both the antecedent and the consequent of the rule could have more than one Item. The formal definitions of these two metrics are:

$$\text{Support, } s\ (X \Rightarrow Y) = \Sigma(X \cup Y)/N \quad (1)$$

$$\text{Confidence, } c\ (X \Rightarrow Y) = \Sigma(X \cup Y)/\Sigma X \quad (2)$$

Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of literals called items and D be a set of transactions where each transaction T is a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier, called its TID. We say that a transaction T contains X, a set of some items in I, if $X \subseteq T$ [4].

Association rule mining process could be divided into two main phases to enhance the implementation of the algorithm.

The phases are:

1. Frequent Item Generation: This is to find all the itemsets that satisfy the minimum support threshold. The itemsets are called frequent itemsets.

2. Rule Generation: This is to extract all the high confidence rules from the frequent itemsets found in the first step. These rules are called strong rules.

According to [18] association analysis is based on the rule that specifies in the form: If item A is part of an event then X % of the time (confidence factor) item B is part of the same event. For instance:

- If a customer buys snacks, there may be 85% probability that the customer will also buy soft drinks or beer.
- If a person buys vacation airline tickets for an entire family, there may be 95% probability that he or she will rent a full-size car at the vacation location.

3.2 Market Basket Analysis

Let us assume that the supermarket sells the following products: Apple, Black Jam, Biscuit, Blue Band, Chocolate, Cream Soda. We consider only the five transactions and generalize these on the remaining 25 transactions considered in the study in order to simplify the concept.

Table 1: Transaction table

Transaction ID	Items
10	Apple, chocolate
20	Apple, Biscuit, Chocolate
30	Biscuit, Black Jam, Cream Soda
40	Cream Soda, Apple, Chocolate, Blue Band
50	Apple, Cream soda

Table 2: Representing transactions as a binary item list

TID	Apple	Chocolate	Biscuit	Black Jam	Cream Soda	Blue Band
10	1	1	0	0	0	0
20	1	1	1	0	0	0
30	0	0	1	1	1	0
40	1	1	0	0	1	1
50	1	0	0	0	1	0

Table 3: List of all itemsets and their frequencies

Itemsets	Frequency
Apple	4
Chocolate	3
Biscuit	2
Black Jam	1
Cream Soda	3
Blue Band	1
{Apple, Chocolate}	3
{Apple, Biscuit}	1
{Apple, Black Jam}	0
{Apple, Cream Soda}	1
{Apple, Blue Band}	1
{Apple, Chocolate, Biscuit}	1
{Apple, Chocolate, Black Jam}	0
{Apple, Chocolate, Cream Soda}	0
{Apple, Chocolate, Blue band}	1
{Apple, Biscuit, Black Jam}	0
{Apple, Biscuit, Cream Soda}	0
{Apple, Biscuit, Blue Band}	0
{Chocolate, Biscuit, Blue Band}	0
{Chocolate, Biscuit, Black Jam}	0
{Chocolate, Blue Band, Cream Soda}	0
{Cream Soda, Apple, Chocolate, Blue Band}	1

Let us assume that the minimum support of 50% is taken into consideration, we find the itemsets that occur in at least two transactions. Such itemsets are called frequent itemsets. The list of frequencies shows that all six items Apple, Chocolate, Biscuit, Black Jam, Blue Band and Cream Soda are frequent. The frequency goes down as we look at 2-itemsets, 3-itemsets, and 4-itemsets.

Table 4: The set of all frequent itemsets

Itemsets	Frequency
Apple	4
Chocolate	3
Biscuit	2
Cream Soda	3
{Apple, Chocolate}	3

We can now proceed to determine if the 2-itemset {Apple, Chocolate} lead to association rules with required confidence of 75%. Every 2-itemset {X, Y} can lead to two rules X => Y and Y=>X, if both satisfy the required confidence. As defined earlier, confidence of X => Y is given by the support for X and Y together divided by the support for X. We therefore have two possible rules and their confidence as follows:

Apple => Chocolate with confidence of ¾ = 75%

Chocolate => Apple with confidence of 3/3 = 100%

Therefore, both rules have confidence of minimum of 75% required and qualify. Rules that have more than the user-specified minimum confidence are called confident.

3.3 Apriori Algorithm

Apriori algorithm [16], is the most fundamental and important algorithm for mining frequent things. Apriori is used to find all frequent things in a given database, that is, it provides a way to find association rules on large scale. The significant of Apriori algorithm is to produce multiple passes over the database. It uses a repetitive approach called a breadth-first search (level-wise search) through the search room, where *K*-things are used to explain (K+1)-things. Apriori is designed to operate on databases containing transactions. For instance, collections of items bought by customers, or details of a website frequentation. Each transaction is seen as a set of items known as itemsets. Given a threshold *C*, the Apriori algorithm identifies the itemsets which are subsets of at least *C* transactions in the database [7].

Apriori uses a bottom up approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length *K* from itemsets of length *K-1*. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent *K*-length itemsets. After that, it scans the transaction database to determine frequent itemsets among the candidates.

The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties, as we shall see as following. Apriori employs an iterative approach known as a level-wise search, where *k*-itemsets are used to explore (*k+1*)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L_1. Next, L_1 is used to find L_2, the set of frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent *k*-itemsets can be found. The finding of each L_k requires one full scan of the

database. To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property is used to reduce the search space. This property states that "All nonempty subsets of a frequent itemset must also be frequent" [4]. Apriori algorithm is a two-step process of join and prune as follows:

1. The join step: To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted C_k. Let l_1 and l_2 be itemsets in L_{k-1}. The notation l_i [j] refers to the jth item in l_i (e.g., $l_1[k\text{-}2]$ refers to the second to the last item in l_1). By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. For the (k-1)-itemset, l_i, this means that the items are sorted such that $l_i[1] < l_i[2] < \ldots.. < l_i[k\text{-}1]$. The join, $L_{k-1} \times L_{k-1}$, is performed, where members of L_{k-1} are joinable if their first (k-2) items are in common. That is, members l_1 and l_2 of L_{k-1} are joined if $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \ldots \wedge (l_1[k\text{-}2] = l_2[k\text{-}2]) \wedge (l_1[k\text{-}1] < l_2[k\text{-}1])$. The condition $l_1[k\text{-}1] < l_2[k\text{-}1]$ simply ensures that no duplicates are generated. The resulting itemset formed by joining l_1 and l_2 is $l_1[1], l_1[2], \ldots, l_1[k\text{-}2], l_1[k\text{-}1], l_2[k\text{-}1]$.

2. The prune step: C_k is a superset of L_k, that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in C_k. A scan of the database to determine the count of each candidate in C_k would result in the determination of L_k (i.e., all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to L_k). C_k, however, can be huge, and so this could involve heavy computation.

To reduce the size of C_k, the Apriori property is used as follows. Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence, if any (k-1)-subset of a candidate k-itemset is not in L_{k-1}, then the candidate cannot be frequent either and so can be removed from C_k [4].

4. RESULTS AND DISCUSSION

The proposed data mining software was developed using Java programming language. Figure 1 shows the transaction menu to access the interface where different transactions can be made.

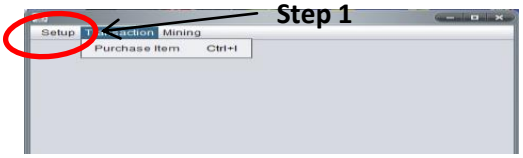

Figure 1: Transaction menu

Figure 2 is used by the system administrator to supply items bought into the database. The available items are displayed alongside their prices and quantities in the store. In this interface, step 3 indicated the number of purchased items by a customer. These items are added to the database by clicking on Process and Save buttons in Figure 3 and 4 respectively.

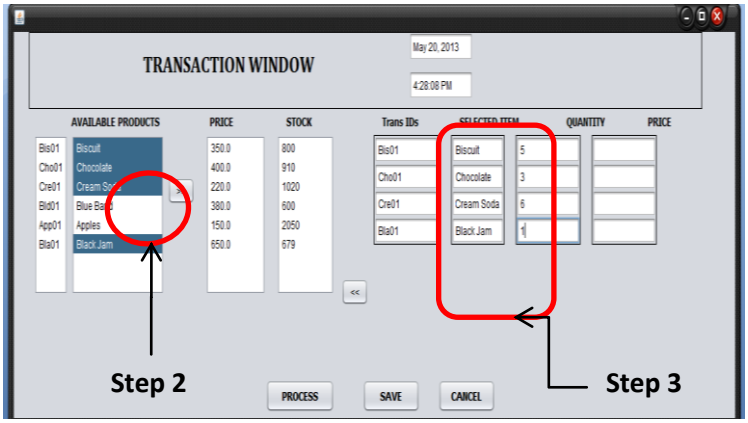

Figure 2: Transaction Interface

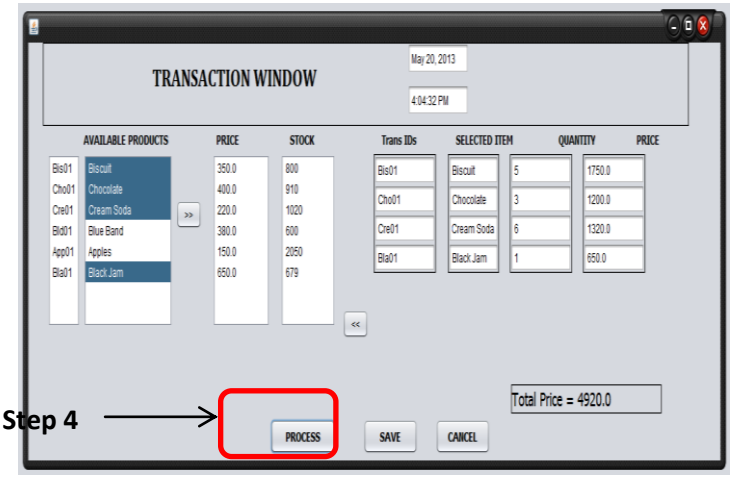

Figure 3: Transaction Interface after clicking on Process button

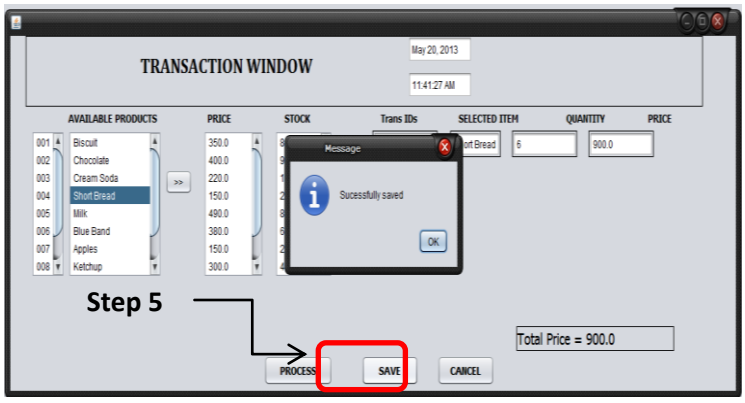

Figure 4: Transaction Interface after clicking on Save button

Figure 5 shows the menu to access the Association mining interface in Figure 6.

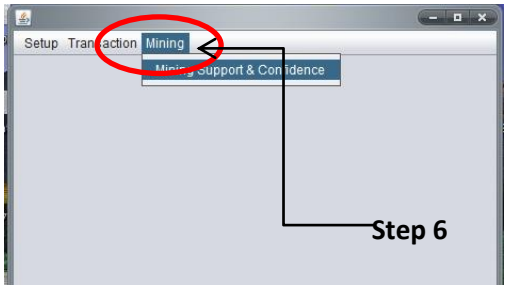

Figure 5: Menu for Association mining

Figure 6 shows the list of all itemsets and their frequencies, the controls for setting minimum support and minimum confidence, the Association rules and the generated strong rules using minimum support of 50 and minimum confidence of 75.

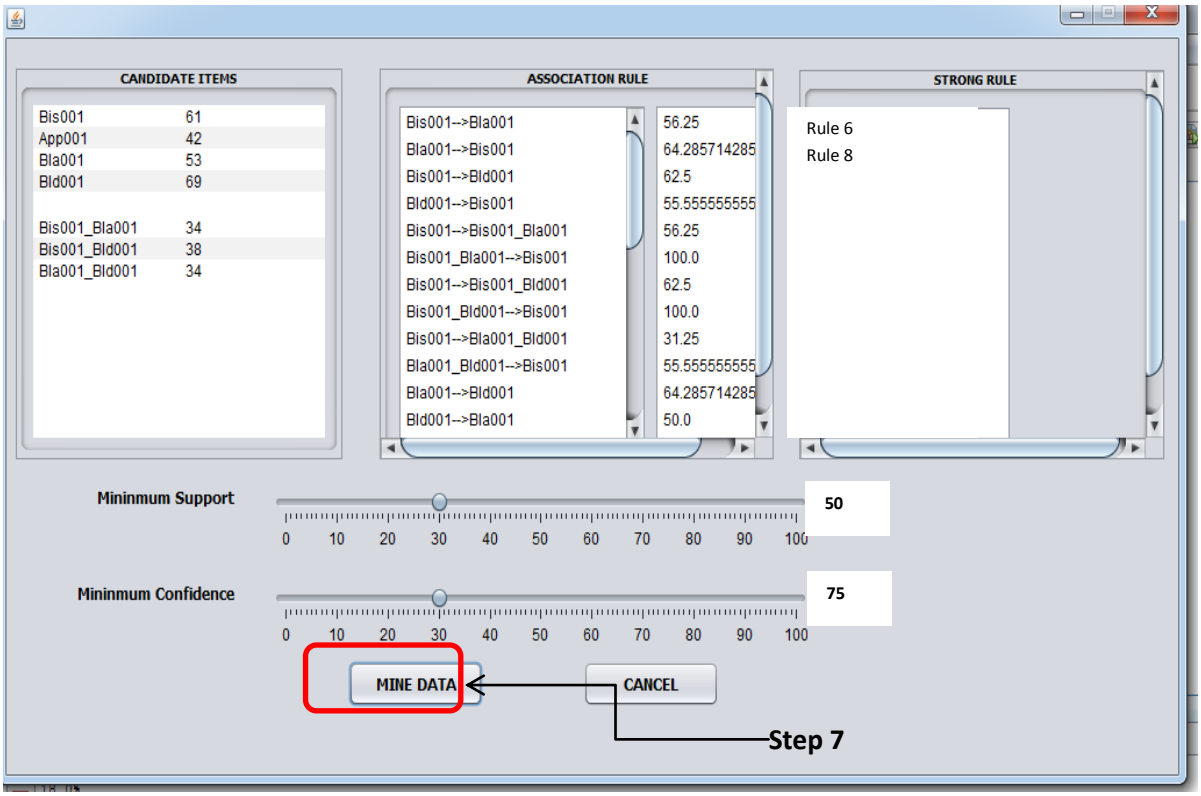

Figure 6: Association mining interface

5. CONCLUSION

This paper focused on understudying how market basket analysis could help provide leverage for business intelligence through Association rule mining. Although, the findings of this paper are applicable to any business that is involved in the sale of goods and services, the sales pattern of a supermarket were specifically analyzed. A transactional database was created and mined using Apriori algorithm implemented in Java programming language. Other Association rule mining algorithms can be applied to mine transactional database for the purpose of performance comparison.

6. REFERENCES

[1] Mulhern, F. N. and Leone, R.P. 1991. Implicit Price Bundling of Retail Products: A Multiproduct Approach to Maximizing Store Probability, Journal of Marketing, 55, pp. 63-76.

[2] Chib S., Nardari F. and Shephard N. 2002. Markov Chain Monte Carlo Methods for Stochastic Volatility Models, Journal of Econometrics, 108, pp. 281-316.

[3] Russell, G. J. and Petersen A. 2000. Analysis of Cross Category Dependence in Market Basket Selection, Journal of Retailing.

[4] Han, J. and Kamber, M. 2006. Data Mining: Concepts and Techniques, 2nd Edition, Morgan Kaufmann Publishers.

[5] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. 1996. From data mining to knowledge discovery in databases. AI Magazine, 17(3), pp. 37-54.

[6] Oladipupo, O. O. and Oyelade, O. J. 2010. Knowledge Discovery from Students' Result Repository: Association Rule Mining Approach, International Journal of Computer Science and Security, 4(2), pp.199-207.

[7] Agarwal, P. Yadav, M. L. and Anand, N. 2013. Study on Apriori Algorithm and its Application in Grocery Store, International Journal of Computer Applications, 74(14), pp. 1-8.

[8] Tissera, W.M.R., Athauda, R.I and Fernando, H.C. 2006. Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining, IEEE International Conference on Information Acquisition, pp. 57-62.

[9] Phani, Prasad J. and Murlidher, Mourya 2013. A Study on Market Basket Analysis using a Data Mining Algorithm, International Journal of Emerging Technology and Advanced Engineering, 3(6), pp. 361-363.

[10] Dhanabhakyam, M. and Punithavalli, M. 2011. A Survey on Data Mining Algorithm for Market Basket Analysis, Global Journal of Computer Science and Technology, 11(11), pp. 23-28.

[11] Bartik, V. 2009. 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.

[12] Sumithra, R. and Paul, S. 2010. Using Distributed Apriori Association Rule and Classical Apriori Mining Algorithms for Grid Based Knowledge Discovery, International Conference on Computing Communication and Networking Technologies, pp. 1-5.

[13] Qiang Niu, Shi-Xiong Xia, and Lei, Zhang 2009. Association Classification Based on Compactness of Rules, WKDD 2009, Second International Workshop on Knowledge Discovery and Data Mining, pp. 245-247.

[14] Lin, L. and Pei-qi, L. 2001. Study on an Improved Apriori Algorithm and its Application in Supermarket.

[15] Tan, P. N., Steinbach M. and Kumar V. 2006. Introduction to Data Mining, Addison Wesley.

[16] Agrawal, R., Imielinski, T. and Swami, A. 1993. Mining Association Rules between Sets of Items in Large Databases, Proceedings of the ACM SIGMOD International Conference on Management of Data.

[17] Association Rule Learning 2011. Retrieved January 12, 2014 from http://en.wikipedia.org/wiki/Association_rule_learning

[18] Larissa, T. M. 2003. Introduction to Data Mining, London: Oxford University Press.