VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Review on: Techniques for Predicting Frequent Items

Himanshu A. Chaudhari[1], Darshana S. Vartak[1], Nidhi U. Tripathi[1], Sunita Naik[2]
[1]*(B.E. Computer Engineering, VIVA Institute of Technology/Mumbai University, India)*
[2]*(Assistant Prof. Computer Engineering, VIVA Institute of Technology/Mumbai University, India)*

Abstract : *Electronic commerce(E- Commerce) is the trading or facilitation of trading in products or services using computer networks, such as the Internet. It comes under a part of Data Mining which takes large amount of data and extracts them. The paper uses the information about the techniques and methods used in the shopping cart for prediction of product that the customer wants to buy or will buy and shows the relevant products according to the cost of the product. The paper also summarizes the descriptive methods with examples. For predicting the frequent pattern of itemset, many prediction algorithms, rule mining techniques and various methods have already been designed for use of retail market. This paper examines literature analysis on several techniques for mining frequent itemsets.The survey comprises various tree formations like Partial tree, IT tree and algorithms with its advantages and its limitations.*

Keywords *– Association Rule Mining, Data Mining, Frequent Itemsets, IT tree, Market Basket Data, Prediction.*

1. INTRODUCTION

We live in a world where huge amount of data are collected each and every day. Analyzing such data is an important need. Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining is also known as Knowledge Discovery in Data (KDD). There are huge amount of data generated in the various organizations. Therefore organizer has to take number of decisions during extraction of useful data from the huge amount of data. But it is difficult to take out each and every record, so organizer finds frequently occurring data in the database. Pattern mining is a subfield of data mining. An interesting pattern is a pattern that appears frequently in a database. The purpose of frequent itemsets mining is to identify all frequent itemsets, i.e., itemsets that have at least a precised minimum support, the percentage of transactions containing the itemsets. Frequent patterns as a name suggest are patterns that occur frequently in data.

A frequent itemsets typically refers to a set of items that often appear together in a transactional dataset. For example, customer tends to purchase first laptop, followed by a digital camera and then a memory card, is a frequent pattern. Mining frequent patterns leads to the discovery of interesting association and correlation within data. Association rule mining is meant to find the frequent itemsets, correlations and associations from various type of database such as relational database, transaction database, sequence databases, streams, strings, spatial data, graphs, etc. Association rule mining tries to find the rules that direct how or why such items are often bought together in a given transaction with multiple items. The main application of association rule mining is market basket data. Association rule can be defined as $X \rightarrow Y$ where X, Y are itemsets with antecedent and consequent respectively. Market Basket analysis[5] consist of support and confidence where support is used to identify how frequently itemsets appears in dataset and confidence is used to identify how frequently the rule has been found to be true. The **support of a rule** is the number of sequence containing the rule divided by the total number of sequences. $\mathrm{Supp}(X \rightarrow Y) = p\ (A \cup B)$. The **confidence of a rule** is the number of sequence containing the rule divided by the number of sequences containing its antecedent. $\mathrm{Conf}(X \rightarrow Y) = \mathrm{supp}\ (A, B)/\mathrm{supp}\ (A)$. By using Support and confidence values, one can generates rules on incoming queries and more precised prediction can be determined using prediction algorithm.

2. TECHNIQUES FOR PREDICTION OF FREQUENT ITEMSETS

Frequent patterns are itemsets, subsequence, or substructures that appear in a data set with frequency no less than a user-specified threshold. Frequent itemsets are a form of frequent pattern. Discovery of all frequent itemsets is a typical data mining task. The original use has been as part of association rule discovery. By finding frequent itemsets, a retailer can learn what is commonly bought together. Especially important are pairs or larger sets of items that occur much more frequently than would be expected were the items bought independently. In this section, the methods for mining the simplest form of frequent patterns are given.

2.1 Prediction of Missing Item Set In Shopping Cart [1]

Author invented IT-Tree (Itemsets Tree) technique. In this paper proposed algorithm makes use of flagged IT-Tree. IT-Tree created from training data set. In this algorithm incoming itemsets were considered as input and depend on that return graph that defines the association rule. In this algorithm they first identified all high support, high confidence rules that have antecedent a subset of itemsets. Then after his it combines consequent of all these rules and then created a set of items which are frequently bought. This method mainly identify repeated occurrence of items and sort them accordingly. And most identical that is root items are indicated with Flagged items. But there are two major drawbacks like time taken for execution is more and this method requires more memory for processing.

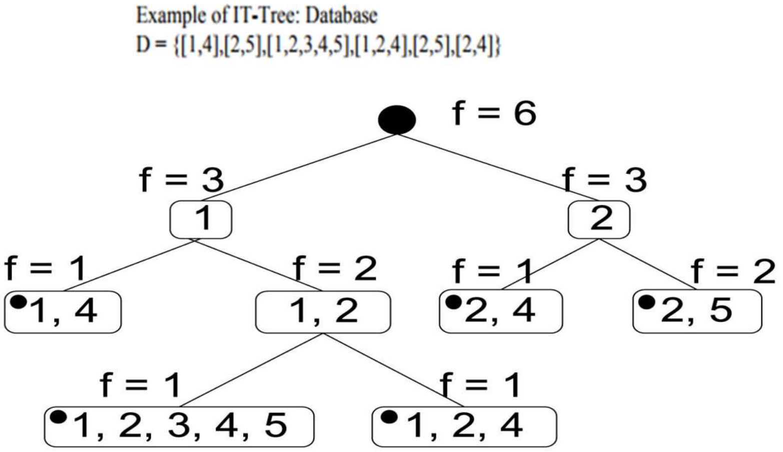

Figure 1: Construction of IT tree from given database [1]

Overall paper gives brief idea about generating the IT tree to scan the dataset and sort into identical itemset. It is Advantageous for computing itemset generation and can be used for generating candidate item sets.

2.2 Data Structure for Association Rule Mining: T -Tree and P Tree [2]

This paper demonstrates structure and algorithm using T tree and P tree with Advantage of storage and execution time. The Total support tree (T tree) method is used to create an object node. After this method tree is converted into array. The array format presents Partial support tree (P tree).This system proposed that the partial support tree is increases the performance of storage and execution time. It also overcomes the Apriori algorithm. In T tree and P tree structure branches are considered as independent therefore this structure can be used in parallel or distributed Association Rule Mining.

Thus paper finally concluded with two different types of tree formation method in which it first form tree and then convert it into array format which consumes memory and gives better performance in terms of support calculation.

2.3 Itemset Trees For Targeted Association Querying [3]

The paper proposed querying the database is made even faster by rearranging the database using the IT-tree data structure. This becomes handy especially in the batch mode Prediction (i.e., when you have to predict `missing items' for several shopping carts). IT-tree is a compact and easily updatable representation of the transactional dataset. Also, the construction of the itemset tree has O(N) space and time requirements. So, this data structure is used to speed up the proposed predictor.

The paper gives detail architecture of itemset tree and experiment with dataset. The experimental result is fast for query answering and it can use for large dataset. But it required more memory.

2.4 Finding Localized Associations In Market Basket Data [4]

The author introduced about market basket analysis which contain support and confidence values. One basket tells you about what one customer purchased at one time. It is a basically a theory that if you buy certain group of items, you are likely to buy another groups of items. Market basket analysis is used in most of all frequent mining concepts. Authors give clustering and indexing algorithms which are used to find significant correlations for association mining.

The paper includes clustering algorithm makes computational process simple with variable type of data. But it will increase its complexity if the problem size is increased.

2.5 An Approach for Predicting Missing Item from Large Transaction Database [5]

The system designed about the architecture to utilize the knowledge of incomplete constituents of a "shopping cart" for the guessing of what more the customer is likely to purchase. Author takes synthetic data obtained from IBM generator. Next step is taken as classification of clusters using Naïve Bayes text classifier and hierarchical document clustering which is simple to implement and used for large database. These clusters are then used for graph construction in form of Hash list which is combination of Hast table and Linked list. And finally Combo Matrix is used for prediction purpose.

Overall the concept of clustering in paper is useful to reduce memory requirement as it does not generate candidate set. But clustering has inability to recover from database corruption and it can arise problem due to data scarcity.

2.6 Review On: Prediction of Missing Item Set in Shopping Cart [6]

The paper reviews for prediction of frequent items in shopping cart. Predicting the missing items from dataset is indefinite area of research in data mining. In this paper some algorithm is introduced to identify the frequently co-occurring group of items in the transaction database for prediction purpose. In this paper author explains the existing approach which contain IT flagged tree. After getting IT tree the main root and identical items sets are indicated by black dot. They modified the original tree building algorithm by flagging each node that is identical to at least one transaction. This is called "Flagged IT tree". Disadvantage of this approach is it generates candidate itemsets which acquire memory space. It uses multiple passes over database. Author proposed Dempster Combination Rule which is used to combine all the rules.

Paper actually gives overall idea about predicting the missing items in shopping cart. Paper is focused on Dempster Shafer combination rule which is used to combine rules formed by rule generation. Proposed system described in paper is more flexible than other system. For e.g. processing speed with IT tree is much better than clustering the items.

2.7 Sequential Approach for Predicting Missing Items in Shopping Cart Using Apriori Algorithm. [7]

The author described sequential approach to predict the missing items in shopping cart using Apriori algorithm. The main objective of this paper to maintain the limitation of excessive wastage of time to hold a huge amount of candidate sets with much frequent itemsets .This system proposed to increase the performance of support value. The authors defined the disadvantages of Apriori algorithm that it generates the number of candidate items. The main disadvantage of this proposed system is wastage of memory.

The proposed system uses sequential approach for prediction using Apriori algorithm. It is basic algorithm and can be applied on any type of dataset. The system gives 65% accuracy with respect to prediction time. But it has disadvantages like storage capacity and I/O load so it cannot be use for long time.

2.8 Data Mining Approach For Retail Knowledge Discovery [8]

This Paper introduced the data mining techniques that are used in retail market for knowledge discovery are describes as following: Market Basket Analysis: Data mining association rules, also called market basket analysis, is one of the application areas of Data Mining. Consider a market with a collection of huge customer transactions. An association rule is X→Y where X is called the antecedent and Y is the consequent. X and Y are sets of items and the rule means that customers who buy X are likely to buy Y with probability %c where c is called the confidence. The algorithms generally try to optimize the speed since the transaction databases are huge in size. This type of information can be used in catalog design, store layout, product placement, target marketing, etc. Basket Analysis is related to, but different from Customer Relationship Management (CRM) systems where the aim is to find the dependencies between customer's demographic data.

The paper is all about review of literature based on techniques used in data mining for retail market knowledge discovery. And theoretically conclude with best approach as Apriori algorithm. But it cannot be used for larger dataset.

2.9 Comparing Data Set Characteristics That Favour The Apriori, Eclat Or FP-Growth Frequent Itemset Mining Algorithms. [9]

Existing system compares the frequent itemset mining techniques with respect to dataset characteristics. Author mainly focused on 3 main algorithms that are Apriori, Eclat and FP-Growth. All three algorithms are used to predict frequent item sets. Paper comprises survey on each algorithm with figure and example. Author gives advantage and disadvantages of each algorithm. In this paper, accuracy detected with respect to parameters as basket size vs. Runtime etc. By analysing these algorithms, author concludes that Apriori is basic and simplest algorithm. But Apriori has serious scalability issues and exhausts available memory much faster than Eclat and FP-Growth. Most frequent itemset applications should consider using either FP-Growth or Eclat.

This paper has beneficial for next version of Eclat or FP-Growth algorithm which decreases the complexity of both algorithms. The survey paper shows that Eclat and FP-growth algorithm is much better than Apriori in all cases.

2.10 An Enhanced Prediction Technique for Missing Itemsets in Shopping Cart [10]

This system proposed the shopping cart prediction architecture. Based on passed transaction we can easily construct a Graph structure from which association rules are generated in consideration of new incoming instances in new transaction. Then based on threshold value set by the user and kept dynamic, the prediction algorithm predicts the new item set to be considered for purchase. Threshold value is the minimum support value that a particular pair has to be present before getting predicted.

2.11 Predicting Missing Items In Shopping Cart Using Associative Classification Mining [11]

This paper describes generation of Boolean matrix using AND operation. And also introduced new concept BBA (Baysian Belief Argument) and rule selection which is used to select the rules from association rule where all the rules are identified using support and confidence values. After getting all possible generated rules, decision making algorithm that is Dempster Shafer algorithm is used for prediction.

Thus the system Combine all the rules using Dempster Shafer algorithm according to BBA and Rule selection technique.

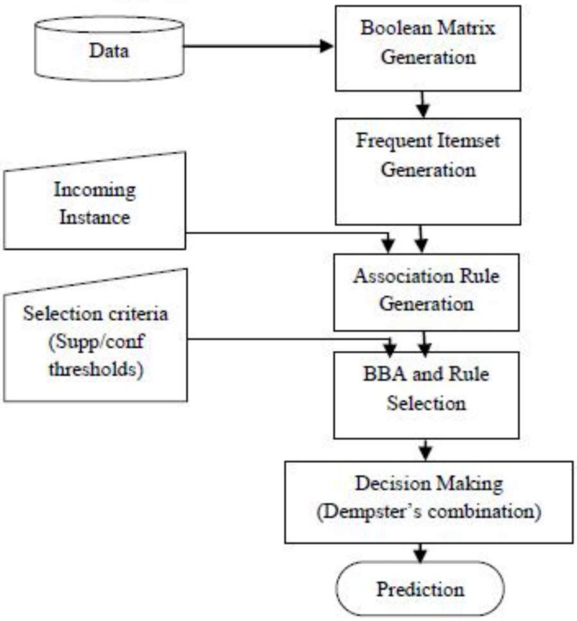
Data
Boolean Matrix Generation
Frequent Itemset Generation
Incoming Instance
Association Rule Generation
Selection criteria (Supp/conf thresholds)
BBA and Rule Selection
Decision Making (Dempster's combination)
Prediction

Figure 2: Shopping cart prediction architecture [11]

2.12 Missing Item Prediction And its Recommendation Based On Users Approach In E-Commerce [12]

This system proposed the algorithm in this spectrum use fast and effective technique. The system uses association rule mining techniques. This method produces high support and high confidence rules. This technique proves to appear better than the traditional techniques in association rule mining. But the cons of this technique are complexity increases with the increase in average length of items. The alternative method to predict missing items uses Boolean vector and the relational AND operation to discover frequent itemsets without generating candidate items. It directly generates the association rules.

By this proposed system, one can gain the information of predict the missing items uses Boolean matrix and AND relation operation.

2.13 A Survey on Approaches for Mining Frequent Itemsets [13]

Paper described algorithms for mining from Horizontal Layout Database for non-frequently bought items. Direct Hashing and Pruning (DHP) Algorithm: DHP can be derived from Apriori by introducing additional control. To this purposes DHP makes use of an additional hash table that aims at limiting the generation of candidates in set as much as possible. DHP also progressively trims the database by discarding attributes in transaction or even by discarding entire transactions when they appear to be subsequently useless. This method, support is counted by mapping the items from the candidate list into the buckets which is divided according to support known as Hash table structure. As the new itemset is encountered if item exist earlier then increase the bucket count else insert into new bucket. Thus in the end the bucket whose support count is less the minimum support is removed from the candidate set.

2.14 Association Rule Mining Using Improved Apriori Algorithm [14]

Author explained that Apriori algorithm generates interesting frequent or infrequent candidate item sets with respect to support count. Apriori algorithm can require to produce vast number of candidate sets. To generate the candidate sets, it needs several scans over the database. Apriori acquires more memory space for candidate generation process. While it takes multiple scans, it must require a lot of I/O load. The approach to overcome the difficulties is to get better Apriori algorithm by making some improvements in it. Also will develop pruning strategy as it will decrease the scans required to generate candidate item sets and accordingly find a valence or weightage to strong association rule. So that, memory and time needed to generate candidate item sets in Apriori will reduce. And the Apriori algorithm will get more effective and sufficient.

This Paper gives advantages of Improved Apriori algorithm is that it has less complex structure and less number of transaction as it scans the dataset less number of times than Apriori. But then also it has limitation of multiple scan with limited memory capacity.

2.15 An Efficient Prediction of Missing Itemset In Shopping Cart. [15]

The system proposed the shopping cart prediction architecture. Based on passed transaction we can easily construct a Graph structure from which association rules are generated in consideration of new incoming instances in new transaction. Then based on threshold value set by the user and kept dynamic, the Prediction algorithm predicts the new item set to be considered for purchase. Threshold Value is the minimum support value that a particular pair has to be present before getting predicted.

3. ANALYSIS

The papers are analyzed by techniques which are studied in module 2. The table analyzes according to techniques with respect to the parameters like support values, prediction time, transaction length, execution time etc.

Table 1: Analysis Table

Sr. No.	Title	Technique/Methods	Parameter	Accuracy
1	Prediction Of Missing Item Set In Shopping Cart [1]	Specific IT flagged tree , BBA	Transaction length vs. Prediction time and support threshold vs. Execution time	Minimum support, execution time = $56*10^3$s, Threshold =30%, if prediction time is 40s then average transaction length is 15.
2	Data Structure For Association Rule Mining :T -Tree And P Tree [2]	T tree and P tree formation in Apriori algorithm	support vs. Time, support vs. Storage, time vs. No. of records	If support is 4% then time required is 1 s. If time require is 30s then no. of records are $300*10^3$
3	Itemset Trees For Targeted Association Querying [3]	IT tree formation, association rule using Market Basket	Basket size vs. Time	For 10,000 distinct items, if there are 4000 baskets then time required to prediction is 10s
4	Finding Localized Associations In Market Basket Data [4]	Clustering algorithm, merging operation	No. of cluster vs. runtime and	N.A.
5	An Approach For Predicting Missing Item From Large Transaction Database [5]	Association rule (market basket analysis)	Length of transaction vs. Avg. Size of transaction	N.A.
6	Review On: Prediction Of Missing Item Set In Shopping Cart [6]	Flagged IT tree, Dempster combination rule (DCR)	N.A.	N.A.
7	Sequential Approach For Predicting Missing Items In Shopping Cart Using Apriori Algorithm. [7]	Apriori algorithm	N.A.	N.A.
8	Data Mining Approach For Retail Knowledge Discovery [8]	Market basket analysis and Apriori algorithm	N.A.	N.A.

9	Comparing Data Set Characteristics That Favour The Apriori, Eclat Or FP-Growth Frequent Itemset Mining Algorithms. [9]	Eclat , Apriori, FP-growth, naive brute method	density of frequent item vs. Runtime and size of basket vs. Runtime	N.A.
10	An Enhanced Prediction Technique For Missing Itemset In Shopping Cart [10]	Prediction accuracy measure to find prediction and recall	transaction length vs. Prediction time and execution time vs. Minimum support	N.A.
11	Predicting Missing Items In Shopping Cart Using Associative Classification Mining [11]	Association rule and Dempster Shafer theory	N.A.	N.A.
12	Missing Item Prediction And its Recommendation Based On Users Approach In E-Commerce. [12]	Association rule and Boolean matrix	N.A.	N.A.
13	A Survey On Approaches For Mining Frequent Itemsets [13]	Association rule	N.A.	N.A.
14	Association Rule Mining Using Improved Apriori Algorithm [14]	Improved Apriori Algorithm	Number of scan the dataset and time	No of scan to Ap =272 while no. of to improved Apri
15	An Efficient Prediction Of Missing Itemset In Shopping Cart. [15]	Association rule mining	Precision, recall, F-value and prediction time	Time required to predict item is les than existing syste

4. CONCLUSION

The goal of data mining is to predict the future or to understand the past. The paper includes analysis of various techniques used for predicting the frequent item set in shopping cart. The paper is all about review of literature based on techniques used in data mining for retail market knowledge discovery. Paper defines methods to find association rules with calculation of support and confidence values to get the rules. New algorithms like Improved Apriori as well as modifications of existing algorithms are often introduced thoroughly. From the above literature review on different techniques of frequent itemset, paper concludes as Improved Apriori is better for generating candidate items. To combine each rule DS-ARM is used based on threshold value to get predicted item. The limitations found in literature are unnecessary generation of candidate itemsets which takes more utilization of memory. Besides the technical limitations of any decision making (DS-ARM) its usability and popularity among practitioners should be a matter of concern Also found that the algorithms like Apriori make multiple scans in database. The drawbacks can be overcome by using less utilization of memory and less number of scan can decrease the execution time which will be better for performance of prediction of items.

REFERENCES

[1] K. Wickramaratna and M. Kubat, "Predicting Missing Item In Shopping Cart", *IEEE Transactions On Knowledge And Data Engineering, Volume 21 Issue 7*, July 2009.

[2] F. Coenen, P. Leng, and S. Ahmed, "Data Structure for Association Rule Mining: T-Trees and P-Trees", *IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 6*, June 2004.

[3] M. Kubat, A. Hafez, V. V. Raghavan, J. Lekkala, And W. K. Chen, "Itemset Trees For Targeted Association Querying", *IEEE Transactions On Knowledge And Data Engineering, Vol. 15, No. 6*, November/December 2003.

[4] C. Aggarwal, C. Procopiuc and P. Yu, "Finding Localized Associations In Market Basket Data", *IEEE Transactions On Knowledge And Data Engineering, Vol. 14, No. 1*, January/February 2002.

[5] P. Meshram, D. Gupta, P. Dahiwale, "An Approach For Predicting The Missing Items From Large Transaction Database", *IEEE Sponsored 2nd International Conference On Innovations In Information Embedded And Communication Systems Iciiecs'15*.

[6] S. Yende, P. Shirbhate, "Review On: Prediction Of Missing Item Set In Shopping Cart", *International Journal Of Research In Science & Engineering, Volume 1, Issue 1,* April 2017.

[7] R. Bodakhe, P. Gotarkar, A. Dahiwade, P. Gosavi, J.Syed, "A Sequential Approach For Predicting Missing Items In Shopping Cart Using Apriori Algorithm"*, Imperial Journal Of Interdisciplinary Research (IJIR) Volume 3, Issue4,* 2017.

[8] J. Vohra, "Data Mining Approach For Retail Knowledge Discovery", *International Journal Of Advanced Research In Computer Science And Software Engineering, Volume 6, Issue 3,* March 2016.

[9] J. Heaton, "Comparing Dataset Characteristics That Favour the Apriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms", 30 Jan 2017

[10] M. Nirmala, V. Palanisamy, "An Enhanced Prediction Technique For Missing Itemset In Shopping Cart", *International Journal Of Emerging Technology And Advanced Engineering, Volume 3, Issue 7,* July 2013 .

[11] K. Kumar, S. Sairam, "Predicting Missing Items In Shopping Cart Using Associative Classification Mining", *International Journal Of Computer Science And Mobile Computing, Volume 2, Issue 11,* November 2013.

[12] H. Deulkar, R. Shelke, "Implementation of Users Approach for Item Prediction and Its Recommendation In Ecommerce", *International Journal Of Innovative Research In Computer And Communication Engineering, Volume 5, Issue 4,* April 2017.

[13] S. Neelima, N. Satyanarayana and P. Krishna Murthy, "A Survey On Approaches For Mining Frequent Itemsets", *IOSR Journal Of Computer Engineering (IOSR-JCE), Volume 16, Issue 4, Ver. Vii,* (Jul – Aug. 2014), Pp 31-34.

[14] M. Ingle, N. Suryavanshi, "Association Rule Mining Using Improved Apriori Algorithm", *International Journal Of Computer Applications, Volume 112,Issue 4,* February 2015.

[15] M. Nirmala. and V. Palanisamy, "An Efficient Prediction Of Missing Itemset In Shopping Cart", *Journal Of Computer Science, Volume 9 (1)*, 2013, pp 55-62.