Mining Customer Data for Decision Making using Optimized Apriori Algorithm

Mr. Ravi Kumar Verma
Dept of Computer Science and Eng.
Parthivi College of Engineering & Management
Bhilai, Chhattisgarh, India

Mr. Sunil Kumar Sahu
Dept of Computer Science and Eng.
Parthivi College of Engineering & Management
Bhilai, Chhattisgarh, India

Mr. Hemant Sahu
Dept of Computer Science and Eng.
Parthivi College of Engineering & Management
Bhilai, Chhattisgarh, India

***Abstract*— This The huge amounts of data continuously being collected and stored in databases, many companies and firms are becoming interested in mining association rules from their databases to increase their profits. Frequent pattern discovery from customer data is playing an important role for business support and improvement. Punctually identification of new emerging trends is very important in business process. Sales patterns from inventory data signalize market tendency and can be used in forecasting which has great potential for decision making, strategic planning and market competition. The objective of this paper is to know the customer behavior at the time of purchase, how easily provide them what they want? The proposed approach makes use of the traditional Apriori algorithm to generate a set of association rules from a database and some improvement over Apriori for fast scanning of the database.**

Keywords-Frequent pattern; Confidence; mining; Association; Transaction

I. INTRODUCTION

In this age everyday customers purchase a lot of product from purchasing website and supermarket. These purchasing data are continuously collected and stored in databases. If we analyze this database we observe some common purchasing pattern of customer which is done by the customer. Frequent patterns stored in the stock or inventory data are very significant for business improvement and decision making. Continuously identification of new emerging trends is also significant in business improvement. Sales patterns are stored in the database indicate market trends and that can be used in forecasting which product has great potential for business improvement, production planning and market competition. The objectives of this research are to take better decision making for improving sale, services and quality as to identify the fast-moving products which is useful mechanisms for business support, increase product sales

A. Customer navigational behaviors

Every time customers navigate the web page for purchase items from purchasing a website that navigation is called customer navigational behavior. Customer's behavior data are continuously being collected and stored in databases in the form of log files. The main aim of frequent pattern mining is to extract the knowledge which is hidden in the log files on a web server. Customer navigational behavior data stored in the database in the form of log files, this is also called web log data. By use knowledge mining techniques to the web log data, frequent pattern concerning the user's navigational behavior may be known, like user and item clusters, further as potential correlations between purchase things.

B. Web log data log file

Web log is a log file where customer navigational behavior data are stored. Each access to an online page is recorded within the log files or the web server. The entries of web log file comprise fields that follow predefined format. Weblog report is a report format of user's navigation behavior. All this information is not required so we have to apply web mining task to extract transaction database.

C. Web mining approach

Web mining consist a wide range of applications that goal is discovering and extracting hidden information in data stored on the Web. Web mining additionally provides numerous techniques through that user or client will access the data simply and with efficiency. In the third step, approach is to search out the data which may be extracted from the directional patterns of the shoppers or users, that pattern are hold on in a web server within the style of log files. Thus, web mining method is categorized into 3 totally different categories supported that a part of the web is to be deep-mined. These three categories are (i) Web content mining, (ii) Web structure mining and (iii) Web usage mining.

D. Web usage mining (WUM)

Web usage mining is that the method of applying data processing techniques to find user navigation patterns from internet knowledge to know and higher serve the wants of users navigating to the online and supplies them terribly straightforward. Figure 1.3 shows the method of web usage mining accomplished as a case study during this work. As is seen, the input of the method is that the log knowledge. The

information need to be pre-processed so as to own the acceptable input for the mining algorithms.

As each data processing task, the method of web usage mining conjointly consists of 3 main steps: (i) pre-processing, (ii) pattern discovery and (iii) pattern analysis.

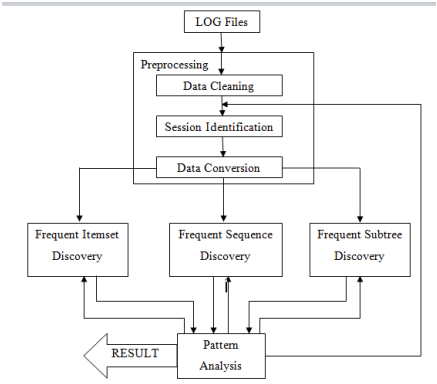


Figure 1 Process of web usage mining

In data pre-processing step data that return from the log file are tending to be screaming, incomplete and inconsistent. During this part, the info accessible ought to be treated per the wants of future part. It includes information cleansing, information integration, and information transformation and information reduction.

Pattern discovery suggests that applying the introduced frequent pattern discovery ways to the dealings info. For this reason the information got to be converted within the pre-processing part specified the output of the conversion are often used because the input of the algorithms. Several completely different ways and algorithms like statistics, data processing, machine learning and pattern recognition may be applied to spot user patterns.

Pattern analysis means that understanding the results obtained by the algorithms and drawing conclusions. This method targets to know, visualize and provides an interpretation of those patterns. Web usage mining depends on the collaboration of the user to permit the access of the web log records. After the invention has been achieved, the analysis of the patterns follows. The whole mining method is an associate unvaried task that is delineated by the feedback.

II. LITERATURE REVIEW

This section presents a comprehensive survey, primarily targeted on the study of analysis strategies for mining the frequent item sets and association rules with utility issues. Most of the prevailing work paid attention to perform the website and memory perceptions of the information.

A. Frequent Pattern Mining(FPM)

Frequent pattern mining is associate degree rising, active and heavily researched area within the field of data mining, with a good vary by application. Pattern mining is to use frequent pattern discovery strategies in journal information. Various Frequent patterns mining method discussed below

In 1993. Agrawal. R, Imielinski. T. & Hindoo. A. projected a replacement algorithmic program AIS algorithmic program uses candidate generation to find the frequent itemsets. They bestowed this rule "Mining association rules between sets of things in massive databases "in the ACM SIGMOD Conference on Management of information, Washington, D.C.

SETM algorithmic rule was really created by Huntsman and Hindu in Gregorian calendar month 1993 and enclosed during an analysis report whereas they were operating within the IBM Almaden center except for some reason it had been formally free solely in 1995. The algorithmic program additionally generates candidates on the fly supported the dealing scan from the info, similar to the AIS algorithmic program.

Agrawal R Imielinski.T & Swami. A proposed new algorithm which is called as Apriori algorithm in their research work "Mining association rules between sets of items in large databases". Apriori could be a classic rule for frequent item set mining and association rule learning over transactional databases. This algorithmic program initially projected in 1994 and remained the quality reference of all algorithms for locating all association rules.

The FP-Growth Formula, projected by Han dynasty for an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth, using an extended prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree.

In 1995 Park J. et al, proposed a new work in the ACM SIGMOD International Conference on Management of Data. DHP uses a hash technique that produces it terribly economical for the generation of candidate item sets, above all for the big two-item sets, therefore greatly rising the performance bottleneck of the full method.

In the 3rd International conference on knowledge discovery and data mining, Zaki. M. Parthasarathy. S., Ogihara M. & Li. W. proposed a new algorithm for fast discovery of Association Rules in 1997. It is the first algorithm that uses a vertical data (inverted) format. ECLAT is extremely economical for giant item sets however less economical for little ones. The frequent item sets area unit determined victimization straightforward tid-list intersections in an exceedingly depth-first graph.

III. METHODOLOGY

In this study we proposed a new work for improving business improvement and an improved apriori algorithm for mining patterns of huge stock data to predict factors affecting the sale of products. There are various data mining algorithm suggested by various researcher. To implement this proposed work we are going to choose Apriori algorithm, which is very popular data mining technique in the field of data mining research area. Apriori algorithm has potency to find out lagest common frequent pattern and it is easily implementable algorithm. And we have also tried to improve the efficiency of algorithm as suggested in literature review section.

A. An improved Apriori algorithm

The complexes of an association rule mining system scarcely depend on the identification of frequent item sets. The most well-liked rules for execution this identification is that the Apriori algorithm. Agrawal et al determined a remarkable downward closure property, known as Apriori, among frequent k-item sets: a k-item set is frequent provided that all of it sub item sets area unit frequent. Accordingly, the frequencies 1-itemsets are invented in the first access to the database, and then the frequent 1-itemsets are applied to produce candidate frequent 2-itemsets, and scan opposite the database to retrieve the frequent 2-itemsets. Normally, the frequent (k-1) –item sets square measure accustomed generate candidate frequent k-item sets, and check against the info to retrieve the frequent k-item sets. This method iterates till no additional frequent k-item sets may be discovered for a few k. this is often the outline of the Apriori algorithmic program. it's represented as follows:

Algorithm 1. Apriori

Input: D: a database; ms: minimum support;

Output: F: a set of frequent itemsets of interest;

(1) **Let** F ← {};

(2) **Let** L1 ← {frequent 1-itemsets}; F ← F ∪ L1;

(3) **For** (k = 2; (Lk-1 = {}); k++) **do begin**

//Generate all possible frequent k-item sets of interest in D.

(3.1) let Temk ← {{x1,. . . xk-2, xk-1, xk}|{x1, . . . , xk-2, xk-1} ∈ Lk-1 ∧{x1, . . . , xk-2, xk} ∈ Lk-1};

(3.2) for each transaction t in D do begin

//Check which k-itemsets are included in transaction t.

Let Temt ← the k-itemsets in t that are also contained in Temk;

For each itemset A in Temt **do**

Let A.count ← A.count + 1;

End for

(3.3) let Lk ← {c|c ∈ Temk ∧ (p(c) = (c.count/ |D|) ≥ ms)};

(3.4) let F ← F ∪ Lk;

End (3)

(4) Output F;

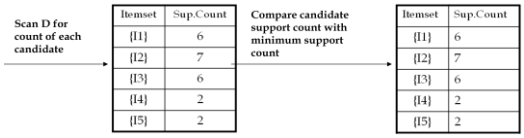
(5) Return.

B. Working of Apriori Algorihm

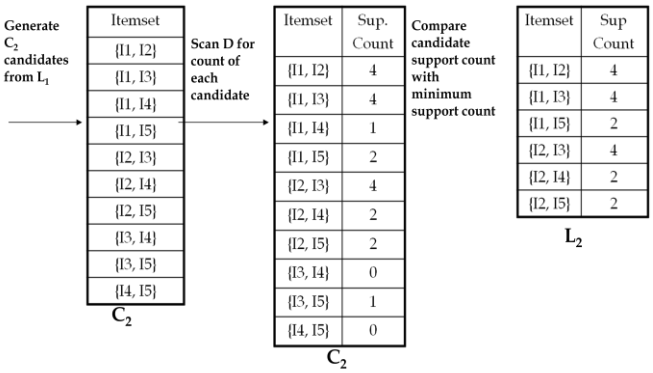
There are some steps to generate Most Frequent Pattern (MPF) from the Transaction Database D. The steps are followed-

- Consider an information D, consisting of nine transactions.
- Imagine min. support count required is two (min_sup = 2/9 = 22 %)
- Suppose minimum confidence expected is 70%. We got to 1st of all we've to come up with frequent item set exploitation Apriori algorithmic rule. Then, Association rules are going to be generated by exploitation min. support & min. confidence

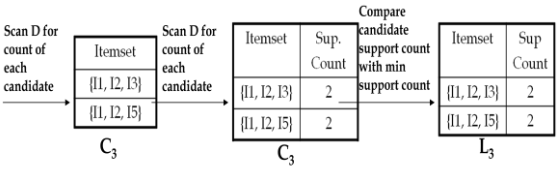
Step 1: Generating 1-Itemset frequent pattern.



Step 2: Generating 1-Itemset frequent pattern.



Step 3: Generating 3-Itemset frequent pattern



C. Optimization in Apriori

The optimization is principally some way of reducing question frequencies and storage resources. The designed to boost the Apriori algorithmic rule that mines frequent item sets while not new candidate generation. In this optimization we are going to overcome the problem suggested in problem identification section. A lot of researches have given their effort to improve the efficiency of Apriori algorithm. We are also going to suggest to some optimization techniques and implementing in this proposed work. In this proposed work we are going to reduce the size of the database and scan time of the database.

D. Screening Null Transactions

Null transactions may overburden the non-null transactions in any real time marketing. For example, from a supermarket, customers may buy neither buy beer nor chips, if these item sets are assumed to be two of the frequent item sets. If customer not purchase anything from the shopping website of supermarket then also these transactions are stored in database D. Algorithm have to scan all these null transactions these null transaction will not give any result and this scanning process is time consuming. Null transactions also

effect the various association and correlation scale. Assume during a grocery has a hundred transactions during which 40% are null transactions. Apriori or the other connected technique would scan all the a hundred transactions, the projected optimization technique effort to scale back the transactions by considering the sole valid sixty transactions when screening or avoiding the forty null transactions. This planned technique reduces the lots of computation time and reduce the database size also. Thus, in this proposed methodology effort has been made to delete the null transactions thereby, attempting to reduce the scanning time for discovering frequent k-item sets. Discovering null transactions and later eliminating them from the forthcoming idea of things is the first part of this proposed framework. This technique surely improves the efficiency of the algorithm.

E. Screening Single Transaction

A single transaction may also overburden for this algorithm. If a customer buys only beer from a supermarket it also store in the database. These single transactions are occupying the memory space in the database. The main objective of this proposed work is to find the frequent items et in the database, which is useful in improvement in marketing and sales. Due to this single transaction algorithm gives unambiguous result and it also consumes the searching time of the database.

F. Generating Association Rule

To show the way to generate association rules from a given info (the second step of the supp–conf framework), we have a tendency to use the higher than frequent item sets known from the dataset in Table.

For elucidative within the detail, we tend to describe a way to generate association rules from the frequent item set BCE in F, with p (BCE) = 50% = ms.

Because p (BCE)/p (BC) = 2/2 = 100 per cent, that is larger than minimum confidence mc = 60%, BC → E is extracted as a legitimate rule. Similarly, because p (BCE)/p (BE) = 2/3 = 66.7%, which is greater than mc, BE →C can be extracted as another valid rule; and because p(BCE)/p(CE) = 2/2 = 100% is greater than mc, CE → B can be extracted as a third valid rule. The association rules with 1-item consequences generated from the BCE area unit listed in Table 4.7.

Also, as a result of p (BCE)/p(BE) = 2/3 = 66.7% is bigger than megacycle, B →CE may be extracted as a sound rule. Similarly, as a result of p (BCE)/p(C) = 2/3 = 66.7% is bigger than megacycle, C →BE may be extracted as a sound rule; and p(BCE)/p(E) = 2/3 = 66.7% is bigger than megacycle, E →BC may be extracted as a sound rule.

IV. Result And Discussion

The optimized Apriori algorithmic rule is projected to scale back the disadvantage of classical Apriori algorithmic rule. Through pruning candidate item sets by rare item sets, this algorithmic program will cut back the amount of information scanning and also the redundancy whereas generating subtests and validating them within the info. Screening of null transaction also reduces the database scanning because null and single transaction, which is not useful in a frequent pattern generation. This proposed methodology increases the efficiency of Apriori algorithm and saves the precious time of customers to find the items in which they are interested in. This will surely improve the business profit.

V. Conclusion

In this paper, the matter of pattern discovery from stock data mining is self-addressed. Association rule mining encompasses a big selection of relevancy such market basket analysis, medical diagnosis/ analysis, web site navigation analysis, Office of Homeland Security, education, money and business domain then on. In this paper we tend to gift associate example of information mining technique diagrammatically by association rule mining refers to as market basket analysis. In the market basket analysis example the prevailing information was analysed to spot doubtless fascinating patterns. The target isn't solely to characterize the prevailing information. The simple marginal and conditional chances are a unit too little to inform the North American country regarding causative relationships a lot of subtle techniques area unit needed. Association rule mining is simple to use and implement and may improve the profit of firms. The procedure value of association rule mining represents an obstacle and future work can specialize in reducing it.

References

[1] Agrawal. R. Imielinski. T. and Swami. A. 1993. "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. 207–216.2006, 934-937.

[2] Agrawal. R. and Srikant. R. 1994. "Fast algorithms for mining association rules in large databases". Proceedings of the Twentieth International Conference on Very Large Databases. 1994, 487–499.

[3] Agrawal. R. and Srikant. R. 1994. "Fast Algorithm for Mining Association Rule". Proceedings of 20th VLDB Conference, Santiago, Chile 1994.

[4] Ahuja. L. and Kumar. E. 2011. "Multilevel Index Algorithm in Search Engine". CIS Journal. Journal of Emerging Trends in Computing and Information Sciences, 2, ISSN 2079-8407. No.:3.

[5] Chen. Y., Chen. J. and Tung. C. 2006. "A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales". Elsevier B.V. 0167-9236.

[6] Eirinaki. M. And Vazirgiannis. M. 2003. "Web Mining for Web Personalization". ACM Transactions on Internet Technology. 3, No. 1. Pages: 1–27.

[7] Fayyad. U. Shapiro. G. and Smyth. P. 1996. "From data mining to knowledge discovery: an overview". Adv Knowledge Discov Data Min 1996, 1–34.

[8] Frawley. W. Shapiro. G and Matheus C.1992. "Knowledge discovery in databases: An overview". AI Magazine. 1992, 13:57–70.

[9] Garg. D. and Sharma. H. "Comparative Analysis of Various Approaches Used in Frequent Pattern Mining". (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence. Pages: 141-147.

[10] Gunaseelan. D. and Uma. P. 2012. "An Improved Frequent Pattern Algorithm for Mining Association Rules". International Journal of Information and Communication Technology Research.

[11] Han. J., Cheng. H. And Xin. D. 2007. "Frequent pattern mining: current status and future Directions". Springer Science+Business Media, LLC 2007.

[12] Ivancsy. R. and vajak. I. 2006. "Frequent Pattern Mining in Web Log data". Acta Polytechnica Hungarica. 3, No. 1.Jadon. R. and Jain. R. 2010, "An Implementation of Frequent Pattern Mining Algorithm using Dynamic Function". International Journal of Computer Applications. 9, No.9. ISSN:0975 – 8887.

[13] Kibum. K. and Carroll. J. 2002. "An Empirical Study of Web Personalization Assistants : Supporting End-Users in Web Information Systems". Proceedings of the IEEE Symposia on Human Centric Computing Languages and Environments. (HCC'02) 0-7695-1644-0/02.

[14] Khan. A., Baharudin. B. and Khan. K. 2011. "Mining Customer Behavior data for Decision Making Using New Hybrid Classification Algorithm". Journals of theoretical and applied information technology. 27 No.1, ISSN:1942-8645.

[15] Khanchana. R. and Punithavalli. M. 2011. "Web page prediction for Web personalization". Global Journal of Computer Science and Technology. Publisher: Global Journals Inc. (USA). 11. ISSN: 0975-4172 & ISSN: 0975-4350. Issue 7 Version 1.0.

[16] Khanchana. R. and Punithavalli. M. 2011."W*eb Usage Mining for Predicting Users' Browsing Behaviors by using FPCM Clustering*", IACSIT International Journal of Engineering and Technology, 3, No. 5