Second International Conference on Nexgen Technologies
Sengunthar Engineering College, Tiruchengode, Namakkal Dist. Tamilnadu (India)
8th - 9th March 2019 www.conferenceworld.in ISBN : 978-93-87793-75-0

DISCOVERING HIGH UTILITY ITEMSETS FROM TRANSACTIONAL DATABASE

Mr. K.Tharani[1], D.Poorani[2]

[1]Assistant Professor, Department of Computer Science and Engineering,
Velalar College of Engineering and Technology, Tamilnadu (India)
[2]II Year M.E, Department of Computer Science and Engineering,
Velalar College of Engineering and Technology, Tamilnadu (India)

ABSTRACT

Ecommerce-oriented Data mining is a very promising area. It can automatically predict trends in customer spending, market trends which guide company to build personalized business intelligence web site, bring huge business profits. Data mining is combined with e-commerce systems with the appropriate data transformation bridges from the transaction processing system to the data warehouse and vice-versa to take advantage of this domain. Association rule learning in data mining is used to discover interesting relations between variables in large databases. Traditional model treats all the items in the database equally by only considering if an item is present in a transaction or not. The existing system uses FP-tree algorithm to find the frequent itemsets bought by the customers. But there is no way of knowing the quantity of the purchase. Utility pattern tree is proposed to mine high utility itemsets from a transactional database. It is achieved by employing large scale data sampling and length constraint strategy, reducing the number of candidate sets of frequent item sets and the global sensitivity.

Keywords*: *Association rule learning, FP- tree, frequent itemset, systolic tree, utility pattern mining,.

I. INTRODUCTION

Data mining is widely used to retrieve the hidden patterns and their relationships in a data by utilizing a variety of data analysis tools. It is a versatile subfield of computer science and statistics which has the overall goal of extracting information from a data and transforming them into a meaningful structure for further use. It involves database and database management aspects, data preprocessing, interestingness metrics, complexity considerations, post processing of discovered structures and online updating. Data mining and data warehouse are integrated together to retrieve information as in fig 1.

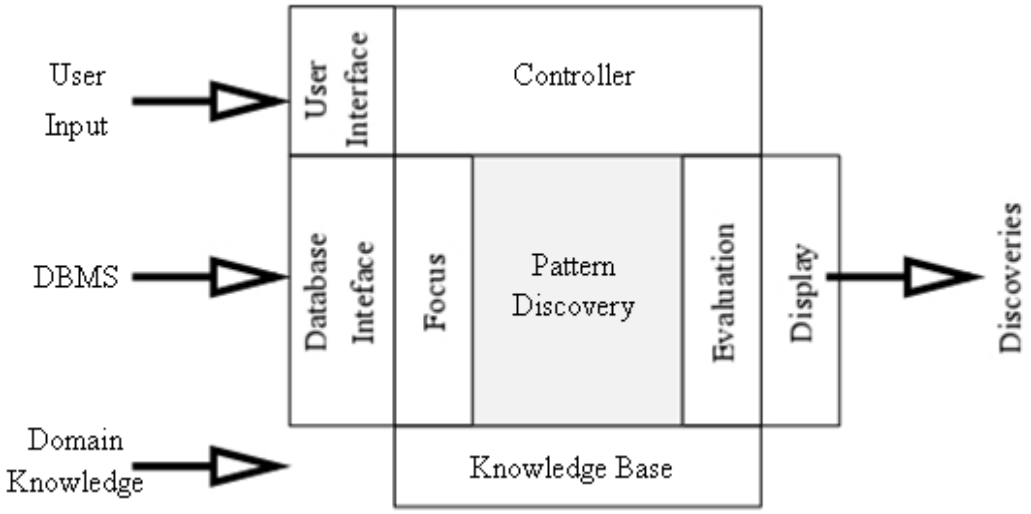


Figure 1: mining data from a database.

The first step in analysis of data in data mining is to characterize the data- compile its statistical attributes, reviewing it visually using charts and graphs, then look for possible meaningful links among the variables. In data mining process, collection, exploration and selection of the correct data is critically important. A predictive model must be constructed based on patterns because data description alone cannot determine an action plan. To verify the constructed model is the final step.

The data mining is usually applied to large scale data as well as any applications of computer decision support system, including artificial intelligence and business intelligence. Data mining consists of techniques like classification, association, outlier detection, clustering, prediction and more. But in terms of business perspective only few analysis is used which provides different outcomes and different insights. Anomaly, association rule learning, clustering, classification and regression analysis is the five most used technique for helping create business value.

II. MATERIALS AND METHODS

Association rule learning is rule based machine learning technique used to find interesting relations between variables in large databases. It identifies a strong rule in database using a measure of interestingness, which in terms will generate new rules as it analyzes more. In a point of sale (POS) systems, association rule learning discovers routines of product transactions in supermarkets.

Architecture diagram of high utility itemset is given in figure 2.

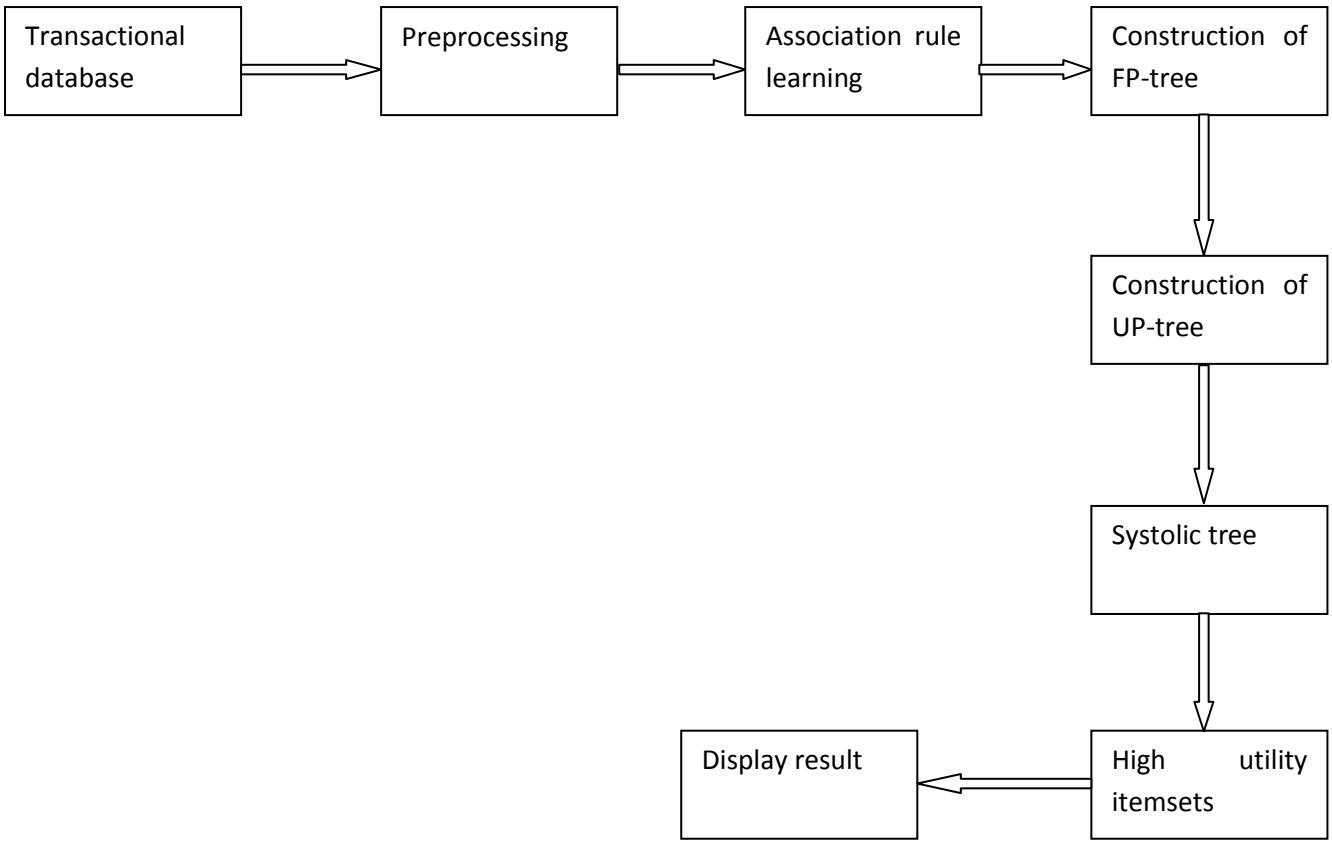


Figure 2: architecture of high utility itemset

Transactional database is taken and is preprocessed, thus converts the raw data into data useful for mining, also are split into columns. It uses preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

2.1 Market Basket Analysis

The technique is used to discover concurrence relationships among the products transactions by the customer in the supermarket. It is applied in general to the recorded transaction of an individual. It calculates support and confidence. Support is an measure of transaction in which the probability of product A and product B bought together.

Support $(A \rightarrow B) = P(A \cup B)$

Confidence is the measure of probability of having product B when product A is brought.

Confidence $(A \rightarrow B) = P(B|A)$

2.2 FP growth tree

Frequent pattern growth algorithm is an efficient method for mining complete set of frequent patterns by pattern using an extended prefix-tree structure which stores compressed and crucial information about frequent patterns. Transactional utilities are multiplied with their respective profit of a product using equation (1).

$$\mathbf{U(X)} = \sum \mathbf{U} \qquad (1)$$

Every multiplied product and transaction utilities are added together. The itemset weights are applied. For a specific product that involves in transaction have their utility values added together by using equation (2)

Itwu(x) = ∑ tu (Tq) (2)

A minimum threshold value is given to eliminate the profit values coming under the threshold value using equation (3)

min_util = ∂ X Σ tu (Tq) (3)

Where Sigma value is given as the user's preferred value of percentage. Based on their profit the products are arranged in descending order. The transactions and their routes are reorganized.

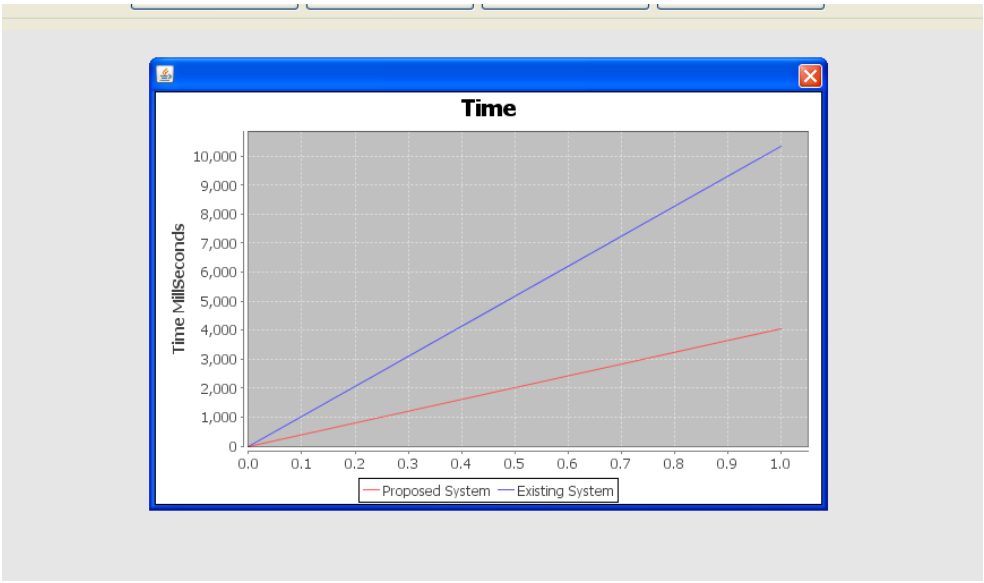
2.3 Utility pattern mining

Utility pattern mining is used to overcome the problem of the frequent itemset mining. The problem of the frequent itemset mining is that it does not take the quantity of the product into the account and the product appears as either 1 or 0 in transaction which denotes whether they are bought or not respectively. Every product are viewed with same utility weight. By applying strategy DGU, an UP tree is constructed. For every route support count and utility is estimated. A minimum threshold is applied that removes below thresholds.

The tree is rearranged with new paths and respective paths support count and utility.

III. RESULTS AND DISCUSSION

The algorithm presented has been experimented with real time databases Real-Store database. The graphical representation of running time and memory requirement values for various min_util values for retail-store database is shown in fig 2.



IV. CONCLUSIONS

There are many approaches to improve the efficiency of the algorithm. The efficiency would improve to a greater extent if the number of unwanted item sets is reduced. The proposed efficiency of the proposed algorithm reduces if the number of items and its support increases with the number of transaction being constant.

Future work includes mining top-k high utility itemsets from data streams and mining high utility itemsets from data streams with constraints of time spent in verifying the patterns depends on the number of candidates generated and reduces the efficiency of the algorithm provide the complete set of top-k high utility itemsets from every sliding window without missing any itemsets. With the evolution of the new application, the data processed may be in the continuous dynamic data streams. As the data streams come with high speed and are continuous and unbounded, mining result must be generated as fast as possible and make only one pass over a data.

REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Mag., vol. 17, no. 3, p. 37, 1999.

[2] H. Yang, K. Huang, I. King, and M. R. Lyu, "Localized support vector regression for time series prediction," Neurocomputing, vol. 72, nos. 10–12, pp. 2659–2669, 2009.

[3] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 6, pp. 601–618, Nov. 2010.

[4] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques. Amsterdam, The Netherlands: Elsevier, 2011.

[5] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, "Robust semi-supervised subspace clustering via non-negative low-rank representation," IEEE Trans. Cybern., vol. 46, no. 8, pp. 1828–1838, Aug. 2016.

[6] M. Peña, F. Biscarri, J. I. Guerrero, I. Monedero, and C. León, "Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach," Expert Syst. Appl., vol. 56, pp. 242–255, Sep. 2016.

[7] Y. Guo, F. Wang, B. Chen, and J. Xin, "Robust echo state networks based on correntropy induced loss function," Neurocomputing, vol. 267, pp. 295–303, Dec. 2017.

[8] H. Lim and H.-J. Kim, "Item recommendation using tag emotion in social cataloging services," Expert Syst. Appl., vol. 89, pp. 179–187, Dec. 2017.

[9] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2006, pp. 754–759.

[10] L. Sweeney, "K-anonymity: A model for protecting privacy," Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 557–570, 2002.

[11] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 571–588, 2002.

[12] A. Meyerson and R. Williams, "On the complexity of optimal K-anonymity," in Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst., 2004, pp. 223–228.

[13] Y. Zhang et al., "Embedding cryptographic features in compressive sensing," Neurocomputing, vol. 205, pp. 472–480, Sep. 2016.

[14] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," ACM Trans. Knowl. Discovery Data, vol. 1, no. 1, 2007, Art. no. 3.

[15] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-anonymity and l-diversity," in Proc. IEEE 23rd Int. Conf. Data Eng., Apr. 2007, pp. 106–115.