© 2015 IJSRSET | Volume 1 | Issue 6 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

High Utility ITEMSET Mining from Large Database

Annusooya A*, T. Seeniselvi
Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, Tamilnadu, India

ABSTRACT

Frequent itemset mining is one of the main problems in data mining. It has practical importance in a wide range of application areas such as decision support, Web usage mining, bioinformatics, etc. A number of relevant algorithms have been proposed in recent years for the fast access of data from the database. Mining high utility itemsets from a large database refers to the discovery of itemsets with high utility like profits. The proposed work is to mine the high utility items from the large database. The traditional association rule mining algorithm is used to find out the frequently occurring patterns of item sets. Apriori algorithm is used to find the high utility itemset. Data about the products are collected and stored in a database. Whenever customers buy the same product repeatedly the frequent pattern is formed and the infrequent items are separated. The high utility itemset is based on the user-specified utility threshold or it is a low-utility itemset. Admin maintain the entire system process like workers details, user details, product sales, raw materials. Admin can generate report based on the product sales. Admin can generate the Apriori products based on the threshold value. Admin can generate the graph for the frequently purchased products.
Keywords: Frequent Data Mining (FDM), Association Rule Mining, Apriori Algorithm.

I. INTRODUCTION

Frequent Itemset Mining

Recently, data mining or knowledge discovery from data bases has received much attention which deals with automatic discovery of implicit information within the databases. In which frequent itemset mining [1] plays a vital role in discovering all high utility itemsets with higher utility values than the minimum threshold in a transaction database [2] is one of the main problem facing in data mining.

High Utility Mining

High utility itemset mining has been resolved which is the extension of frequent itemset mining overcomes the problems of frequent itemset mining. Identifying the itemsets that occur frequently in transaction [3, 4, 5] with highest utilities is the main objective of utility mining with the considerations like user preferences such as profit, quantity and cost. Some of the standard methods for mining association rules [1, 6] that finds frequent itemsets are based on the support confidence model. And the problem of frequent itemset mining [7] is to find out the complete set of itemsets that appears with high occurrence in transactional databases. However, it identifies the utility of an itemset like profit, quantity, cost, weight which are important factors for addressing the real world decision problems that require maximizing the utility of an organization.

The meaning of itemset utility is interestingness, importance or profitability of an item to users [15]. Utility of items in a transaction database consists of two aspects:

i. Importance of distinct items called as external utility
ii. Importance of items in transactions called as internal utility.

Table 1.1: An Example Database

Profit	5	2	1	2	3	5	1	1
Item	A	B	C	D	E	F	G	H

This paper mines the high utility items from the large database with the traditional association rule mining algorithm to find out the frequently occurring patterns of item sets. Apriori algorithm is used to find the high utility itemset. Data about the products are collected and stored in a database. Whenever customers buy the same product repeatedly the frequent pattern is formed and the infrequent items are separated. The high utility itemset is based on the user-specified utility threshold or it is a low-utility itemset. The administrator produce the graph based on the utility of the itemset in the transaction.

II. METHODS AND MATERIAL

A. Literature Survey

R. Agrawal and R. Srikant et al [8] studies about the various algorithms for online mining of frequent itemsets (FIs). He reviews that unbounded memory requirement and the high data arrival rate of data streams in online are the combinatorial explosion of itemsets exacerbates the mining task is one of the scan nature.

Agrawal, R, Imielinski, T., Swami, A. [1] proposed a frequent itemset mining algorithm that uses the Apriori principle based on Support Confidence Model. An antimonotone property is proposed in the system to reduce the search space. Yao, H., Hamilton, H.J., Buzz, C.J. [9] proposed Umining and another heuristic based algorithm to find high utility itemsets. Pruning strategies is applied based on the mathematical properties of utility constraints which is efficient than previous researches.

Gouda, K. [10] proposed a backtracking search based algorithm for mining maximal frequent itemsets with a novel technique called progressive focusing that maximalist and diffset propagation to perform fast frequency computation. S. Shankar [11], presented a novel algorithm Fast Utility Mining to finds all high utility itemsets based on the given utility constraint threshold and suggest a novel method of generating different types of itemsets such as

i. High utility and high Frequency itemsets,
ii. High utility and Low Frequency itemsets
iii. Low Utility and High Frequency itemsets
iv. Low Utility and Low Frequency itemsets with the algorithm.

C.-W. Lin [12] designed a high utility pattern tree (HUP tree) algorithm to derive high utility patterns effectively and efficiently. The system integrates two-phase procedure for utility mining and the FP-tree concept to utilize the downward-closure property and to generate a compressed tree structure.

G.-C. Lan, T.-P. Hong, and V. S. Tseng [13] propose an efficient utility mining approach to adopt indexing mechanism the traditional method to speed up the execution and reduce the memory requirement in the mining process. It shows superior performance on real data. Gopalan, and N.R. Achuthan [6] CTU-PROL algorithm for mining high utility itemsets from large datasets. It followed the pattern growth approach [14]. The algorithm first finds the dataset if it is small, it creates compressed utility pattern tree for mining high utility itemsets. If the datasets is too large then the algorithm creates subdivisions using parallel projections that can be subsequently mined independently.

B. Problem Definition

Mining is a challenging task due to the possibility of long transactions. Users find difficult to comprehend the results in a very large number of high utility itemsets

- It may cause the algorithms to become inefficient in terms of memory and time requirement, or it may even run out of memory.
- For high utility itemsets, the algorithms consume more processing.
- Performance of the mining task decreases greatly based on minimum utility thresholds.

C. Proposed System

The proposed system is designed to achieve high efficiency for the mining task and provide a concise mining result to users, we propose a novel framework. The frequent pattern items are retrieved from the database by using the association rule mining. Once the threshold value is specified frequent pattern items are displayed in a sorted order. This reduces the memory space which in turn reduces the searching time. So the execution time becomes faster. Finally, the high utility pattern items should satisfy the threshold value. Primary constraints are:

1. The proposed representation is lossless and compact
2. It efficiently recovers all the high utility itemsets.

TID	Transaction	TU
T1	(A,1) (C,10) (D,1)	17
T2	(A,2) (C,6) (E,2) (G,5)	27
T3	(A,2) (B,2) (D,6) (E,2) (F,1)	37
T4	(B,4) (C,13) (D,3) (E,1)	30
T5	(B,2) (C,4) (E,1) (G,2)	13
T6	(A,1) (B,1) (C,1) (D,1) (H,2)	12

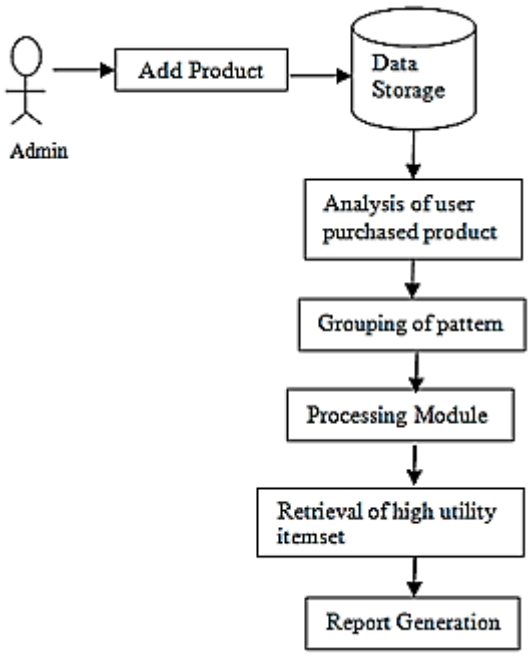

Figure 1 : Architecture diagram

Customer Seek

A user can search for a product of their choice. The user registers their details for login. The system will display the products which matches the selected search criteria. Customer can purchase the products by providing their credit card number which is stored in the database.

Administrator

Administrator wants to get access to all the functionalities of online product Store by his/her login Id. Administrator can update the products by deleted or adding the products which will be updated in the database. Admin maintain the entire details. Admin can generate the report based on the sales.

Frequent Item Mining

The frequent pattern items are retrieved from the database by using the association rule mining. Once the threshold value is specified frequent pattern items are displayed in a sorted order. This reduces the memory space which in turn reduces the searching time. So the execution time becomes faster.

Apriori Selection

By using the Apriori algorithm we can find out the high utility itemset. The high utility pattern items should satisfy the threshold value. From the list, admin can also know the products that are not frequent.

Report generation

Admin generates report based on the product sales. Admin can generate the report individually for the products and can view all the transaction details in month wise.

Apriori Algorithm

Apriori algorithm is designed for identifying frequent item set mining and association rule learning over the transactional databases. It keeps identifying the frequent individual items in the database and it extends them to larger and larger item sets till those item sets appear sufficiently often in database. The applications with Apriori algorithm is used for market basket analysis.

Apriori is a "bottom up" approach, where frequent subsets extend one item at a time and groups of candidates are tested against the provided data. When no further successful extensions are found the algorithm gets terminates.

Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. The algorithm generates candidate itemsets of length K from itemset K-1. Then it prunes candidates that contains infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-l length item sets. Finally, it scans the transaction database to determine frequent item sets among the candidates.

Pseudo-Code:

C_k: Candidate itemset of k size
L_k : frequent itemset of k size
L_1 = frequent items;
for (k = 0; L k != ∅; ++k) do begin
C_{k+1} = candidates generated from itemset L_k;
for each transaction t1 in database do
increments the count of all candidates in C_{k+1} that are contained in t1
L_{k+1} = candidates in Ck+1 with min_support end
return $\cup_k L_k$;

III. RESULTS AND DISCUSSION

Retrieval of high utility itemset efficiency by mining a large database and provide a concise mining result to users by a novel framework. Associations rule mining is used to retrieve frequent pattern items from database and if once the threshold value is specified, the frequent pattern items are displayed in a sorted order which reduce the search time.

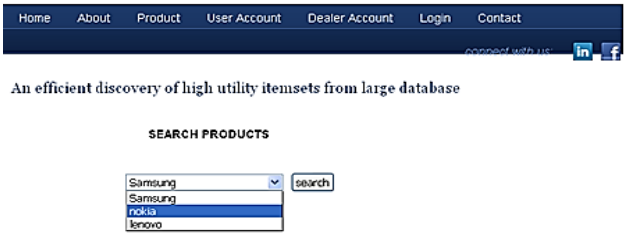

Figure 2: shows search product in online

Figure 2 shows the search product selected by a customer in online with their login id based on which the system provides the products which matches the selected search criteria. The user can purchase the item with credit card that have already stored in database while registering the personal data.

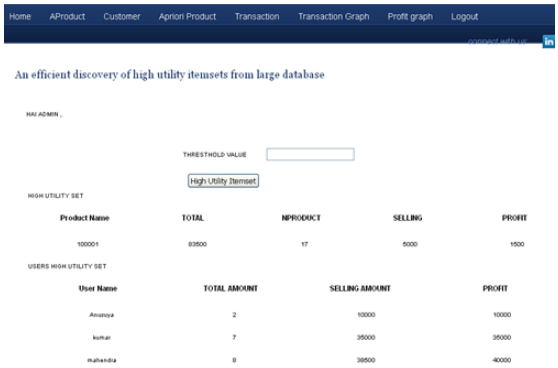

Figure 3: shows high utility itemset and threshold value

Figure 3 shows threshold value and high utility itemset which is detected with the proposed novel network from the large database. The itemsets are identified by means of profit achieved and no of items sold.

Figure 4: shows the chart of high utility itemset

Figure 4 shows the graph of high utility product from the product list with the code p1. The Top-N product char is display with x axis as product name and y axis as product price. The high utility itemset have been detected from the large database with the lossless and compact manner.

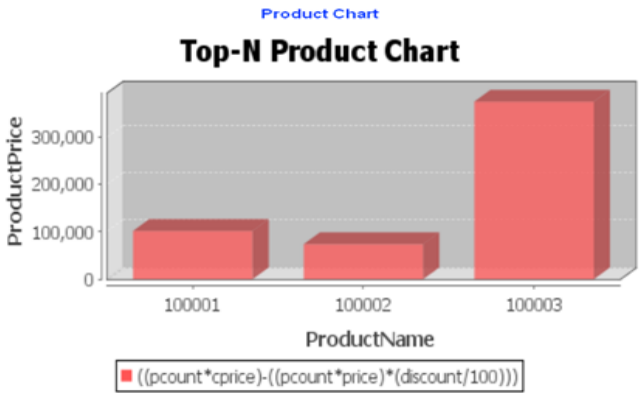

Figure 5: shows Top-N product chart

Figure 5 shows the Top-N product chart which is the frequent pattern items retrieved from the database by using the association rule mining. The threshold value specifies frequent pattern items are displayed in a sorted order which reduces the memory space which in turn reduces the searching time that makes the execution faster. Based on this approach high utility itemsets are achieved efficiently.

IV. CONCLUSION

The proposed system addresses the problem of high utility itemsets. To find out internet users web usage mining is the procedure is used. It can be described as the sighting and scrutiny of user ease of access pattern, during mining of files in internet which recognize and better serve up the desires of Web based applications. From the proposed system admin can know the high utility itemsets of user. This helps in many organizations to improve their profit and to analyze the user needs and expectations.

V. REFERENCES

[1] Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Database. In: ACM SIGMOD International Conference on Management of Data (1993).

[2] A. Erwin, R.P. Gopalan, and N.R. Achuthan,"Efficient Mining of High Utility Itemsets from Large Datasets", T. Washio et al. (Eds.): PAKDD2008, LNAI 5012, pp. 554–561, 2008. © Springer-Verlag Berlin Heidelberg 2008.

[3] Chan , Q.Yang,Y.D Shen, Mining high utility itemsets, in: Proceedings of the 3rd IEEE International Conference on Data Mining , Melbourne , Florida, 2003, pp.19-26.

[4] J Han, J.Pei, Y.Yin ,R. Mao Mining frequent Patterns without candidate generation:a frequent - pattern tree approach , Data Mining and Knowledge Discovery 8(1)(2004) 53-87

[5] J.Hu, A. Mojsilovic , High-utility pattern mining :A method for discovery of high-utility ietmsets,in :Pattern Recognition 40(2007) 3317-3324

[6] Erwin, A., Gopalan, R.P., Achuthan, N.R, "A Bottom-Up Projection Based Algorithm for Mining High Utility Itemsets", In: International Workshop on Integrating AI and Data Mining. Gold Coast, Australia (2007).

[7] Yao, H., Hamilton, H.J., Buzz, C. J., "A Foundational Approach to Mining Itemset Utilities from Databases", In: 4th SIAM International Conference on Data Mining, Florida USA (2004).

[8] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th Int. Conf. Very Large Data Bases, 1994, pp. 487–499.

[9] Yao, H., Hamilton, H.J., "Mining itemset utilities from transaction databases", Data & Knowledge Engineering 59(3), 603–626 (2006).

[10] Gouda and M. J. Zaki, "Efficiently mining maximal frequent itemsets," in Proc. IEEE Int. Conf. Data Mining, 2001, pp. 163–170.

[11] S.Shankar Dr. T .purusothaman, Kannimuthu s a novel utility and frequency based itemset mining approach for improving crm in retail business 2010 international journal of computer applications (0975 - 8887) volume 1 – no. 16

[12] C.-W. Lin, T.-P. Hong, and W.-H. Lu, "An effective tree structure for mining high utility itemsets," Expert Syst. Appl., vol. 38, no. 6, pp. 7419–7424, 2011.

[13] G.-C. Lan, T.-P. Hong, and V. S. Tseng, "An efficient projectionbased indexing approach for mining high utility itemsets," Knowl. Inf. Syst, vol. 38, no. 1, pp. 85–107, 2014.

[14] Han, J., Wang, J., Yin, Y., "Mining frequent patterns without candidate generation", In: ACM SIGMOD International Conference on Management of Data (2000).

[15] Smita R. Londhe , Rupali A. Mahajan , Bhagyashree J. Bhoyar "Overview on Methods for Mining High Utility Itemset from Transactional Database" International Journal of Scientific Engineering and Research (IJSER), Volume 1 Issue 4, December 2013