RESEARCH ARTICLE OPEN ACCESS

An Efficient Algorithm for Identification of Most Valuable Itemsets from WebTransaction Log Data

Litty Tressa George*, Asha Rose Thomas**
*(Computer Science and Engineering, Adi Shankara Institute of Engineering and Technology, Kalady, India)
** (Computer Science and Engineering, Adi Shankara Institute of Engineering and Technology,Kalady, India)

--************************----------------------------------

Abstract:

Web Utility mining has recently been a bloomingtopic in the field of data mining and so is the web mining, animportant research topic in database technologies. Thus, theweb utility mining is effective in not only discovering thefrequent temporal web transactions & generating high utilityitemsets, but also identifying the profit of webpages. Forenhancing the web utility mining, this study proposes a mixedapproach to the techniques of web mining, temporal highutility itemsets& On-shelf utility mining algorithms, toprovide web designers and decision makers more useful and meaningful web information. In the two Phases of thealgorithm, we came out with the more efficient and moderntechniques of web & utility mining in order to yield excellentresults on web transactional databases. Mining most valuableitemsets from a transactional dataset refers to theidentification of the itemsets with high utility value as profits.Although there are various algorithms for identifying highutility itemsets, this improved algorithm is focused on onlineshopping transaction data. The other similar algorithmsproposed so far arise a problem that is they all generate largeset of candidate itemsets for Most Valuable Itemsets and alsorequire large number database scans. Generation of largenumber of item sets decreases the performance of mining withrespect to execution time and space requirement. This situationmay worse when database contains a large number oftransactions. In the proposed system, information of valuableitemsets are recorded in tree based data structure called UtilityPattern Tree which is a compact representation of items intransaction database. By the creation of Utility Pattern Tree,candidate itemsets are generated with only two scans of thedatabase. Recommended algorithms not only reduce a numberof candidate itemsets but also work efficiently when databasehas lots of long transactions.

Keywords —Utility Mining, Itemset utility, Valuable itemsets, Most valuable itemsets.

--************************--------------------------

INTRODUCTION

Extracting or “mining” knowledge from large amounts ofdata stored in databases, data warehouses or any otherinformation repositories is called data mining. A dataset can be defined as any named group of records upon which datamining is performed. Groups of items that appear togetherin any transaction datasets can be called as Itemsets.Frequent itemsets are set of items that appear in a data set frequently. For example, a set of items, like milk and bread, appears frequently together in a transaction data set. Findingsuch frequent patterns plays an essential role in dataclassification, mining associations, correlations, identifyingmany interesting relationships among data. Thus, frequent

pattern mining has turn out to be an imperative data miningtask and a focused theme in data mining research. Frequentpattern mining searches for recurring relationships in agiven data set.The problem with frequent pattern mining was that, individual importance of each pattern is not considered. Infrequent itemset mining which is a type of frequent patternmining, where the patterns are itemsets, only the occurrenceof items are considered. Unit profits as well as purchasedquantities of the items were not taken into consideration.Individual importance of each item is not considered infrequent itemset mining this is the main limitation offrequent itemset mining. Therefore, it cannot satisfy therequirements of users who are interested in discovering theitem sets with high sales profits, since the profits in thesense- unit profits or weights, or even purchased quantities. For example, each item in a supermarket has different priceor profit and multiple copies of an item can be sold in atransaction. Hence, the most profitable item sets cannot befound in those frameworks since profit of an item set can becalculated by multiplying unit profit of each item in the itemset by the quantities in transactions including the item set.To find the most valuable item sets, both the importance andquantity of each item have to be reflected. In view of this, utility mining arises as a main topic in data mining field.

I. STATE OF ART

A. Frequent Pattern Mining

Frequent pattern mining is concerned with the mining ofmost frequently appearing patterns within a dataset. Herethe problem is to discover the complete set of patternssatisfying a minimum support in the transaction database.The entire dataset is pruned on the basis of downwardclosure property to identify the infrequent patterns. Thedownward closure property states that if a pattern is
infrequent, then all of its super patterns must also beinfrequent. The Apriori algorithm [2] was the first solutionto mine frequent patterns. It is a breadth first searchalgorithm. The drawback was that it suffers from a level wisecandidate generation and test problem and also it needsseveral database scans. That is for the first database scan,the Apriori discovers all the one-element itemsets and on thebasis of that produces the candidates for two-elementfrequent itemsets. In the second database scan, Aprioriidentifies all of the two-element frequent itemsets, andbased on that, generates the candidates for three-elementfrequent itemsets and so on. Thus if the size of the largestfrequent itemset is 'n',thenApriori needs 'n' database scans.In order to overcome this limitation, later FP growth [3]algorithm was proposed. It was a depth first searchalgorithm. It needed only two database scans for generatingfrequent patterns without any candidate generation.

B. Weighted Frequent Pattern Mining

Weighted frequent pattern mining deals with binarydatabases where frequency of each item in each transactioncan be either 1 or 0. W. Wang [5] et al in proposedweighted association rule mining algorithm WAR .In WAR,we discover first frequent itemsets and the weighted association rules for each frequent itemset are generated .Theweight of a pattern ‘p’ is defined as the proportion of thesum of all its items’ weight value to the length of ‘p’. Theforemost challenge in this area is the weighted frequency ofa pattern does not satisfy the downward closure property.Thus the mining performance cannot be improved. In orderto address this problem, Cai.et al. [4] first proposed theconcept of a weighted downward closure property. Bymeans of the transaction weight, weighted support can notonly reveal the importance of an itemset but also retains thedownward closure property in the course of the miningprocess. Although weighted association rule miningconsiders the importance of items, in some applications, such as transaction databases, items’ quantities intransactions are not taken into considerations yet. Also thenon-binary occurrence of items in transactions is notconsidered. Thus, the matter of high utility itemset miningcame into scene.

C.*High Utility Pattern Mining*

High utility pattern mining focus on mining the highlyutilized or most valuable patterns (it can be itemsets) from adataset. Identification of only frequent patterns in a databasecannot achieve the requirement of identifying the most valuable itemsets that add to the majority of the total profitsin a retail business. This gives the inspiration to develop amining model to determine those itemsets contributing tothe majority of the profit. To find the most valuable itemsets, both the importance and quantity of each item have tobe reflected. Identification of the item sets with high utilitiesis called as High Utility Item set Mining. Here, the meaningof item set utility is interestingness, importance, profitabilityor any other user relevant feature exhibited by the items ofthe dataset under consideration.

1) Two Phase Algorithm: Liu et al. proposed an algorithm named Two Phase [6]which is mainly consists of two mining phases. In phase I, itincludes an Apriori-based level-wise method which is abreadth first search strategy, to enumerate High TransactionWeighted Utility Itemsets (HTWUI). Candidate itemsetswhich are of length k are generated from length k-1HTWUIs, and their TWUs are calculated by scanning thedatabase once in each pass. After these steps, the whole setof HTWUIs is collected in phase I. In phase II, HTWUIsthat are high utility itemsets are recognized with anadditional database scan. The authors have well-defined thetransaction-weighted utilization (twu) and by that theyproved it is possible to sustain the downward closureproperty. In the initial database scan, the algorithmdiscovers all the one-element transaction-weightedutilization itemsets, and based on that result, it produces thecandidates for two element transaction-weighted utilizationitemsets. In the second database scan, it discovers all thetwo-element transaction-weighted utilization itemsets, andbased on that result, it creates the candidates for threeelementtransaction weighted utilization itemsets, and so on.At the final scan, the Two-Phase algorithm discovers thereal high utility itemsets from the high transaction-weightedutilization itemsets. This algorithm suffers from thedifficulty of the level-wise candidate generation-and-testmethodology. Although two-phase algorithm decreasessearch space by using TWDC property, it still produces too many candidates to obtain HTWUIs and needs multipledatabase scans.

2) Incremental High Utility Pattern (IHUP) Algorithm: To efficiently produce HTWUIs in phase I and avoidscanning database multiple times, Ahmed et al. [9]proposed a tree-based algorithm, called IHUP, for mininghigh utility itemsets. It includes an IHUP-Tree to maintainthe information of high utility itemsets and transactions.Every node in IHUP-Tree consists of an item name, asupport count, and a TWU value. The structure of thealgorithm consists of three parts: (1) The construction ofIHUP-Tree, (2) the generation of HTWUIs and (3)identification of high utility itemsets. In step 1, items in thetransaction are reorganized in a fixed order likelexicographic order, support descending order or TWUdescending order. Then, the reorganized transactions are putin into the IHUP-Tree. In step2, HTWUIs are producedfrom the IHUP-Tree by applying the FP-Growthalgorithm. Thus, HTWUIs in phase I can be retrieved morecapably without producing candidates for HTWUIs. In step3, high utility itemsets and their utilities are recognized fromthe set of HTWUIs by scanning the original database once.Even though IHUP finds HTWUIs without producing anycandidates for HTWUIs and achieves a better performancethan Two-Phase, it still results in too many HTWUIs inphase I since the overestimated utility calculated by TWU istoo long. IHUP and Two-Phase produce the similar numberof HTWUIs in phase I, since they use Transaction-WeightedUtilization mining (TWU) model to overestimate theutilities of the itemsets. However, this model mayoverestimate too many low utility itemsets as HTWUIs andproduce too many candidate itemsets in phase I. Such ahuge number of HTWUIs reduces the mining performancein phase I in terms of execution time and memoryconsumption. Besides, the number of HTWUIs in phase Ialso decrease the performance of the algorithms in phase IIsince the more HTWUIs are generated in phase I, the moreexecution time is required for recognizing high utilityitemsets in phase II.

3)Utility Pattern Growth (UP) Algorithm: The framework of the UP Growth method proposed by V.S.Tseng [12] consists of three parts: (1) construction of UP Tree,(2) generation of potential high utility itemsets fromthe UP-Tree by UP-Growth, and (3) identification of highutility itemsets from the set of potential high utility itemsets.In this algorithm, a new term called potential high utilityitemsets (PHUIs) is used to distinguish the discoveredpatterns found by up growth approach from the HTWUIssince our approach is not based on the traditional framework of transaction-weighted utilization mining model. UP Growthefficiently generates PHUIs from the global UPTree with two strategies, namely DGU (Discarding GlobalUnpromising items) to eliminate the low utility items andtheir utilities from the transaction utilities and DGN(Decreasing Global Node utilities) node utilities which arenearer to UP-Tree root node are effectively reduced. Eventhough it successfully generates the PHUI's by the abovetwo strategies, the problem is that there will be more no ofPHUI's.

II. METHODOLOGY

In this paper, we are considering one web transaction as one visit to the website which may include ordering a single product (item) or multiple (different) products (otherwise itemsets i.e., group of items) of specified quantities.Usually, a single webpage may be allotted for a singleproduct. A web transaction is said to occur not just byhitting a web page of a company's product details orspending some time on it; but by ordering or paying for aproduct and thus making some profit to the company. So ourwork will be focusing on analyzing which item or itemsetsare providing maximum profit value i.e. the most valuableones. The framework of the intended system consist of analgorithm UP-Growth+ which is an improved version of thestate of the art algorithm UP-Growth, used for mining high utility itemsets. The entire algorithm consists of three steps: 1) Scanning of the database twice to construct a global UPTree with the two strategies DGU and DGN. 2) Recursivegeneration of valuable itemsets from global UP-Tree andlocal UP-Trees by UP-Growth with the strategies DNU andDNN. 3) Identifying the most valuable itemsets from the setof valuable itemsets. Thus the entire process can be described as two phases. Both phases consists of twostrategies each.

A. Problem Statement

Given a transaction dataset D and a client specifiedminimum utility threshold min_util, the task of mining themost valuable itemsets from D is to discover the completeset of the items whose utilities are larger than or equal tominimum utility threshold value without any miss, andanalyzing the importance (value) of each individual item, sothat the most valuable itemsets among them, can beretrieved.

B. Preliminary

An itemset X is a set of k distinct items {i1, i2, ...,ik}. Anitemset with length k is called a k-itemset. A transactiondatabase D = {T1, T2, ...Tn} contains a set of transactions,and each transaction has a unique identifier d, called TID.Each item ip in transaction Td is associated with a quantityq (ip, Td), that is, the purchased quantity of ip in Td.Consider the following small sample set of items retrievedfrom the web transactional data of an online shopping site.Each item is followed by its count, i.e. number ofoccurrences in that corresponding transaction. So, one complete transaction includes different products in differentweb pages of different quantities ordered by a single customer in one single visit to the site.

TABLE I
A SAMPLE TRANSACTION DATASET.

T id	Transaction
T_1	(A,1) (C,1) (D,1)
T_2	(A,2) (E,6) (C,2) (G,5)
T_3	(A,1) (B,2) (C,1) (D,6) (E,1) (F,5)
T_4	(B,4) (C,3) (D,3) (E,1)
T_5	(B,2),(C,2) (E,1) (G,2)

TABLE 2

PROFIT TABLE

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

Definition 1: Utility of an item ip in a transaction Td isdenoted as u (ip, Td) and is defined as pr(ip)×q (ip, Td).

Definition 2: Utility of an itemset X in Td is denoted as u(X,Td) and defined as Σip□X∨X⊆ Td u (ip, Td).

Definition 3: Utility of an itemset X in D is the sum of theutilities of all occurrences of item X in the entire dataset Dand is denoted as u(X) and defined as ΣX⊆Td∧Td⊆Du(X,Td).

Definition 4.Transaction Utility of a transaction Td isdenoted as TU(Td) and is defined as the sum total of theutilities of all items in that transaction Td.

Definition 5.Transaction Weighted Utility of an itemset X isthe sum of the transaction utilities of all the transactionscontaining X, in the dataset and is denoted as TWU(X).

Definition 6: An item or itemset is called a Valuable Itemsetif its utility is not less than a user-specified minimum utilitythreshold or else they are called less-utility or less valuableitemset and is denoted as VI.

Definition 7: An item or itemset is called the Most ValuableItemsetif it is having a utility value greatest of all utilityvalues of the highly Valuable Itemsets and is denoted asMVI.

C. *Proposed Framework*

The framework of proposed method consists of twophases.

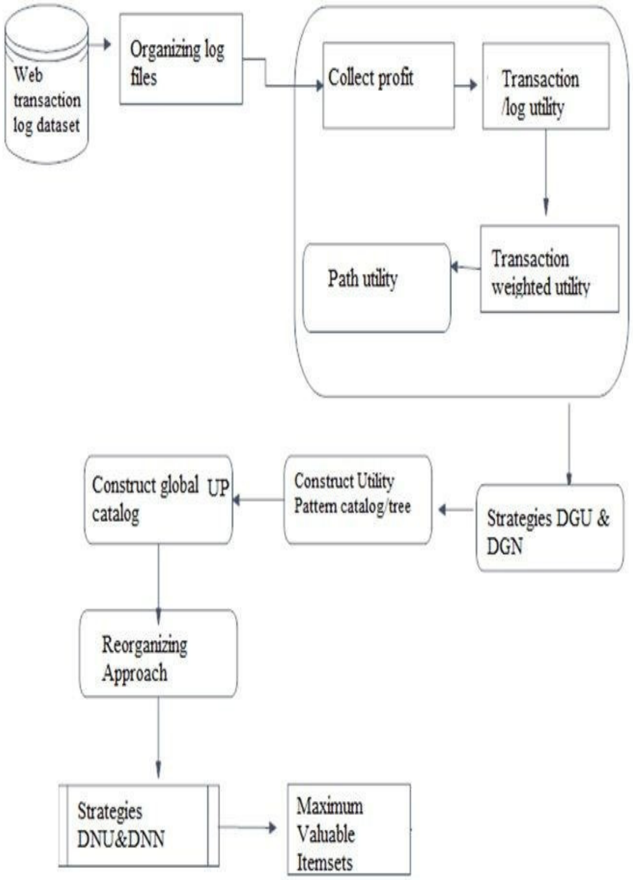
Web transaction log dataset
Organizing log files
Collect profit
Transaction /log utility
Transaction weighted utility
Path utility
Strategies DGU & DGN
Construct Utility Pattern catalog/tree
Construct global UP catalog
Reorganizing Approach
Strategies DNU&DNN
Maximum Valuable Itemsets

Fig. 1Illustration of the proposed framework

Phase 1: Scan the database twice to construct a global UPTree with the first two strategies.

Phase 2: Recursively generate potential highly ValuableItemsets (abbreviated as VI's) from global UPlocal UP-Trees by UP Growth+ with the last two strategies.

Phase 1

The Proposed Data Structure: UP-Tree

To facilitate the mining performance and avoid scanning original database repeatedly, we will use a compact tree structure, named UP-Tree (Utility Pattern Tree), to maintain the information of transactions and high utility itemsets.

Elements of UP-Tree

In an UP-Tree, each node N consists of *N. name*: the node's item name, *N.count*: the node's support

count, *N.nu*: the node's node utility, *N .parent*: the parent node of N, *N.link*: a node link which points to a node whose item name is the same as *N.name*.Two strategies are applied to minimize the
overestimated utilities stored in t he nodes of global UP-Tree. In following subsections, the elements of UP-Tree are first defined. Next, the two strategie s are introduced

Strategy 1: Discarding Global Unpromising Items – DGU

TU of each transaction is computed. Then the TWU of each single item is also accumulated. On the basis of TWU global unpromising items are then discard ed. Thus the unpromising items are removed from the transaction and their corresponding utility values are also eliminated from the initial TU of the transaction. The remaining promising items in the transaction are sorted in the descending order of TWU and is classified under the name Re -organized Transactions. By reorganizing the transactions, n ot only less information is needed to be recorded in UP-Tree, but also smaller overestimated utilities foritemsets a regenerated.

Strategy 2: Decreasing Global Node Utilities – DGN

Re-organized Transactions are then inserted into UP-Tree in a particular manner. Initially, the first reorganized transaction (e.g. T'1) is retrieved. Then the first node representing the first item in that transaction is created with Na.item= {A} and Na.count=1 and Na.nu is calculated as RTU { T'1 } minus the sum of the utilities of the rest items that are behind {C} in T'1. All the reorganized transactions are then inserted in the same way.

Phase 2

Strategy 3: *Discarding local unpromising items and their estimated Node Utilities from the paths and path utilities of conditional pattern bases – DNU*

Here the MNU's of the nodes are calculated and is recorded in a table named Minimum No de Utility table. Then conditional pattern bases CPB's of each nodes are generated. By scanning the CPB once the path utility of each local item is calculated. Then DNU is applied. That is the local unpromising items are found out and their MNU's are discarded from the path utilities of their associated paths and the path utilities are recalculated and the items in each path are reorganized by descen ding order of path utility of local items.

D. Experiment and result

1) Practical environment.

In this section, the input dataset and its type, practical results and environment is described.

Input.

A transaction dataset and profit table corresponding to the items in the dataset are used for the experiment.

Hardware Requirements:-

1) Operating System: windows XP/ Win7
2) Processor: Pentium IV or advanced
3) RAM: 2 GB
4) HDD: 160 GB

Software Requirements:-

1) Programming Language: Java
2) Framework: Net beans 6.8 or more
3) Development Kit: JDK 1.6 or more
4) Database: My SQL

Output

All Maximum Valuable Itemsets in the input dataset.

2) Data Collection

Any web transaction dataset with the itemset details including their count can be collected from any of the web transaction sites. It should be then tabulated or orderly arranged along with corresponding counts. Datasets can also be collected according to the requirement from the FIMI Repository [13].

3) Result Analysis

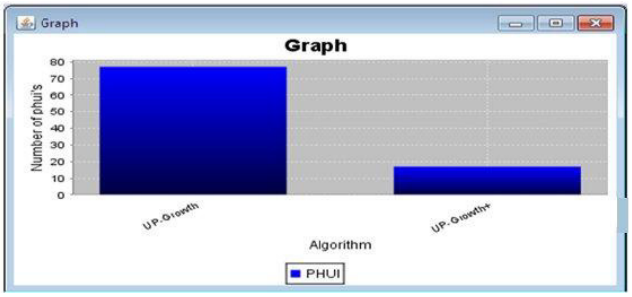

Figure 2: PHUI generation comparison

From the figure 2 which is a graph that illustrates the PHUI's generated (in our method it is called as Valuable Itemsets) by the state of the art algorithm UP growth Vs. the UP growth plus algorithm. From the graph it's clear that the UP growth plus effectively reduce s the number of valuable itemsets and thus makes it easy to mine the Most Valuable Itemsets from these.

III. CONCLUSION

To present a new scheme for high utility itemset miningfrom web transactional data, aiming to be with highperformance in terms of performance, scalability and time. This method is very much useful where continuous updating goes on appearing in a database. If the data is continuously added to the original transaction database, then the database size becomes larger and mining the entire process would take high computation time, hence this scheme will mine only the updated portion of the database. It will use previous mining results to avoid unnecessary calculations. High-utility item sets can be generated from UP-Tree efficiently with only two scans of original databases. This can not only decrease the overestimated utilities of PHUIs but also greatly reduce the number of candidates. This scheme overtake other individual algorithms substantially in term s of execution time, especially when databases contain lots of long transactions or low minimum utility thresholds are set, by the use of a two efficient strategies DNU and DNN. Results show that the methods significantly improved performance by reducing both the search space and the variety of candidates

REFERENCES

[1] Vincent S.Tseng, Bai-en S hie, Cheng-Wei Wu, and Philip S.Yu,"Efficient Algorithms for Mining High Utility Itemsets from Tran sactionalDatabases"IEEE Transactions on Knowledg e And Data Engineering, Vol.25, No.8, AUGUST 2013 .

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", In *Proc. 20th Int'l Conf. Very LargeData Bases* , pp. 487-499, 19 94.

[3] J.Han, J.Pei, and Y.Yin," Mining Frequent Patterns", Proc.ACM-SIGMOD Int'l Conf.Management of data, pp. 1-12, 2000.

[4] C.H. Cai, A.W.C. Fu, C.H. Cheng, and W.W. Kwong, "Mining Association Rules with Weighted Items," Proc. Int'l Database Eng. and Applications Symp. (IDEAS '98), pp. 68-77, 1998.

[5] W. Wang, J. Yang, and P. Yu, "Efficient Mining of Weighted Association Rules (WAR)," Proc. ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD '00), pp. 270-274, 2000.

[6] Y. Liu, W. Liao and A. Choudhary, "A fast high utility itemsets mining algorithm," In Proc. of the Utility-Based Data Mining Workshop, 2005.

[7] Yao H and Hamilton H J, "Mining itemset utilities from transaction databases", IEEE Data & Knowledge Engineering, pp. 59: 603-626, 2006

[8] Jieh-shanyeh, Yu-Chiang Li and Chin - Chen Chang, "Two-Phase Algorithms for a Novel Utility-Frequent mining Model', PAKDD Workshop, pp. 433-444, 2007".

[9] C.F. Ahmed, S.K. Tanbeer, B.-S. Jeong and Y.-K. Lee, "Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 12, pp. 1708-1721, Dec. 2009.

[10] C.F. Ahmed, S.K. Tanbeer, B.-S. Jeong and Y.-K. Lee, "Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 12, pp. 1708-1721, Dec. 2009.

[11] V. S. Tseng, C. J. Chu and T. Liang, "Efficient Mining of Temporal High Utility Itemsets from Data streams," in Proc. of ACM KDD Workshop on Utility-Based Data Mining Workshop (UBDM'06), USA, Aug., 2006.

[12] V.S. Tseng, C.-W. Wu, B.-E. Shie, and P.S. Yu, "UP-Growth: An Efficient Algorithm for High Utility Itemsets Mining," Proc. 16th ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD '10), pp. 253-262, 2010

[13] Frequent Itemset Mining Implementations Repository, http:// fimi.cs.helsinki.fi/, 2012.