ISSN:0975-9646

A Novel Approach of THUI (Temporal High Utility Item sets) Algorithm on On-Shelf Web Log Data

D.N.V.S.L.S. Indira [1] , JYOTSNA SUPRIYA P [2], S.Narayana[1]

[1]*Department of Computer Science & Engineering,*
Gudlavalleru Engineering College,
Gudlavalleru, Krishna Dt., AP., India

[2]*Department of Computer Science & Engineering,*
Swarnandhra Institute of Engineering & Technology
Narsapur-534275, WGDT, AP., India

Abstract – **Web Utility mining has recently been an emerging topic in the field of data mining and so is the web mining, an important research topic in database technologies. Thus, the web utility mining is effective in not only discovering the frequent temporal web transactions & generating high utility itemsets, but also identifying the profit of webpages. For enhancing the web utility mining, this study proposes a mixed approach to the techniques of web mining, temporal high utility itemsets & On-shelf utility mining algorithms, to provide web designers and decision makers more useful and meaningful web information. In the two Phases of the algorithm, we came out with the more efficient and modern techniques of web & utility mining inorder to yield excellent results on web transactional databases.**

Keywords - WTOS, PWTU, Temporal high-utility itemsets, On-Shelf data.

I. INTRODUCTION

Utility mining has recently been an emerging topic in the field of data mining. It finds out high-utility itemsets by considering both the profits and quantities of items in transactions. In real applications, however, utility mining may have a bias if items are not always on-shelf. On-shelf utility mining is then proposed, which considers not only individual profit and quantity of each item in a transaction but also common on-shelf time periods of a product combination.

Web transactions are now of great importance as we need to buy products which may not be available locally to our area and need to communicate with the websites or the companies across the world with more profits. These are even more useful for the business men who try to import products of multiple numbers from companies of other countries. Such web transactional utilization is gaining more importance now-a-days in this world of communications. This gave us the inspiration in choosing this topic of web transactional utility mining.

When a businessman (person) orders particular needed products of specified quantity , he may also need to have the knowledge of discovering the items or itemsets which are of high utilities & profits in web transactions across the world. Now, the utilities of the products, their usage int earlier days/weeks/months can also be displayed on the webpage along with the product details & the details of its profit when compared to that of other companies by the company's website itself giving more knowledge about the product to its customers visiting the website. This would add more interest to the customer to visit that company's website more confidently & would dare to order undoubtedly. The company's website also gains good recognition among the customers across the world.

II. RELATED WORK

Association-rules mining [1] is an important issue in the field of data mining due to its wide applications. Agrawal *et al.* first proposed the most well-known algorithm, namely Apriori, for mining association rules from a transaction database [1] . Traditional association rules are, however, derived from frequent itemsets, which only consider the occurrence of items but do not reflect any other factors, such as price or profit.

Besides, seasonal or on demand data mining has emerged and attracted much attention in these years because of its practicality. For example, assume there is an association rule like "In the summer, customers usually purchase air-conditioner, fridge together". The itemset {air-conditioner, fridge} may be not frequent throughout the entire database, but may be with a high frequency in summer. Mining time-related knowledge is thus interesting and useful.

According to the above scenario, in the past, Lan *et al.* thus proposed a new kind of patterns, namely high on-shelf utility itemsets [2], which could take the selling periods of a set of products into consideration.

In this paper, we propose to apply the same above [3] algorithm to be applied for the web log data, especially to those that are transactional in nature. All such weblog records are to be applied this [3] algorithm in order to utilize the web data much more effectively. This paper also states the two methods in which this algorithm can be applied : one at the itemset or transactional level (utility values given to the itemsets generated in a single webpage); other at the web page level (utility values given to the web pages in the log). But, to this paper, we limit the description only to the former.

III. PROBLEM DEFINITION

To the best of our knowledge, that is the first paper concentrating on several profitable & interesting features of a web page indicating the transactional weighted utilities of the items or itemsets of a company's web-based transactions. Many of the authors concentrated on the concepts like transactions of a super market , web traversing patterns, adding utilities to the products/webpages etc., But our paper is a mixed approach to these concepts together as it deals with both the webpages as well as the transactional details of it.

The terms transaction & utility sounded slightly different in our paper in comparative to others. From the Zhou's et. al. & Yu-cheng's et. al. researches, the meaning of a transaction includes either a hit of a webpage or time period spent on it and the data therefore involved the details of URL, protocols, the date & time, hit ratio and also the confidence & support values etc., for the calculation of the utilities of the webpages.

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 ie., group of items) of specified quantities. A single webpage may be a single product. One complete transaction hence includes one visit to the site with all products of different cost prices ordered multiplied to their quantities specified. Now, this resembles the market basket problem.

In our sense, a web transaction is said to occur not just by hitting a web page of a company's product details or spending some time on it; but by ordering or paying for a product and thus making some profit to the company. Unless a transaction occurs, the utility cannot be enhanced to the product on a webpage. Hence, our contribution concentrates not on how many times a web page is hit by the customer, or how much time he spent on it; but rather on how much profit he makes to the company by placing an order to or buying its product & the enhancement of utility value to the respective product.

Now, we apply the two-phase algorithm on the web log details obtained from the transactions of a single website in a day utilizing a system clock for seasonal & periodical processes. The visited pages of ordered products can be collected from the web log records as they are automatically noted in the log of a system when used by the customer.

In this web transactional log, the database of products is divided as on-shelf & off-shelf items which are otherwise called temporal data items. Here, the term temporal is not just similar to that applied market basket problem ie., hourly-based transactions of itemsets. But, the term is used in analogous to seasonal items which last for a certain long periods ie., of months together. For example, the air-conditioners, refrigerators much often preferred to be ordered just prior to summer season while the room warmers, microwave ovens, etc., preferred before or in the winter seasons. Another example is the computer systems, computer cabinets& chairs are preferred by many educational institutions, other offices or organisations when the financial year begins. These type of items are considered as on-shelf items.

One web site of a company's products may contain thousands of web pages. If it is a product display website for a single company, each product information can be displayed in web page.If that company has nearly 50000 products (approx.)then that company website has nearly 50000webpages. The items(products) in the website when placed seasonally, the size of data bae on which the process of calculating the utility values gets automatically reduced and the executions becomes faster. The On-shelf products in one season become the off-shelf items in another season. Few items that are non-seasonal are considered as all-time available or general items.

Now the web transactional log consists of several transactions occurred in one whole day, where each is a transactional visit to that website by a single user/customer. So, one complete transaction includes different products in different web pages of different quantities ordered by a single customer in one single visit to the site. This transaction may not be of minutes rather of hours together. Hence, the time- periods & their partition as in THUI-mine algorithm are not considered here.

The web transactions though is conceptually same as that of the market basket , it differs in certain aspects like the time for transactions, frequency of buying products available ina web page , the type of payment & details etc.,

IV.EXPERIMENTAL RESULTS

A..WTOS :

The Web Transactional On-shelf Table is known to us where the total period is divided into no of on-shelf perids. Consider 4 on-shelf time periods with 5 items in the database is shown in the table1 below:

	Wp1	Wp2	Wp3	Wp4
A	1	0	1	1
B	1	1	0	1
C	0	1	0	1
D	0	0	0	1
E	1	1	1	1

Table 1 : An example of the WTOS table.

The on-shelf & off-shelf statement of product A is simply known from the bit string (1,0,1,1)Similarly for B,C,D,E the bit strings are (1,1,0,1), (0,1,0,1), (0,0,0,1) & (1,1,1,1) respectively. The itemset {A,D} has the on-shelf bit string is (0,0,0,1), which is obtained by applying AND operation between the bit strings of A &D that means both the items A& D are collectively found during the timeperiod wp4 only.

In this way, the seasonal items of the on-shelf periods and common periods for the combination of items can easily be extracted from the above table. This table of on-shelf availability of items can directly be displayed on the website itself. The very technique of AND operation between bit strings of different items, web on-shelf periods and transactions can be applied on the data for calculating the on-shelf utilities at various stages.

B. Phase I :

To find out the Periodical total web- transactional Utility(PWTU), we achive this in 3 stages.
Stage 1 :- Item-level utility
Stage 2 :- Transactional-level utility
Stage 3 :- Web- Period level utility
To undergo all these processes, we first gather the date of transactions & items of a particular time period ie., the entire data of items are divided according to their respective on-shelf timeperiods and is maintained in a tabular format as described below.

Let the no. of divided on-shelf periods be 4 as in the above table of WTOS ie., wp1,wp2,wp3 &wp4. Let us now consider any one of the four on-shelf periods say,wp1. This web on-shelf time period wp1, if supposed to be a season like summer or winter or so, the total transactions in the complete seasons are tobe considered which is quite difficult task to do, so, for simplicity, let us consider the season as no of days. Now we calculate the utility of all the transactions occurred in one day and the accumulate the values with each other days's transactions till the end of the season to obtain complete periodical total transactional utility.

A day's Total transactional utility is calculated by considering the data of transactions occurred in that day using the following data table :

Transactional data of day d1 of wp1:
In this wp1, the available itms on-shelf are only A,B & E only and are placed in the table 2 with their quantities of the respective transactions.
The subjective values of the available items are placed in another table3 which tells about the profit per unit item.

	T1	T2	T3
A	20	0	2
B	0	0	4
E	12	1	7

Table 2 : A Transaction database of the on-shelf items

Item	S(i)
A	3
B	10
E	6

Table 3 : The utility table

1) Item-level utility : In this stage, we calculate the utility values of individual items of different transactions as
$u(A,T_3)$ =2 x3 =6
u(A) = u(A,T1) + u(A,T3) = 20x3 +2x3 =66.
Defn: $u(i_p)=\sum u(i_{p,} T_j)$
Where $u((i_, T_j) = s(i) \times q(i,T_j)$, ie., q(i) is quantity & s(i) is the profit per unit.

2) Transactional-level utility: The itemlevel utility is now improved to transaction level utility for we get only the utilities of only the single itemsets in the earlier stage. This is now upgraded using the definitons for itemsets.

Utility of item/itemset,X in transaction T_j is

$$u(X, Tj) = = \sum_{I \text{€} X} u(i_{p,} T_q)$$

And the Transaction utility is

$$tu(T_q) = \sum_{T_q \text{€} D_y} u(i_{p,} T_q)$$

Where D_y is the respective day of web time period.

For ex, tu(T1) = u(A,T1) + u(E, T1)
= 20 x 3 + 12x 6 = 132.
When the utility is calculated at transaction-level, the itemsets can easily be identified for the high transactional values. This can be understood from the table4 below :

TID	Transactional Utility (tu)
T1	132
T2	6
T3	88

TABLE 4 : An example for the calculated transaction utilities

The high utility itemset sequences can be identified easily by setting a minimum utility threshold ε', which is user-specified value generally can be taken as the minimum utility value.This process is done in the phase II.

3) Web Period level utility: Daily total web-transaction utility value is obtained by adding all transactional utilities in that particular day in item-wise or transaction-wise also.

$$DTWU(X,D_y) = \sum_{T_j \text{€} D_y} tu(T_j)$$

$$\text{Where,} \quad tu(T_j) = \sum_{I \text{€} T_j} tu(i, T_j)$$

This can be extended to the period level by adding the utilities of all the days' total transactions. And is denoted by PWTU table 5 as below:

Period	Periodical Web Total Transaction Utility
pwtu1	97
pwtu2	53
pwtu3	112

TABLE 5 : An example of the PWTU table.

C.. Phase II:

In Phase II, one database scan is required to select the high web utility sequences from high transaction-level utility sequences identified in Phase I. The number of the high web transaction-level utility sequence is small when ε' is high. The number of the high transaction-level utility sequence is small when □' is high.

A novel method, namely ***THUI*** **(Temporal High Utility Itemsets) –*Mine was*** proposed by **V.S. Tseng et al** in [7]**,** for mining temporal high utility itemsets from data streams efficiently and effectively. The novel contribution of ***THUI-Mine*** is that it can effectively identify the temporal high web utility itemsets by generating fewer temporal high transaction-weighted web utilization 2-itemsets such that the execution time can be reduced substantially in mining all high utility itemsets in data streams. In this way, the process of discovering all temporal high web transactional utility itemsets under all time windows of data streams can be achieved effectively with limited memory space, less candidate itemsets and CPU I/O time. This meets the critical requirements on time and space efficiency for mining data streams.

In the phase II, this novel method among several others available, is chosen for its proven efficiency among others and also its suitability to our present problem definition (ie. the kind of temporal web transactional data we worked on). The experimental results show that *THUI-Mine*[7] can discover the temporal high utility itemsets with higher performance and less candidate itemsets compared to other algorithms under various experimental conditions. Moreover, it performs scalable in terms of execution time under large databases. Hence, *THUI-Mine* is even more promising in our paper, for its mining capability temporal high utility web transactional itemsets in web data streams or logs.

CONCLUSION AND FUTURE WORK

In this paper, we presented "high on-shelf web utility mining", which introduced the concept of "utility" into on-shelf web log. As utility measures the "interesting" or "usefulness" of a webpage, thus satisfies the Web Service Providers in quantifying the user preferences of ease in web data transactions. Hence, we explored a Two-Phase algorithm that discovered high on-shelf utility data on web pages highly efficiently, in which both the phases are carried out with effective algorithms and became responsible in giving us the effective results. We also demonstrated the interesting areas we observed, as well as their significance to the decision making process.

On-shelf utility mining considered not only individual profit and utility of each item in a web transaction but also common on-shelf time periods of a product combination. In this study, a new on-shelf web utility mining algorithm was preferred in order to speed up the execution efficiency for mining high on-shelf utility web transactional itemsets.

The experimental results also showed that the proposed high on-shelf utility approach had good impact when compared to the other traditional utility mining approaches. In the future, we will attempt to handle the maintenance problem of high on-shelf utility mining of transactions at the webpage level. Besides, the results from on-shelf utility mining on web transactional log are independent of the order of transactions. Another kind of knowledge called sequential patterns depends on the order of transactions.

We will also extend our approach to mining out this kind of knowledge in the future. A number of interesting problems are open to discuss in the future research. For example, the accuracy, effectiveness and scalability of the proposed idea applied to larger databases need to be evaluated. Other factors can also be explored as utility in web trasactional pattern mining. Besides, how to combine frequency and utility together to improve web transactional ease is still a problem need to be studied.

REFERENCES

1. R. Agrawal and R. Srikant, "Fast algorithm for mining association rules in large databases," The 20th International Conference on Very Large Data Bases, pp. 487-99, 1994.
2. G. C. Lan, T. P. Hong, and Vincent S. Tseng. "A two-phased mining algorithm for high on-shelf utility itemsets." The 2009 National Computer Symposium, pp. 100-5, 2009.
3. G. C. Lan, T. P. Hong, and Vincent S. Tseng. "Reducing Database Scans for On-shelf Utility Mining". IETE Tech Rev 2011, vol 28,no. 2, pp.103-12, 2011
4. Yu-Cheng Chen and Jieh- Shan Yeh. "Preference utility mining of web navigation patterns." IET International Conference on Frontier Computing. Theory, Technologies & Applications (CP568) Taichung, Taiwan, pp.49-54, 2010.
5. H. Yao, H. J. Hamilton,and C.J. Butz, "A foundational approach to mining itemset utilities from databases," *Proceedings of the 3rd SIAM International Conference on Data Mining*, pp. 482-486, 2004.
6. L. Zhou, Y. Liu, J. Wang, and Y. Shi, " Utility-based web path traversal pattern mining", *Proceedings of the 7th IEEE International conference on Data Mining Workshops*, pp. 373-378, 2007.
7. V. S. Tseng, C.J. Chu, T. Liang, "Efficient Mining of Temporal High Utility Itemsets from Data streams", *Proceedings of Second International Workshop on Utility-Based Data Mining, August 20, 2006.*