A Conceptual Approach to Temporal Weighted Item set Utility Mining

Jyothi Pillai
Senior Lecturer
Bhilai Institute of Technology
Durg-491001, Chhattisgarh,
India

O.P. Vyas
Professor
Indian Institute of Information Technology, Allahabad,
India

Sunita Soni
Reader
Bhilai Institute Of Technology,
Durg-491 001, Chhattisgarh,
India
University

Dr. Maybin Muyeba
Senior Lecturer
Dept. of Computing and Mathematics
Manchester Metropolitan

ABSTRACT
Conventional Frequent pattern mining discovers patterns in transaction databases based only on the relative frequency of occurrence of items without considering their utility. Until recently, rarity has not received much attention in the context of data mining. For many real world applications, however, utility of itemsets based on cost, profit or revenue is of importance.
Most Association Rule Mining (ARM) algorithms concentrate on mining frequent itemsets from crisp data and recently, use of discrete utility values. Unfortunately, in most real-life applications, use of discrete valued utilities alone is inadequate. In many cases where these values are uncertain, a fuzzy representation may be more appropriate.
An interesting extension to ARM is including the temporal dimension. Traditional ARM does not use time; however, the real application data always changes with time. Discovering temporal association rules that hold in given time intervals may lead to more useful information. As real-world problems become more complex, temporal rare itemset utility problems become inevitable to solve. To handle uncertainty, temporal itemset utility mining with fuzzy modeling allows item utility values to assume fuzzy values and be dynamic over time. In this paper, we present a theoretical conceptual approach to Temporal Weighted Itemset Utility Mining.

Keywords:
Association Rule Mining, Utility, Temporal, Frequent Pattern Mining

1.0 INTRODUCTION
An important area of data mining research deals with the discovery of association rules. The mining of Association rules for finding the relationships between data items in large datasets is a well-studied [22].
The basic bottleneck to association rule mining is rare item problem. Most association rule mining algorithms implicitly consider the utilities of the itemsets to be equal [1]**.** A utility is a value attched to an item depending on its evaluation., e.g. if coke has support 20 and its profit is 2%, cookies may have support 10 but with a profit of 20%, then utility of cookie is higher than coke}. Similarly, most association rule algorithms [22] use simple *support–confidence model* i.e. first find all *frequent itemsets* with support of at least *minsup* and then generate all association rules with confidence of at least *minconf*. The frequency of an itemset may not be a sufficient indicator of interestingness because it does not reveal the *utility* of an itemset, which can be measured in terms of cost, profit, or other expressions of user preference.
In many applications, some items appear very frequently in the data, while others rarely appear. If frequencies of items vary, two problems may be encountered –(1) If minsup is set too high, then rules of rare items will not be found (2) To find rules that involve both frequent and rare items, minsup has to be set very low. This may cause combinatorial explosion.
Another feature worth considering is handling temporal data. Temporal association rule mining discovers valuable relationships among items in the temporal database [23]**.** This incorporation is especially necessary if we want to extract useful knowledge from dynamic domains, which are time varying in nature.
Further, the modeling of imprecise and qualitative knowledge, as well as the transmission and handling of uncertainty at various stages are possible through the use of fuzzy sets. Fuzzy logic is capable of supporting, to a reasonable extent, human type reasoning in natural form [24]**.** In many applications, for example, data streams, use of discrete valued utilities alone is inadequate. In many cases where values are uncertain, a fuzzy representation may be more appropriate. The utility value itself can be discrete or fuzzy and thus demands a new theoretic conceptual approach to the problem, which has been proposed in this paper.
The rest of paper is organized as follows. In section 2, we discuss some related works. In section 3, we detail our

proposed theoretical conceptual approach. Section 4 presents conclusion & future work.

2. Related Work

Utility mining is now an important association rule mining paradigm. In [1], a good foundational and theoretical model to utility itemset mining is introduced where a utility table UT<I,U> is defined by items I and their utilities U computed for each transaction and termed local utility of a transaction. This approach is improved in [5]. Some utility approaches have considered performance enhancements to enable handling of large candidate sets, for example in [6] which is adopted theoretically from [1].

Some of the more recent works in temporal ARM arises from utility data streams e.g the FTP-DS algorithm [7], THUI (Temporal High Utility Itemsets)-Mine [21] and some others [2] [6]. In these approaches, a time window is usually employed to mine temporal utility itemsets efficiently. We note that even in these dynamic approaches, where dynamic means data changes with respect to time, the utility or weight of an item does not change with time or during any given time window. The approach in [8] does not consider the utility of an item changing within the same period i.e. they allow dynamicity of a utility value between time partitions and not within time partitions. Further, [8] only considers discrete-valued utilities.

In many applications, for example stock markets or data streams, use of discrete-valued utilities alone is inadequate. In cases where the values are uncertain, a fuzzy representation may be more appropriate. However, we note that even in normal temporal association rule mining where the data is analyzed from a static point of view but with a given temporal measure, there needs to be a theoretical foundation on the problem of temporal utility mining with weighted utility measures (fuzzy or non-fuzzy) redefined.

In [1], a foundational approach is given from a static ARM approach (without consideration of temporal or fuzzy features). We build on this foundation.

3. Problem Statement

In this section, a problem definition for utility mining from a temporal and fuzzy perspective is presented with formal definitions and examples to illustrate the approach.

DEFINITION 3.1 (*Utility Mining*) Let D be a given transaction database with a set of transactions T, a set of items $I = \{i_1, i_2, i_3, \cdots, i_m\}$ where each item $i \in I$ has a set of time measures with defined granularities $P = \{p_1, p_2, p_3, .., p_{|k|}\}$ and utilities defined $U = \{u_1, u_2, u_3, .., u_{|k|}\}$. Utility mining is the problem of finding all itemsets in a transactional database above a minimum utility threshold τ satisfying a given a minimum support s and confidence c.

DEFINITION 3.2 (*Utility Table*) A utility table UT is a triple $UT = < \mathrm{I}, \mathrm{U}, \mathrm{P} >$ where each item i has some utility value u_j in $U = \{u_1, u_2, .., u_{|k|}\}$ for some $k > 0$. These utility values are set accordingly and correspond to a given set of time partitions or periods $P = \{p_1, p_2, .., p_{|k|}\}$ i.e. u_j corresponds to some p_j. Note that some $u_j = u_i$ at time p_i and p_j respectively, and if $u_1 = u_2 = .. = u_{|k|} = 1$ at all times, the problem becomes a standard ARM problem.

These definitions may apply more appropriately to discrete-valued utilities [8]. In [8] the authors present the first dynamic utilities in ARM where a time partition p (e.g. transaction 1 to transaction k) has one set of utility values without allowing item utilities within the transaction block $p_1, .., p_k$ to change. Naturally, our approach is more realistic in real applications because not all utilities may remain the same in a given time partition. Alternatively, utility values should be allowed to have continous or even fuzzy values. For example, for stock exchange data, share prices of one particular product may remain constant in same period of time that one other product's share price changes ten times. Utility mining frameworks ought to reflect such dynamicity within the time partitions being considered.

We now define item utility, itemset utility, transaction utility and the corresponding utility function according to [1] as follows:

DEFINITION 3.3 (Transaction Utility) The transaction utility value in a transaction, denoted $u_o(\mathrm{i_p}, \mathrm{T_q})$, is the value of an item $\mathrm{i_p}$ in a transaction $\mathrm{T_q}$. The transaction utility reflects the utility in a transaction database.

DEFINITION 3.4 (External Utility) The external utility value of an item is a numerical value $\mathrm{s}(\mathrm{i_q})$ associated with an item $\mathrm{i_q}$ such that $\mathrm{s}(\mathrm{i}_q) = U(i_q)$, where U is a utility function, a function relating specific values in a domain to user preferences.

DEFINITION 3.5 (Utility Function) A utility function $f(o, s)$ is a two variable function, that satisfies:

(1) $f(o, s)$ monotonically increases in $f(o, s)$ for fixed o.

(2) $f(o, s)$ monotonically increases in $f(o, s)$ for fixed s.

DEFINITION 3.6 The utility of an item i_q in a transaction T_q, denoted $U(i_q, T_q)$, is $f(o(i_q, T_q), s(i_q))$, where $o(i_q, T_q)$ is the transaction utility value of i_q, $s(i_q)$ is the external utility value of i_q, and f is a utility function.

Given these definitions, however, the dynamicity of utility values over time can add to the interestingness of the association rules discovered. Here we can further distinguish between cyclic, semi-cyclic and non-cyclic utilities.

DEFINITION 3.7 ***(Cyclic utility)*** A cyclic utility $u^c{}_j$ of an item u_j is when an item's utility repeats in given periods of time.

For example, utility of alcoholic beverages may change repeatedly between festive and non-festive periods. In contrast, utility of a fridge may not be cyclic in festive seasons.

DEFINITION 3.8 ***(Semi-Cyclic utility)*** A semi-cyclic utility $u^{s-c}{}_j$ of an item u_j is when an item's utility, on average, can appear repeatedly.

DEFINITION 3.9 ***(Non-Cyclic utility)*** A non-cyclic utility $u^{n-c}{}_j$ of an item u_j is when an item's utility never repeats but assumes different values all the time.

Non-cyclic utilities are those where items are not following a repeating pattern and hard to predict. We note that cyclic and semi-cyclic utility can be useful for investors in the stock market. Also, every type of utility however can be captured by a consumer price index for benchmarking the economy's weighted average of prices of a basket of consumer goods.

Following formulations from [1], we present an example of a typical utility mining problem where the database records quantitative items i.e. numbers of items bought per transaction. Table 1 shows a transaction database with item quantities sold per transaction. Suppose that external utilities for the three items are A=4, B=7 and C=2.

TID	*A*	*B*	*C*
1	20	5	50
2	10	4	12
3	8	0	25
4	3	1	0
5	3	3	8
6	4	0	10
7	1	2	1
..	..	..	..
12	..	..	..

Table 1. Quantitative Transaction Database

Item	Utility	P_u
A	{4, 5, 4, 15}	{<p1>, <p2>, <p3>,<p4>}
B	{7, 4, 10, 2}	{<p1>, <p2>, <p3>,<p4>}
C	{2, 1, 5, 8}	{<p1>, <p2>, <p3>,<p4>}

Table 2. Temporal Static Utility values

In transaction 1, the transaction weight of item A, (A, T_1)= 20, item B, (B, T_1)=5 and item C, (C, T_1)=50.

P_u=1				P_u =2				P_u =3				P_u =4			
ITEM	Utility			ITEM	Utility			ITEM	Utility			ITEM	Utility		
	TID				TID				TID				TID		
	T1	T2	T3		T4	T5	T6		T7	T8	T9		T10	T11	T12
A	4	4	4	A	5	5	5.3	A	4	4	4	A	15	15	15
B	7	7	8	B	4	4	4.5	B	10	10	10	B	2	2.1	3
C	2	2	3	C	1	1	1.3	C	5	5.2	5	C	8	8.4	8

Table 3: Temporal Dynamic Utility values within transaction periods (Continuous value)

Given the external utilities of the three items A, B and C as $\{\mu_A, \mu_B, \mu_C\}$ = (4, 7, 2) in period p_1, the weighted utility of item A in transaction T_1 is $u(A,T_1) = \mu_A * \nu_A = 4*20 = 80$. Similarly, utility of C in transaction T_6 is $u(A,T_6) = \mu_C * \nu_C = 2*10 = 20$ and so on. The simple example emphasizes two factors, internal and external utility in determining utility of an item. Since an itemset X is made up of item combinations from different transactions T_X where X occurs, the local utility of an item $x_i \subset X$ is $l(x_i, X) = \sum_{q \in X} (\mu_I * \nu_I, T_q) = \mu_I * \sum_{q \in X} (\nu_I, T_q)$ as μ_I is constant with $T_q \supseteq X$.

In table 1, local utility of B in period p_1 is $l(B,B)$=7*(5+4+1+3+2)=105.

Similarly utility of a k-itemset (e.g. AB is a 2-itemset) in the same period is

$u(X) = \sum_{i=1}^{k} l(x_i, X) = l(A, AB) + l(B, AB)$ =

4*(20+10+3+3+1)+105=4*37+105=253, ignoring transactions 3 and 4.

In table 2 item has static utility values for all transaction in a given time granularity. For example in time granularity p_1 the utility value of item A is 4 for all transactions. An example of dynamic utility within a time partition is shown in table 3.

In table 3 items have different utility values for different transaction in a given time granularity and some are dynamic with a given time period e.g. B has dynamic within T1 to T3. For example in time granularity $P_u = 1$, the utility of item A is 4 in Transactions T1 to T3. Infact, this item has a cyclic utility value in $P_u = 1$ and $P_u = 3$.

	TID	*A*	*B*	*C*
P1	**1**	20	5	50
	2	10	4	12
	3	8	0	25
P2	**4**	1	2	3
	5	6	4	3
	6	13	2	4
P3	7	5	10	2
	8	3	5	1
P4	9	1	3	2
	10	30	50	5
	11	3	12	15

Table 4 shows Transaction Database.

Let Dataset D have transactions $T = (t_1, t_2, .., t_n\}$ and a set of items X= { x_1, x_2,….., x_m }. Let P= {P_1, P_2,….., P_k} be a set of time granularities and $|x_i|$ is the number of such items.

Definition 3.10. Utility Temporal Support is sum of the utilities of the item, presented in all transactions, divided by the total number of transaction in a particular time period $[P_i, P_k]$ and defined as:

$$UTS(x_i) = \frac{\sum_{q=1, x_i \in t_q \in [Pi, P_k]}^{k} |x_i| * U(x_i, t_q)}{|D|}$$

For example, using tables 1 and 3, x_1=A, q=1, 2; where |D|=6, Utility Temporal Support of item A is calculated as:

UTS(A)=(20*4+10*4+8*4+3*5 +3*5+4*5.3) /6
= 33.9

Definition 3.11. Utility Temporal Confidence is the ratio of sum of votes satisfying both AUB to the sum of votes satisfying A. It is formulated as:

$$UTC(A \rightarrow B) = \frac{UTS(A \cup B)}{UTS(A)}$$

Utility Temporal Confidence of (AB) in $[P_1, P_2]$ is calculated as:

UTC (AB) = UTS(A)+UTS(B))/UTS(A)
=(33.9+(5*7+4*7+1*4+3*4)/6)/33.9
=(33.9+13.2)/33.9
= 1.39

In reality, we assume that time granularities are set according to the application domain but in the example of table 4, we set discrete values for simplicity.

Table 4 Transaction Database

A partitioned database is given in table 4 with time granularities (P1, P2, P3, P4) each of which represents variable time periods.

This is typical in real world applications where at particular times some items have higher demands, profitability etc than others.

We can see from table 4 that in period 2, the best utility item is A with 13, followed by B and then C, but in period 3, item B has the highest utility of 10 etc.

Apart from discrete utility values, we can use fuzzy values (Low, Medium, high) to describe utilities. This aspect represents an imprecise value usually given as an estimate of utility of an item. Table 5 shows this for four periods.

Item	*U*	*P*
A	{L, M, H, M}	{1, 2, 3, 4}
B	{M, M, H, L}	{1, 2, 3, 4}
C	{L, L, M, M}	{1, 2, 3, 4}

Table 5. Fuzzy temporal utility table

To illustrate further, table 5 can be a translation of discrete values to fuzzy values where linguistic values used are Low [0, 10], Medium as [11,40] and high as [41-100]. To simplify our understanding, we use table 2. For item A, {4, 5, 4, 15} are utilities for periods 1, 2, 3 and 4 but {L, M, H, M} according to the fuzzy sets defined in table 5.

However, in contrast to table 5 where utilities are allowed to be dynamic within one time period per transaction, an aggregation of fuzzy support and confidence is needed to compute actual supports for such an item. As boundary values in portioning quantitative data is error-some, a normalization of fuzzy support and confidence is more realistic.

A fuzzy representation of fuzzy utilities is shown in table 6.

P	*TID*	*Trans*	*Fuzzy Utility*
			a b c d
1st	T1	a, b, c	L H H L
	T2	b, c	L H H L
	T3	a, b, c, d	H M H M
	T4	a, b, d	H H H L
2nd	T5	b, c, d	H H H L
	T6	a, c, d	L H L L
	T7	b, c, d	M H H L

Table 6. Dynamic fuzzy utility values within transactions in given periods.

Let Dataset Z have transactions $T = (t_1, t_2, .., t_n\}$ and a set of items $X = (x_1, x_2, .., x_m\}$. $P = \{p_1, p_2, .., p_k\}$ is a set of time granularities. Let $< X, A >$ be the itemset–fuzzy set pair, where X is any attribute x_i and A is the set of fuzzy sets a_i with fuzzy temporal utility membership degree μ_i for each itemset $< X, A >$.

The fuzzy utility support of an item x_i in a period P_k for every transaction t_q is given by

$$FTUS(x_i) = \frac{\sum_{k=1, t_q \in P_k}^{|P_k|} \left(\sum_{j=1}^{|t_i|} (\sum_{i=1}^{m} [\mu(x_i) * (t_q . | x_i |)]) \right)}{| Z |}$$

where $| x_i |$ is the weighted value of an item e.g. 20 for item A in P1 of table 1.

We an also calculate fuzzy temporal utility confidence in the same way.

This approach sets a new direction where real applications can use different utilities varying over time, moreover, the flexibility of this approach renders potential for further work and experiments on large real-world data sets with varying time granularities.

4 Conclusions and Future Work

Our work presents a new foundational approach to temporal weighted itemset utility mining where item utility values are allowed to be dynamic within a specified period of time, unlike traditional approaches where these values are static within those times. Moreover, our approach incorporates a fuzzy model where utilities can assume fuzzy values on the other hand. A Conceptual model has been presented that allows development of an efficient and applicable algorithm to real world data and captures real-life situations in fuzzy temporal weighted utility association rule mining.

REFERENCES:

[1] Yao, Hong, Hamilton, H., and Butz, C. J. 2004. *A Foundational Approach to Mining Itemset Utilities from Databases*, Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida, pp. 482-486.

[2] Chu, C., Tseng, V. S., and Liang, T. 2008. *An efficient algorithm for mining temporal high utility itemsets from data streams*. *J. Syst. Softw.* 81, 7 (Jul. 2008), 1105-1117

[3] Hu, J., Mojsilovic, A. *High-utility Pattern Mining: A Method for Discovery of High-utility Item Sets, Pattern Recognition*, Vol. 40, 3317-3324.

[4] Ale, J. M. and Rossi, G. H. (2000). *An Approach to Discovering Temporal Association Rules*. In Proceedings of the 2000 ACM Symposium on Applied Computing, Vol.1, J. Carroll, E. Damiani, H. Haddad, and D. Oppenheim, Eds. SAC '00. ACM Press, New York, NY, pp 294-300

[5] Yao, H. and Hamilton, H, J. 2006. *Mining Itemset Utilities from Transaction Databases*, Data and Knowledge Engineering, 59(3): 603-626

[6] Liu, Y., Liao, W., and Choudhray, A. 2005. *A Fast High Utility Itemsets Mining Algorithm*. Proceedings of the Utility-Based Data Mining Workshop.

[7] Teng, W. G., Chen, M. S., and Yu, P. S. 2003. *A Regression-Based Temporal Pattern Mining Scheme for Data Streams*. Proceedings of the 29th International Conference on Very Large Databases, pp 93-104.

[8] Ahmed, C. F., Tanbeer, S. K., Jeong, B-S, and Lee, Y-K. 2008. *Handling Dynamic Weights in Weighted Frequent Pattern Mining*, IEICE Trans. Information and Systems, Vol. E91-D:2578-2588.

[9] Han, J., Pei, J. and Yiwen, Y. 2000. *Mining Frequent Patterns Without Candidate Generation*. Proceedings ACM-SIGMOD International Conference on Management of Data, ACM Press, pp1-12.

[10] Coenen, F., Leng, P. and Ahmed, S. 2004. *Data Structures for association Rule Mining: T-trees and P-trees*. IEEE Transactions on Data and Knowledge Engineering, Vol 16, No 6, pp774-778

[11] Yun, U. 2007. *Mining lossless closed frequent patterns with weight constraints*. Know.-Based Syst. 20, 1 (Feb. 2007), 86-97

[12] Ning, H. and Yuan, S-C. 2006. *Temporal Association Rules in Mining Method*, First International Multi-Symposiums on Computer and Computational Sciences - Volume 2 (IMSCCS'06) pp. 739-742

[13] Verma, K., Vyas, O. P. and Vyas,R. 2005. *Temporal Approach to Association Rule Mining Using T-Tree and P-Tree, Machine Learning and Data Mining in Pattern Recognition*, 651-659, LNS Volume 3587

[14] Cheng-Yue Chang, Ming-Syan Chen, Chang-Hung Lee, *Mining General Temporal Association Rules for Items with Different Exhibition Periods* , Second IEEE International Conference on Data Mining (ICDM'02),

[15] Yo-Ping Huang; Li-Jen Kao; Sandnes, F.-E. , 2005. *A prefix tree-based model for mining association rules from quantitative temporal data, Systems, Man and Cybernetics*, 2005 IEEE International Conference on Volume 1, Issue , 10-12 Oct. 2005 Page(s): 158 - 163 Vol. 1

[16] Edi Winarko and John F. Roddick,. 2005. *Discovering Richer Temporal Association Rules from Interval-Based Data, Data Warehousing and Knowledge Discovery*, LNCS 3589, 315-325.

[17] Kriegel, H-P et al. 2007. *Future Trends in Data Mining, Data Mining and Knowledge Discovery*, 15:87–97

[18]Gaber, M. M,, Zaslavsky, A. and Krishnaswamy, S. 2005. *Mining data streams: a review*. SIGMODRecords

[19] Lu, S., Hu, H. and Li, F. 2005. *Mining weighted association rules*. Intelligent Data Analysis, 5(3):211–225.

[20] Agrawal ,R. Srikant, R. 1994. *Fast Algorithms for Mining Association Rules*, In: Proceedings of 20th International Conference on Very Large Databases, Santiago, Chile, pp. 487-499.

[21] Yingjiu Li, Peng Ning, X. Sean Wang , Sushil Jajodia R. 2003. *Discovering calendar- based temporal association rules*, Data & Knowledge Engineering volume 4,Elesvier publisher, Volume 44, pp 193-214.

[22] Agrawal R., Imielinski T., and Swami A., "*Mining association rules between sets of items in large databases*", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, pages 207–216, 1993.

[23] Zhai Liang, Tang Xinming, Li Lin, Jiang Wenliang, *Temporal Association Rule Mining based on T-Apriori Algorithm and its typical application*, Proceedings of International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion, 2005.

[24] Sushmita Mitra, Sankar K. Pal, Pabitra Mitra, *Data Mining in Soft Computing Framework: A Survey*, IEEE Transactions On Neural Networks, VOL. 13, NO. 1, January 2002