High Utility Pattern Mining on Data Streams in Data Mining

Bhavya Shukla[1], Dr.R.K Gupta[2]
[1, 2] Department Of Computer & Science
[1, 2] Madhav Institute of Technology and Science, Gwalior,India.

Abstract- *A data stream is continuous, rapid, infinite sequence of data. Mining Frequent pattern in stream data is extremely difficult as results of information are often scanned just one occasion. Because of this reason ancient approach like frequent pattern mining cannot be used for economical information streaming. Discovery of high utility item sets like profit is thought as High utility pattern mining. It is the growing field of knowledge stream mining. During this survey paper we have a tendency to offer an outline of finding consecutive patterns through the construct of window mechanism which can end in the doable outcomes which will differ from ancient approaches that manufacture level –wise-candidate and test generation.*

Keywords- high utility mining, data streams, sliding window.

I. INTRODUCTION

wide-spread use of distributed information systems results in the construction of large data collections in business, science and on the web. These data collections contain a wealth of information, which however needs to be discovered. Businesses can learn from their transaction data more about the behavior of their customers and therefore can improve their business by exploiting this knowledge. Science can get from observational data (e.g. satellite data) new bits of knowledge on research questions. Information on web can be analyzed and exploited to optimize information access [1]. Data mining presents methods that permit extracting from massive data collections unknown relationships among the data items which are useful for decision making. Along these lines data mining generates novel, unsuspected interpretations of data.

According to the Gartner Group [2], "data mining is the system of discovering significant new correlations, patterns and traits by using sifting via huge amounts of data put away in repositories, making utilization of pattern recognition sciences in the same class as measurable and numerical programs." "Data mining is the examination of (in general huge) observational data sets to look for out unsuspected connections and to compress the data in novel techniques. Which can be each understandable and priceless to the data owner". "Data mining is an interdisciplinary subject uniting frameworks from machine learning, pattern awareness, records, databases, and representation to handle the obstacle of know-how extraction from large data bases".

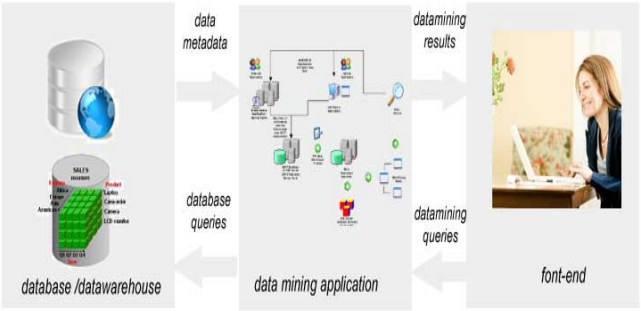

Figure 1. Data mining

II. HIGH UTILITY PATTERN MINING

The Item set Share strategy [3] considers non-binary frequency values of an object in every transaction. Share is the percentage of a numerical whole that is contributed through the items in an item set. These authors defined the problem of finding share frequent item sets and compared the share and support measures to illustrate that share measure can give helpful data about numerical qualities that are connected with transaction items, which is unrealistic by utilizing just the support measure. This approach cannot rely on the downward closure property. Heuristic ways to search out item sets with share values above the minimum share threshold had been developed. They assume the same medical treatment for different patients (different transactions) will have different levels of effectiveness. They can't keep up the downward closure property however they build up a pruning methodology to prune low utility item sets in view of a weaker anti-monotonic condition. The hypothetical model and meanings of high utility pattern mining (HUPM)were given. This approach, known as MEU (mining with expected utility), cannot preserve the downward closure property. They used heuristics to check whether or not an item set must be considered as a candidate item set. Later, comparable makers proposed two new algorithms, UMining and UMining_H, to figure high utility patterns.In UMining, a pruning process founded on the utility higher certain property is used.

UMining_H was designed with one more pruning technique centered on a heuristic process. Moreover, these techniques don't fulfill the downward closure property and could overestimate excessively numerous patterns. They likewise endure generation-and test technique issue.

In light of the meanings of the Two-Phase algorithm was produced to discover high utility item sets. The authors have defined the transaction weighted utilization (TWU) and utilizing it they proved that it is feasible to preserve the downward closure property. At the last output, the Two-Phase algorithm decides the actual high utility item sets from the high transaction weighted utilization item sets.

A. High average utility pattern mining

In traditional utility pattern mining [4], the summation of items' utilities in a pattern is expanded alongside the expansion of the pattern's length. In order to obtain fair utilities of patterns according to their lengths, high average utility mining has been studied. The TWU downward closure property utilized as a part of customary high utility mining can't be specifically connected to high average utility mining. For this reason, TPAU can generate all candidate patterns without losses of high average utility patterns through utilizing this overestimation system. It finds excessive normal utility patterns without an additional database scan for calculating actual traditional utilities of patterns by means of storing all the item utility comprehension of precursors in each node.

B. High utility pattern mining with multiple minimum support

The MHU-development algorithm [4] developed on CFP Growth++ was once proposed to mine high Utility Frequent Item sets (HUFIs) through given that rarity and utility know-how of each and every object. The algorithm constructs its data constitution named MHU-Tree by way of a single database scan. In the interim, not at all like CFP-Growth++, this calculation stores both backings and TWU estimations of things in MHU-Tree. By means of utilizing the info stored in MHU-Tree, the algorithm performs growth procedures to be able to generate a set of candidate patterns with helps and utilities pleasing the rarity and utility conditions on the foundation of the pattern growth approach. In this manner, MHU-Growth requires an extra database filter for checking genuine utilities of competitors comparably to the past HUPM algorithms. Therefore, we need to apply such length factors into the mining process to obtain more meaning full information from databases.

C. Objective of high utility pattern mining

The main objective of high-utility item set mining is to find all those item sets having utility greater or equal to user- defined minimum utility threshold. Every item in the item sets is associated with an additional value, called internal utility which is the quantity (i.e. count) of the item An external utility is hooked up to associate item, showing its quality (E.g. price) with such a utility-based information, high utility item sets (patterns) area unit mined , together with those satisfying the minimum utility. Mining high utility item sets is far tougher than discovering frequent item sets, as a result of the elemental downward closure property in frequent item set mining doesn't hold in utility item sets. information streams. The ascension of continuous information has several challenges to store, Computation and communication capabilities in computer system. The high speed information wants some techniques to perform real time extraction of hidden info. In information streams, information enters at a high speed rate. The system won't be capable for storing the complete stream information. therefore it stores solely a little quantity of knowledge.

Data mining techniques facilitate to seek out attention-grabbing patterns from uncommon sort of information. data processing techniques play an important role in several giant organizations. however today several new techniques and algorithms area unit used for information streams while not dropping the events, formula area unit designed with clear concentrate on the event of essential information. Window mechanism usually, it's assumed that the info area unit static, unless expressly changed or deleted by a user or application.

Database queries area unit dead once issued and their answers mirror this state of the info. However, rising applications, like sensing element networks, period web traffic analysis, and online money commerce, need support for process of infinite information streams.

The fundamental assumption of an information} stream management system is that new data area unit generated frequently, creating it impossible to store a stream in its entireness. At best, a window of recently arrived information could also be maintained, which means that previous information should be removed as time goes on. what is more, because the contents of the slippy windows evolve over time, it is sensible for users to raise a question once and receive updated answers over time.

III. LITERATURE SURVEY

Heungmo Ryang (2016) et al gift that HUPM has been studied as a crucial subject inside the world of pattern mining as the way to satisfy necessities of the many real-world applications that require to method non-binary databases as well as item importance like marketing research. though many list-based algorithms to get high utility patterns while not candidate generation are prompt in recent years, they need an outsized variety of correlation operations.

Our technique facilitates economical mining of high utility patterns with the projected indexed list by effectively reducing the full variety of such operations. Exploratory results on genuine and synthetic datasets demonstrate that the proposed calculation mines high utility examples more productively than the state-of-the-art algorithms [5].

Dhyanesh K.Parmar (2013) et al present that Main purpose of using different data mining techniques is to find novel, potentially useful patterns which can be useful in real world applications to derive best knowledge and using it into useful way. In first area we offered fundamental terms like Data mining, frequent pattern mining, Sequential pattern mining, Time interim Sequential Pattern Mining and utility mining. In second section we summarizes some most important prior study work involving Sequential pattern mining, Time interval centered sequential pattern mining and utility mining. In final part we concluded the survey work via suggesting some future instructional materials for discovering some robust sequential rules that may infer from extracted patterns [6].

Shuning Xing (2015) et al present that UP-Growth is one of the most discussed HUPM algorithms based on the data structure of UP-Tree. However, the process of construction trees needs to scan database several times and spends much time in calculating. To solve these problems, an improved construction process of UP-Tree is proposed by introducing a Fast Utility Tree (FU-Tree). In this method, we introduce the Link Queue to reduce the number of scanning the original database and adopt prefix utility to minimize the overestimated utility. The theoretical analyses and experimental results show that FU-Tree outperforms UP-Tree in the time consumption of construction trees, and enhances the efficiency of mining high utility item sets [7].

Vijay Kumar Dwivedi (2013) et al present that HUPM is valuable for identification of probably the most valuable item sets in incremental databases. We advocate algorithm for establishing IHUP_TWU tree structure making use of trie structure. We additional propose mining algorithm for picking out most valuable item sets through using trie structures. Experiments show that algorithms are efficient as compared to other existing algorithms [8].

Jingyu Shao (2015) et al present that In latest years, the significance of identifying actionable patterns has end up increasingly well-known so that decision-support actions can be inspired by the resultant patterns. A typical shift is on identifying high utility rather than highly frequent patterns. For that reason, high Utility Item set (HUI) Mining ways have emerge as particularly widespread as good as faster as and more riskless than earlier than. Nonetheless, the present study focus has been on improving the efficiency at the same time the coupling relationships between objects are neglected. It is primary to study object and item set couplings inbuilt in the knowledge. here we introduce a new framework for mining actionable high utility association rules, called Combined Utility-Association Rules (CUAR), which aims to find high utility and powerful association of item set combinations incorporating item/item set relations. The algorithm is turned out to be effective per trial results on each real and algorithm [9]

Vijay Kumar Dwivedi (2013) et al reward that prime utility sample mining is useful for identification of the most useful item sets in incremental databases. We advocate algorithm for setting up IHUP_TWU tree structure utilizing tire constitution. We extra propose mining algorithm for opting for most useful item sets with the aid of using trie buildings. Experiments show that algorithms are effective as compared to other present algorithms [10].

Jingyu Shao (2015) et al gift that in recent years,the importance of making a choice on actionable patterns has become increasingly famous so that decision-help moves will also be encouraged via the ensuring patterns. However, the present research focal point has been on making improvements to the efficiency even as the coupling relationships between gadgets are neglected. To this finish, right here we introduce a brand new framework for mining actionable high utility association principles, called mixed Utility-organization rules (CUAR), which ambitions to seek out high utility and robust organization of item set combinations incorporating object/item set family members. The calculation is ended up being proficient per test impacts on both genuine and efficiently and datasets [11].

Jerry Chun-Wei Lin (2016) et al present that just lately, HUPM has been extensively studied. Many procedures for HUPM were proposed in latest years, however most of them intention at mining HUPs for granted for their frequency. This has the main hindrance that any combination of a low utility object with a very high utility pattern is viewed as a

HUP, despite the fact that this combination has low affinity and comprises items that rarely co-occur. As a consequence, frequency should be a key criterion to decide on HUPs. To deal with this hindrance, and derive high utility interesting patterns ((HUIPs) with strong frequency affinity, the HUIPM algorithm was proposed.

However, it recursively constructs a sequence of conditional bushes to supply candidates after which derive the HUIPs. This method is time-consuming and could result in a combinatorial explosion when the minimum utility threshold is set fairly low. On this paper, an efficient algorithm named quick algorithm for mining DHUPs with FDHUP is proposed to comfortably discover DHUPs with the aid of considering that each the utility and frequency affinity constraints. A huge experimental learn indicates that the proposed FDHUP algorithm substantially outperforms the state-of-the art HUIPM algorithm in phrases of execution time memory utilization, and scalability [12].

Guo-Cheng Lan (2014) et al present that Recently, high utility sequential pattern mining has been an emerging preferred hassle because of the consideration of quantities, profits and time orders of items. Consequently, the highest measure is thoroughly used to simplify the utility calculation for subsequences in mining.. The indexing technique can also be developed to swiftly find the critical sequences for prefixes in mining, and consequently unnecessary search time will also be lowered. Ultimately, the experimental outcome on a couple of datasets exhibit the proposed process has good performance in both pruning effectiveness and execution efficiency [13].

Jerry Chun-Wei Lin (2015) et al reward that HUPM is an rising topic in recent years alternatively of association-rule mining to notice extra intriguing and useful information for choice making. Many algorithm had been developed to find high-utility patterns (HUPs) from quantitative databases without on the grounds that timestamp of patterns, exceptionally in contemporary intervals. A flexible minimum-length technique with two specified lifetimes may also be designed to find more efficient UDHUPs based on a users' specification. Analyses are completed to evaluate the efficiency of the proposed two algorithms as far as execution time, memory utilization, and number of created UDHUPs in a number of genuine and synthetic datasets [14].

Heungmo Ryang (2016) et al present that Processing changeable data streams in actual time is among the principal issues within the data mining area due to its vast purposes similar to retail market analysis, WSNs, and stock market prediction. In addition, it is an interesting and challenging problem to deal with the stream data since not only the data have unbounded, continuous, and high speed characteristics but also their environments have limited resources. HUPM mean while is likely one of the primary study themes in pattern mining to beat principal drawbacks of the usual framework for frequent pattern mining that takes just double databases and same item essentialness into thought. This strategy conducts mining approaches by reflecting characteristics of true databases, non-double amounts and relative significance of items. As a result, they consume a huge amount of execution time, which is a significant performance issue since a rapid process is necessary in stream data analysis. On this paper, we propose an algorithm for mining high utility patterns from resource limited environments through effective processing of data streams in order to remedy the issues of the overestimation-centered ways.. The proposed tree is rebuilt by our upgrading technique with brought overestimation utilities down to safeguard redesigned move data at whatever points the present window slides. Our technique additionally has a main influence on expert and intelligent techniques in that it may provide users with more meaningful information han traditional analysis approaches by reflecting the characteristics of real world non-binary databases in stream environments and emphasizing on recent data. Comprehensive experimental results demonstrate that our algorithm it will provide customers with extra meaningful know-how than traditional evaluation strategies by way of reflecting the characteristics of actual world non-binary databases in flow environments and emphasizing on latest information. Comprehensive experimental results reveal that our algorithm beats the predominant sliding window-based one in expressions of runtime efficiency and scalability [15].

IV. PROBLEM DEFINITION

In this subsection, we present an example of utility mining. Let us take a finite set of distinct items and a transaction where Ti be a subset of I . In this we are calculation the internal utility which is calculated by value of item in transaction into its quantity value i.e iu(i,T). And then external utility is calculated by multiplying the internal utility into its item profit i.e eu(i).

Definition 1 Item utility of transaction called Utility value is calculated by multiplying the internal utility of item into the external utility.

Definition 2 Transaction utility of item is calculated which is defined as total profit of the transaction denoted by T.

Consider an example of Data Streams in fig.1 and of profit table in Tabel.1 in which we will calculate the Item utilities and Transaction utilities.

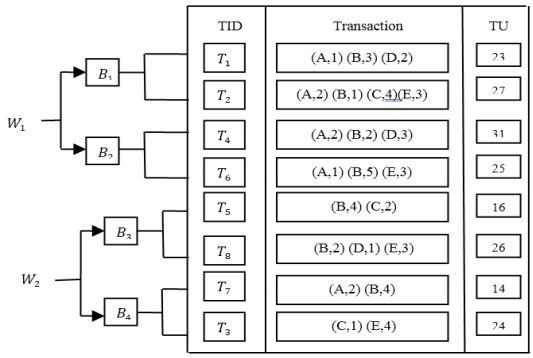

Figure 1. example of data streams

Table 1. profit table

items	A	B	C	D	E
profits	3	2	4	7	5

Def:1(Item utility) In respect to an item, A in T_2, for instance, its external and internal utilities are eu (A)= 3 and in (A, T_2) =2.

Thereby, utilities of A in T_1 is u (A, T_1) =eu (A)×in (A, T_1)
= 3×1 =3
In same way, utilities of other items in T_1 is B,D
u(B, T_1)=2×3=6, u (D, T_1)= 7×2=14
In addition sum of utilities in T_1 is $u(A, T_1) + u(B, T_1) + u(D, T_1)$=14+6+3=23
In other words transaction utilities of T_1 *is* Tu (T_1) =23

Next, we define the problem of mining high utilities pattern in a sliding window over streams data.

Let a data stream
DS= [T_1, T_2,.....T_n] be a sequence of finite transaction and a pattern p be a set of items.
T_2 = eu (A) ×in (A, T_2)=3×2=6
= in (B, T_2)= 2×1=2
= in (C, T_2)= 4×1=4
= in (E, T_2)= 5×3=15
Tu of T_2 = Tu(T_2) =6+2+4+15=27

T_4 = eu (A) ×in (A, T_4) = 3×2 = 6
Utilities of other items in T_4 *is* A,B,D u (B, T_4)= 2×2=4, u(D, T_4)= 7×3=21
Tu(T_4) =6+4+21=31
T_6 =u(A, T_6)=3×1=3
u(B, T_6)= 5×2=10
u(E, T_6)= 3×4=12
$Tu(T_6)$ = 12 + 3 + 10=25
T_5 =u(B, T_5)=2×4=8
u(C, T_5)=4×2=8
$Tu(T_5)$ =16
T_8 =u(B, T_8)=2×2=4
u(D, T_8)=7×1=7
u(D, T_8))=5×3=15
$Tu(T_8)$ =4+7+15=26

T_7 =u(A, T_7)=3×2=6
u(B, T_7)=2×4=8
=8+6=14
$Tu(T_7)$ =14
T_3 =u(C, T_3) =4×1=4
u(E, T_3)=5×4=20
$Tu(T_3)$ =24

2. Prof def.

Here we are calculating the pattern utilities of the following datasets given

Table 2. example of data streams

	Transactional data base
T_1	(A,4), (B,2), (C,8), (D,2)
T_2	(A,4), (B,2), (C,8)
T_3	(C,4), (D,2), (E,2), (F,2),
T_4	(E,2), (F,2), (G,1)

Table 3. external utilities

Items	A	B	C	D	E	F	G
Unit profits	2	3	1	3	4	4	8

Example: Calculating The high utilities item set, we have u (A,C) in the transactional database=T_1 + T_2
=<(4×2+8×1)>+<(4×2+8×1)>=32
We have eu (A) ×in (A, T_1) =4×2=8
i(B, T_1) =2×3=6
eu(C, T_1) =8×1=8
(D, T_1) =2×3=6
Tu(T_1)=8+8+6+6=28
eu(T_2) ×in(A, T_2)=4×2+2×3+8×1=22
eu(T_3) ×in(C, T_3)=4×1+2×3+2×4+2×4=26
eu(T_4) ×in(E, T_3)=2×4+2×4+1×8=24

Table 4.

	Transactional data base	Total
T_1	(A,4), (B,2), (C,8), (D,2)	28
T_2	(A,4), (B,2), (C,8)	22
T_3	(C,4), (D,2), (E,2), (F,2),	26
T_4	(E,2), (F,2), (G,1)	24

Calculate pattern utility

Each pattern P has utility values in a transaction T_i, u(p, T_i) $\sum u(ip, T_i)$,where ip ϵ p and p⊆ T_i *in a batch* B_j , u(p, B_j)= $\sum u(p, T_i)$ where T_i ϵ B_j and p⊆ T_i and in a window W_k, U_{wk} (P)= $\sum u(p, B_j)$, where B_j ϵ W_k.
Utility of {AB} in T_2 is u (AB, T_2) =u (A, T_2) +u (B, T_2) =6+2=8

Each batch consists of two transactions and thus utility of {AB} in a batch B_j that includes two transactions T_1 and T_2 is u (AB, B_1) = u(AB, T_1)+ u (AB, T_2)=9+8=17
Moreover

For e.g. utility of {AB} in a window W_1 is
u W_1 (AB)=u(AB, B_1)+ u(AB, B_2) =17+23=40
u(AB, B_2)=u(AB, T_4)+ u(AB, T_6) =10+13=23

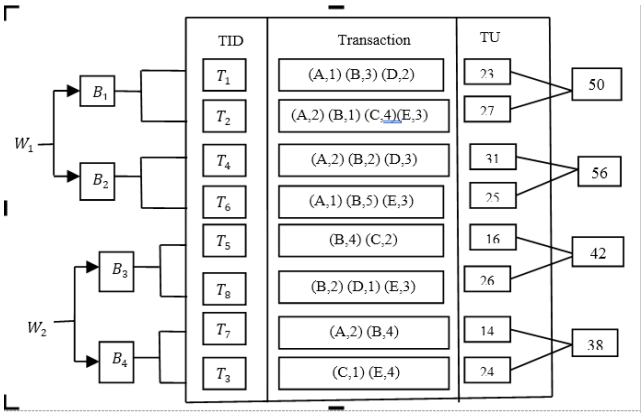

Figure 2.

And meanwhile, the total utility of transaction in W_2= uW_2 (AB,T_7) =14
If a minimum threshold is 22%, minutil W_2 =14×0.22=3.08

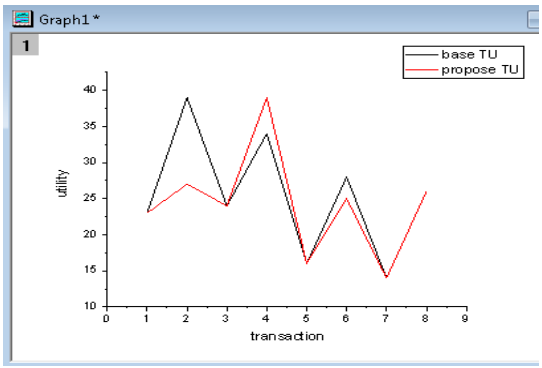

Figure 3.

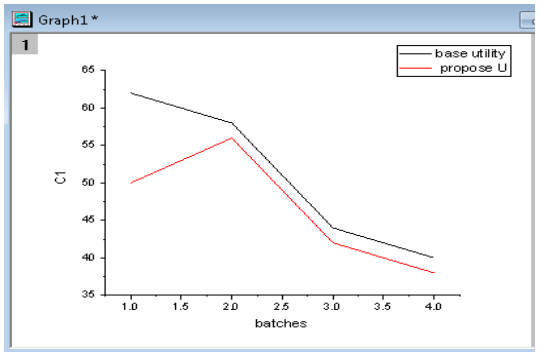

Figure 4.

V. CONCLUSION

We tackled the problem level wise candidate generation using high utility sequential pattern mining over data streams using sliding window mechanism. We have proposed an approximation algorithm, called PSHU Growth algorithm, to discover an efficient tree in which hierarchy is reduced by decreasing the over estimation utilities and increasing the accessing time by decreasing the overall memory utilization and space used.

REFERENCES

[1] Ranshul Chaudhary, Prabhdeep Singh and Rajiv Mahajan," A SURVEY ON DATA MINING TECHNIQUES", International Journal of Advanced Research in Computer and Communication EngineeringVol. 3, Issue 1, January 2014, pp 5002-5003.

[2] Meenakshi Sharma," Data Mining: A Literature Survey", International Journal of Emerging Research in Management &Technology, Volume-3, Issue-2 2014,pp 1-4.

[3] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong and Young-Koo Lee," HUC-Prune: an efficient candidate pruning technique to mine high

utility patterns", Appl Intell (2011) 34: 181–198.

[4] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong and Young-Koo Lee," HUC-Prune: an efficient candidate pruning technique to mine high utility patterns", Appl Intell (2011) 34: 181–198.

[5] Heungmo Ryang and Unil Yun," Indexed list-based high utility pattern mining with utility upper-bound reduction and pattern combination techniques", Department of Computer Engineering, Sejong University, Seoul, Republic of Korea Springer 2016.

[6] Heungmo Ryang and Unil Yun," Indexed list-based high utility pattern mining with utility upper-bound reduction and pattern combination techniques", Department of Computer Engineering, Sejong University, Seoul, Republic of Korea Springer 2016.

[7] Shuning Xing, Fangai Liu, Jiwei Wang, Lin Pang and Zhenguo Xu," Utility Pattern Mining Algorithm Based on Improved Utility Pattern Tree", 2015 8th International Symposium on Computational Intelligence and Design, 2015 IEEE, pp 258-261.

[8] Vijay Kumar Dwivedi," Disk-resident High Utility Pattern Mining: A Trie Structure Implementation", 2013 IEEE, pp 44-49.

[9] Jingyu Shao, Junfu Yin, Wei Liu and Longbing Cao," Mining Actionable Combined Patterns of High Utility and Frequency", 2015 IEEE.

[10] Jerry Chun-Wei Lin, Wensheng Gan, Philippe Fournier-Viger, Tzung-Pei Hong and Han-Chieh Chao," FDHUP: Fast algorithm for mining discriminative high utility patterns", springer 2016.

[11] Guo-Cheng Lan, Tzung-Pei Hong, Vincent S. Tseng and Shyue-Liang Wang," Applying the maximum utility measure in high utility sequential pattern mining", Expert Systems with Applications 41 (2014) 5071–5081.

[12] Jerry Chun-Wei Lin, Wensheng Gan , Tzung-Pei Hong and Vincent S. Tseng," Efficient algorithms for mining up-to-date high-utility patterns", Advanced Engineering Informatics 2015

[13] Heungmo Ryang and Unil Yun," High utility pattern mining over data streams with sliding window technique", Expert Systems With Applications 57 (2016) 214–231.