FREE for Reading & Publication
Diamond Open Access

Analysis of Top-K Utility Pattern Mining Framework

S.S. Naghate*, S.S. Sherekar, V.M. Thakare

P.G. Dept. of Computer Science and Engineering, Sant Gadge Baba Amravati University, Amravati,
Maharashtra 444602, India

***Corresponding author: E-mail: naghatesnehal21@gmail.com**

DOI: 10.5185/amp.2020.030404

A new algorithm of utility-list structure has been proposed in this paper which based on the framework of top-k utility mining analysed. There are many techniques which extracts top keyword within search space relates with various algorithms. Like, High Utility Pattern Mining, which is detection of high utility itemsets in a transactional database; provides the fundamental task in database and data mining community, such as quantity, cost, weight and profits concern, to extract remarkable knowledge and efficient database patterns. Different from the support based mining models; the utility oriented mining framework integrates the utility theory to provide more informative and useful patterns. These could not be directly performed on the utility mining techniques with the help of proposed method in this paper would find a top-k search space itemsets. This paper is focused on analysis of many identical high top-k utility patterns mining algorithm, such as mining for Top-K High Utility Itemsets, Solutions to Utility Big Data Analysis, Negative Sequential Patterns Mining, Efficient High Pattern Mining with Tighter Upper Bounds and Tighter Upper Bound including Average-Utility for Mining High Average-Utility Patterns. However there are some issues that need to resolve. These are discussed in this paper and efficient proposed the analysis of the various utility mining techniques referring to the various frameworks.

Introduction

In the early days, to dominant of frequent itemset mining along with different mining techniques were examine in data mining. Also in association rule mining suggested the set of frequent itemsets for mine. In which the minimum support threshold must be less from occurrence of frequencies are greater than has applied in various real and synthetic applications. To evaluate the association rules for which intimacy was greater than minimum threshold value. Finding interesting patterns like Knowledge Discovering Database (KDD) is essential for variety of applications such as financial data analysis, retail system, genome analysis [**1,2**]. The utility patterns can provide more valuable and more precious product depending on decision-making than the traditional frequency model [**3**]. As the main purpose of data mining is to extract required itemsets which are require at runtime, and potentially useful information from large databases. It can acknowledge frequently many utility itemsets from transactional databases, Top keyword high utility pattern mining significant character in data mining which is an important research issue in data mining [**5**]. Itemset mining is a useful pattern search technique to find correlations between different items as in association rule mining. As more database are collected, the utility companies are now seeking new analytics tools and techniques to address their emerging data mining issues [**6,7**].

The output of the top-k high utility itemsets that is the itemsets that have the highest utility in the transaction database taken as input and it will give the top-k utility itemsets with minimum utility threshold. This paper addresses all of the Top-k utility pattern mining itemsets using research transactional databases and newly arrived datasets at runtime.

Background

Many studies on Top-k utility itemsets mining have been done to develop the utility mining algorithms in past years. Such algorithms are:

An enhanced high utility pattern approach (EHUPA) had improved the Performance of high utility pattern for mining itemsets. This system contains name of the item as node and after calculating transaction utility and transaction weighted utility, the item sets having less utility than predefined minimum threshold utility are identified [**1**].

The traditional high utility pattern algorithms compared to the traditional frequently pattern mining which had suggested precious information. According to the length of attributes of each datasets the high utility pattern mining would considered the causes effect for their utilities obtains. These combined utility with the help to mine frequent itemsets when the length of an itemset is too long [**2**].

In comparison of High utility sequential patterns mining and negative sequential patterns mining, the sequences with once occurred or the uniquely occurred items where considered. It would be considered as the wastage of data and then it plays an imperative act in many real-life applications, such as big data analysis in retail, businesses, and smart courts [**3,4**].

As the extraction of reveal patterns with high-utility has become an important data mining task which engrossed on cluster making. The clustering will defiantly help in these. Since the HUIM could give automatically same sets of items in discrete databases [**5**].

Top-k utility patterns is set too high, no result patterns are found while too small value makes an enormous number of result patterns which cause inefficiencies in terms of computation time and memory usage. Thus, it requires multiple trials, more time for execution and to find an appropriate minimum support value referring to the minimum threshold value [**6**].

Big data technologies are aimed at processing high-volume, high-velocity, and high-variety data and extracting business intelligence for insight and decision making. Utility big data typically undergo a number of transformations during their lifecycle, ranging from collection, storage, analysis, to presentation etc. [**7**].

The paper is organized as follows:

Section I Introduction. **Section II** discusses Background. **Section III** discusses previous work. **Section IV** discusses existing methodologies. **Section V** discusses attributes and parameters and how these are affected on mining techniques. **Section VI is** proposed method. **Section VII** is experimental tests carried out**.** **Section VIII** is outcome and result. **Section IX** is conclusion. Finally **Section X** is future scope of this analytical paper.

Previous work done

In research literature, many data mining technology have been studied to provide various Top-k utility mining algorithms and improve will the top-k utility itemsets with minimum utility threshold.

Arun kumar M. S. *et. al.,* (2018) [**1**] proposed a survey which conducted the work on regular pattern in data mining. In these, the patterns would extract referring to a newly arrived data and enhanced works of Periodic pattern mining. As the author said these makes a countable itemsets produced from a huge datasets.

JIMMY MING-TAI WU *et. al.,* (2018) [**2**] proposed new pruning strategy in EHAUPM, based on the calculation of nodes and the branches in the searching tree. It significantly outperforms on the state of art algorithm and the minimum number of threshold operations according the runtime itemsets had considered.

TIANTIAN XU *et. al.,*(2018) [**3**] proposed the calculation of negative sequences of utility mining and defines the problem from that mining HUNSP along with an efficient method. It will extract high utility sequential patterns from the negative sequential patterns. To generate this, a new data structure PNU-List is to store the related information about negative sequential patterns and efficiently calculate HUNSC's utility.

JERRY CHUN-WEI LIN *et. al.,* (2017) [**5**] proposed a method in which attributes measuring in downward closure property in transaction-weighted performs and the high utility patterns would obtained. As the more would weight of the product, the more would be memory usage.

Vincent S. Tseng *et. al.,* (2016) [**6**] proposed a new efficient top-K high utility itemset mining algorithm; TKUL-Miner using a new framework of the utility-list structure based top-k high utility itemset mining. The utility threshold would be continuously variable and new utility-list structure suggested. There are several strategies are implemented to raise the border minimum utility threshold rapidly.

Jun Zhu *et. al.,* (2016) [**7**] proposed developing a standard-based software framework to address key utility big data issues and faster development of big data analytical applications. Based on the support key shifting, new big data analytical solutions are often rapidly built and deployed to enhance a utility organization.

Existing methodologies

Many techniques and algorithms have been implemented over the last several decades. There are different methodologies that are implemented i.e. on analysis of different high top-k utility pattern mining algorithm, such as mining for Top-K High Utility Itemsets, Utility Big Data Analysis, Negative Sequential Patterns Mining, and High Average-Utility Pattern with Upper Bound for Mining.

TUB-HAUPM Algorithm

TUB-HAUPM called two new tighter upper-bounds first is to greatly reduce the search space for mining. Another one is an addition of upper-bound model. The proposed algorithm discovers the enough value of the top keyword datasets must having a maximum utility in the extracted database which has extract from proposed algorithm. It considered the product of arithmetic average itemsets and addition of upper-bound items. The utility of an item in both transactions is denoted as $u\ (ij;\ Tq)$, and is defined as:

$$u(ij;\ Tq) = q(ij;\ Tq) \text{ x } p(ij)$$

High utility negative sequential pattern mining

HUNSPM algorithm as proposed initially, it mines high utility positive sequential patterns from Q-sequence databases then more and more pattern mining techniques such as USpan, or a new HUSP mining algorithm. In a q-sequence, minimum utility threshold value at last stage of extracting itemsets.

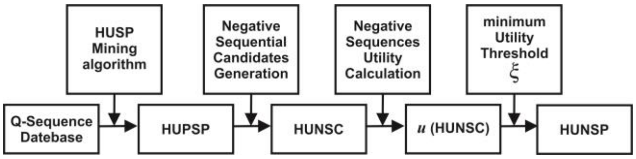
HUSP Mining algorithm
Negative Sequential Candidates Generation
Negative Sequences Utility Calculation
minimum Utility Threshold ξ
Q-Sequence Datebase
HUPSP
HUNSC
u (HUNSC)
HUNSP

Fig. 1. Block diagram of negative sequential pattern mining.

EHAUPM algorithm

In these, during the database scan, the first method is only applied the datasets on a pattern mining and the second strategy is repeatedly applied before catalyzing any new key itemset are designed which is compulsorily greater than 1. These list structure obtained by performing computation and efficient itemsets using a top key set had finds. These transaction utility tu of a transaction Tq is defined as:

$$tu(T_q) = \sum_{i_j \in X} u(i_j, T_q).$$

Algorithms for mining Top-K high utility itemsets

An itemset node of the search tree in the TKUL-Miner is designed to contain more information than that used in other utility-list based algorithms. For an itemset, the TWU indicates sum of transaction utilities of itself in the database. It will find top-k high utility itemsets is the set of the *k* itemsets that have the highest utility in ascending values.

```
Algorithm: TKUL-Miner Algorithm
input : $D$, a transaction database; $minuti;$, $k$.
output: the top-$k$ high-utility itemsets.
Scan $D$ to calculate the TWU of 1-itemsets
$I^* \leftarrow$ each item i such that TWU(i) $\geq minutil$
Sort items in TWU ascending values on $I^*$
Scan $D$ to build the initial utility-list of each item i $\in I^*$ and
  build the EUCS structure
TUL-FirstLevelSearch($\emptyset$, $I^*$, $minutil$, EUCS, $k$)
```

Algorithm1: TKUL-Miner Algorithm

Utility big data analysis

As per the customer requirements, frameworks are comprised of guidance, patterns, shared utility libraries, and collaborative software modules designed to significantly big data analytical application. Utility function *f* is a function of two variables commonly defined as the product of internal and external utility as given as:

$$f\,(x, y){:}\ xp * yp$$

where, *xp* and *yp* noted for utility list *x* and utility list *y* respectively.

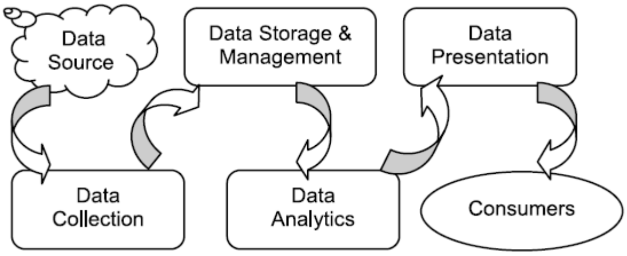

Fig. 2. Life cycle of utility big data.

Analysis and discussion

In HAUPM calculates the high value of threshold, although these cannot trim into more branches to decline the computation, more operations are required for evaluating itemsets. When there is newly arrived datasets enters into the research datasets, the clustering forms and the keyword which has used more search spaces has been suggested as a top keyword referring to the algorithm given by authors [**1,2**].

Utility negative sequential patterns proposed minimum utility threshold ξ for five different datasets, and analyze the performance of HUNSPM algorithm in terms of the running time. It also computes a required time for calculates the sets of similar attributes and forms different types of groups [**3**].

EHAUPM algorithm calculates of each item and the scans the database then for each item in the database, the minimum high average-utility gross the threshold is isolated. The threshold always considered as a minimum value [**5**].

The TKUL-Miner algorithm consists of various parameters which can be calculated from the transaction database and the profit information. The top keyword high utility itemsets is the set of the *k* itemsets that have the highest utility. It is said that the algorithm returns more than *k* itemsets if several itemsets finds the same utility [**6**].

The utility big data application framework is designed to provide developers and end users an open solution development and deployment environment while taking maximum advantage of the underlying infrastructure support. It is a survey of data collection to data presentation in front of customers [**7**].

Table 1. Comparisons between various Top-k utility pattern mining techniques.

Proposed Algorithm Techniques	**Advantages**	**Disadvantages**
TUB-HAUPM Algorithm	The set of frequent itemsets for which affair frequencies are at least the minimum threshold.	It efficiently decreases the number of join operations.
High Utility Negative Sequential pattern mining:	In these pattern growth approaches is to reduce the overall runtime in mining HUSP.	This algorithm cannot support multiple items.
EHAUPM Algorithm	Algorithms were first run on each; the number of patterns can greatly decrease the execution time and memory requirements.	The proposed algorithm can be mine HUSPs in big data, or extending the model to other pattern mining problems.
Algorithms for Mining Top-K High Utility Itemsets	It Greatly raises the border minimum utility thresholds and further good scalability on large datasets.	The algorithm did not report the running time.
Utility Big Data Analysis	The data processing framework shall be readily expendable to respond to the growing computation.	Cloud computing may not be feasible to utility big data analysis.

Proposed methodology

Top-k utility pattern mining is a challenging task as in the recent times, many frameworks have been proposed for top- k itemsets. Top-k utility pattern mining is a technique that finds valuable utility itemsets from a large sequence of database with each itemset having the minimum utility threshold. The representative top- k mining frameworks, high utility calculation, minimum utility itemsets and a minimum utility threshold by summating all the steps of following algorithm.

Steps of algorithm

Algorithm 1: Utility_mining (D, t, B, utilcal, minutil)

Input: A sequence of transaction database D, a minimum utility threshold t, a set of data called batches B, a utility calculations utilcal and a minimum utility itemsets minutil.
Output: The top-k high utility itemsets with minimum utility threshold topk.utility

```
Begin
Scan and Construct D into multiple batches B
    according to ascending entries of database D
Set batches B := utilcal [ ] $\leq$ t
For each utilcal $\rightarrow$ topk.utility(B, minutil,t);
If utilcal $\leq$ t
Then
Utilcal := utilcal + 1
End if
Return utilcal
End for
End
```

Algorithm 2: topk.utility (B, minutil, t)

Input: A sequence of transaction database D, a minimum utility threshold t, a set of data called batches Band a minimum utility itemsets minutil.

```
Begin
D $\leftarrow$ length (B) $\leq$ minutil
For each i $\leftarrow$ length (B)
Do
B $\leftarrow$ B + (minutil $\leq$ t)
Increment i$\leftarrow$ i+1
End for
Return topk.utility
End
```

Diagrammatic representation of proposed method is shown as follows:

Experimental tests

The approach of high utility pattern mining is observing by the purchase quantity and price of each product to discover a set of product generating with high profits. The proposed algorithm is implemented in C# language of visual studio 2010 and executed on computer equipped with an Intel® core™ i3 & i5 processor and running on the 64bit Microsoft windows 7 and 10 operating system. Also the WEKA open source tool might be used for executing the various algorithms on datasets. Experiments were performed on 4 different datasets available on FIMI and UCI repository.

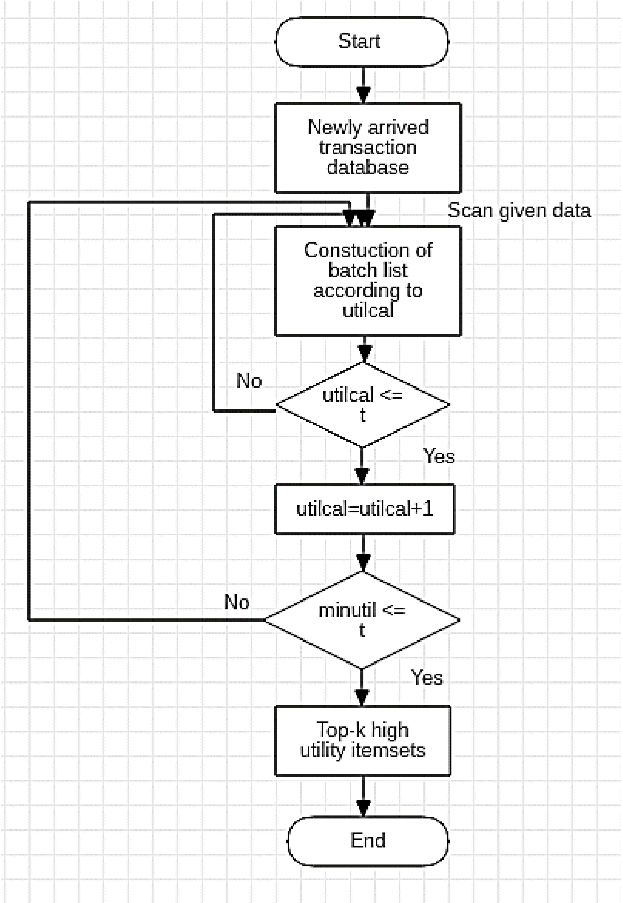
Start
Newly arrived transaction database
Scan given data
Constuction of batch list according to utilcal
No
utilcal <= t
Yes
utilcal=utilcal+1
No
minutil <= t
Yes
Top-k high utility itemsets
End

Fig. 3. Flowchart of proposed methodology.

In this paper, the proposed method performs for the top-k high utility pattern mining when the data is in an information form. With the help of these algorithms, the proposed method calculates the output as the top-k high utility itemsets that is the itemsets that have the highest utility with the minimum threshold in the transaction database taken as input. Since all these datasets are normally used for traditional frequent itemset mining, it had to add transaction and average length values to the characteristics of datasets which is defined in **Table 2**.

Table 2. Characteristics of datasets.

Sr. No.	Datasets	Transaction	Average Length
1	Retail	88,162	10.3
2	BMSPOS	515,597	7.5
3	T10N5D100K	100,00	10
4	UNIFORM_10_50k	50,000	10

Outcomes and results

Also generated a utility table based on minimum utility threshold with the utility values ranging from given values shown in **Fig. 4**. Also, Results of the proposed experiment are shown and the no. of itemsets was generated randomly.

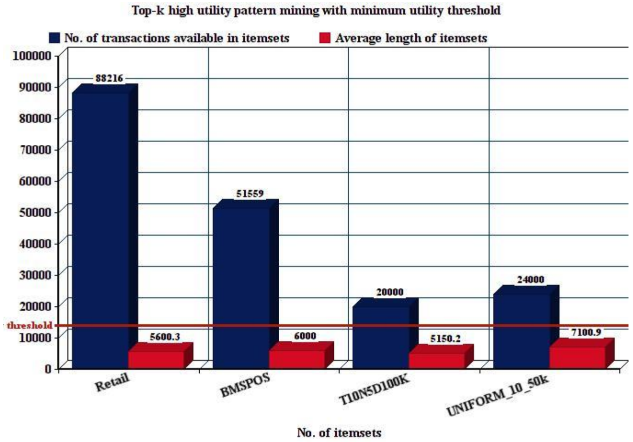

Fig. 4. Distribution of utility value with minimum utility threshold.

Conclusion

To increase the minimum threshold, the proposed algorithm takes itemsets in descending order with first-level searching itemsets. This paper proposed an enhanced High Utility Mining Approach to mine the Top-k High Utility Itemsets with less computation time and less memory space from plenty of unprocessed data. It utilizes the sum of common utilities and the sum of itemset utilities that have zero remaining utilities in order mining effectively. Comparing to all existing methodologies, the standard-based software framework has been proposed which will improve the efficiency and effectiveness by reducing the number of candidates. The scope of improved pattern mining technique for information retrieval of text documents using feature extraction and text mining will be the research papers that provided by the user for summarizing.

Future scope

It is expected that the continuous research and development will eventually result in a number of utility data mining tools. Generally, the proposed algorithm deals with the various real and synthetic datasets and designs mining techniques which will improve the performance and effectiveness of extracted itemsets.

The proposed approach presents a clustering function, which uses minimum time in a mining process with the help of association/ Apriori Algorithm. This approach maintains the datasets in Top Keyword Space Searching Property considering input as an e-commerce datasets and the output will give the top-k utility itemsets with minimum time consumptions and memory used. It might be reduce the computational speed in future and also work on big data.

Acknowledgements

It gives me immense pleasure to express my gratitude to Dr. S. S. Sherekar and Dr. V. M. Thakare, my Supervisors who provided me constructive criticism and a positive feedback during the preparation of this Research Paper. I am indebted to Dr. V. M. Thakare, Head of Department of Computer Science & Engineering at SGBAU and other teaching and non-teaching staff who were always there whenever I needed any help. Without them and their co-operation, completion of this research work would have been inevitable and their presence behind me is totally indispensable.

Keywords

Top-k High utility patterns, TKUL-miner, sequential patterns, frequent itemset mining, knowledge discovery database.

References

1. Sharmila, P.; Dr. Meenakshi, S.; *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, **2018**, *7*, 39.
2. Ming-Tai Wu, Jimmy; Chun-Wei Lin, Jerry; (Member, IEEE), Pirouz, Matin; Fournier-Viger, Philippe; *IEEE*, **2018**, *6*, 18655.
3. Xu, Tiantian; Li, Tongxuan; Dong, Xiangjun; "Efficient High Utility Negative Sequential Patterns Mining in Smart Campus", Special Section on Novel Learning Applications and Services for Smart Campus, **2018**, *Vol. 6*, 23839-23847.
4. Arunkumar, M.S.; Suresh P.; Gunavathi, C.; Preethi, S.; *International Journal of Recent Technology and Engineering (IJRTE)*, **2018**, *7*, 59.
5. Chun-Wei Lin, Jerry; (Member, IEEE), Ren, Shifeng; Fournier-Viger2, Philippe; Hong3, Tzung-Pei; *IEEE*, **2017**, *5*, 12927.
6. Tseng, Vincent S.; (Senior Member, IEEE); Wu, Cheng-Wei; Fournier-Viger, Philippe; Yu, Philip S.; *IEEE Transactions on Knowledge and Data Engineering*, **2016**, *28*, 54.
7. Zhu, Jun; Zhuang, Eric; Fu, Jian; Baranowski, John; Ford, Andrew; Shen, James; *IEEE Transactions on Power Systems*, **2016**, *31*, 2455.
8. Gan, Wensheng; Chun-Wei Lin*, Jerry; Fournier-Viger, Philippe; Chao, Han-Chieh; Yu, Philip S.; *Journal of Latex Class Files*, **2015**, *14*, 1.
9. Pillai, Jyothi; Vyas, O.P.; *International Journal of Computer Applications*, **2010**, *5*, 9.

Authors biography

Ms. Snehal Naghate has completed B.E. Degree in Computer Science & Engineering from Sant Gadge Baba Amravati University, Amravati, and Maharashtra. She is pursuing Master's Degree in Computer Science and Information Technology from P.G. Department of Computer Science and Engineering, S.G.B.A.U. Amravati.

Dr. Vilas M. Thakare is Professor and Head in Post Graduate Department of Computer Science and Engg., Faculty of Engineering & Technology, SGB Amravati University, Amravati. He is also working as a coordinator on UGC sponsored scheme of e-learning and m-learning specially designed for teaching and research. He is Ph.D. in Computer Science/Engg. and completed M.E. in year 1989. He has done his PhD in area of robotics, AI and computer architecture. His area of research is Computer Architectures, AI and IT. He has published more than 150 papers in International & National level Journals and also International Conferences and National level Conferences.

Dr. Mrs. S. S. Sherekar working as a professor in PG department of Computer Science and engineering at SGBAU from 1994 and having 24 years of teaching and research experience. Her area of research is Network security, Image Processing and working as supervisor for Ph.D. guidance. Completed one MRP of UGC one MRP of AICTE is going on. She has published more than 68 research paper in International & National level Journals.