Mathematical Statistician and Engineering Applications
ISSN: 2326-9865
DOI: https://doi.org/10.17762/msea.v70i1.2297

Efficient Algorithms for Mining the Concise and Lossless Representation of High Utility Itemsets

Rakesh Patra

Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Article Info
Page Number: 173-181
Publication Issue:
Vol. 70 No. 1 (2021)

Article History
Article Received: 25 January 2021
Revised: 24 February 2021
Accepted: 15 March 2021

Abstract
Data mining is the process of extracting new, possibly useful information from vast data bases that is not straightforward. Market basket analysis, a kind of data mining used in retail research, is used to analyse client transactions. The association between the things that occur in transactions more frequently was the focus of earlier data mining techniques. They don't take an item's significance or utility into account while often mining an itemset. Utility mining is a new field that has emerged as a result of the limits of common mining goods. When mining, the profitability or utility of an object is taken into account. In a transaction, an item's utility refers to its significance or financial gain. Finding the item set with utility values over the specified threshold is the main goal of mining high utility goods. We give a literature review on several mining methods in this work. The support-confidence framework, used in traditional association rule mining, gives users an objective way to quantify the rules that are important to them. Despite several research being done, the mining operation may perform worse with regard to of processing speed as well as memory. efficiency when consumers are presented with too many high utility itemsets by existing approaches. In this research, the system proposes a unique framework for mining closed high utility itemsets that acts as a lossless yet compact representation of high utility itemsets.

Keywords: Data Mining, Utility Mining, Market Basket Analysis, Client Transaction, Memory Efficiency.

1. Introduction

The practise of extracting non-trivial, heretofore unusual, unknown, and possibly valuable information from huge datasets is known as data mining. In order to find intriguing patterns or connections that would improve knowledge, Furthermore, it is interested in the analysis of enormous amounts of data. Thus, the term "data mining" describes the process of obtaining information from vast volumes of data. Applications for data mining include bioinformatics, genetics, medical, clinical research, education, retail, as well as marketing research. Activities involving data mining integrate methods from artificial intelligence, machine learning, database technology, etc. A basic area of study in data mining is prevalent itemset mining. The itemsets that regularly appear in transactions are known as frequent itemsets.

Finding all of a transaction dataset's frequently occurring itemsets is the aim of frequent itemset mining. Over the past few years, numerous applications have found the challenge of detecting regular patterns in huge databases to be quite beneficial. This work is computationally more expensive, particularly when there are many patterns, and the enormous amount of patterns

mined during the different ways makes it very challenging, allowing the user to select the patterns they find most fascinating. Data mining has been extensively employed in retail research to analyse customer transactions. Market basket analysis, pertaining to the identification of product groups that clients commonly purchase, is one of its well-known applications. Each grocery item has a varied value and price in the real world, and a single client may be interested in purchasing multiples of the same item. Therefore, in order to identify the most value item sets that contribute the most to a retail businesses overall profit, it is not sufficient to detect simply conventional frequent patterns in a database. So we start mining for utilities.

Each item in utility mining has a unit weight thus may be used multiple times during a single transaction. The concept of "utility" describes the significance or usefulness of an item's presence in a transaction as measured by profit, sales, or any other user preference. Internal utility and external utility are the two metrics that make up a transaction database. Among the most intriguing data mining issues with a wide variety of applications is discovering itemsets that have utility that reaches a defined threshold. A collection of transactions in horizontal data format is typically used to construct high utility itemsets.

Since Agrawal et al. released their study on mining association rule in 1991, several strategies have been created and enhanced for locating various sorts of patterns in databases. The method of obtaining data through data mining, a few innovative, nontrivial data points from vast data sources. Data mining is frequently employed in a variety of purposes in diverse industries including medicine, marketing, and so on. This paper offers a thorough analysis of numerous item set mining methodologies and approaches based on utility, frequency, and association rule mining-based research studies. Additionally, it provides a succinct overview of data mining's benefits, a simple explanation of its methodology, a performance evaluation, and recommendations for further study.

Finding intriguing regularities or correlations is done via data mining, a key study field that has grown during the past two decades. Data mining has been heavily utilised in trade research to investigate consumer interactions, or market basket analysis. Data mining is the implicit nontrivial extraction procedure, previously unrecognised, and possibly useful information from data. Knowledge discovery in databases is also known as data mining.

Large data sets are stripped of configured understanding through a method called data mining. It entails communicating the data as an exact reproduction of the dataset's semantic organisation, with the forecasting or classification of the gathered data made simple with the aid of a model. For data mining to have a further influence on corporate operations, it is vital to align its methodology and algorithms with the broad economic goals of the activities it supports. For more than ten years, Data mining research has focused mostly on the problem of periodic pattern mining, spawning a wide range of new fields of study, including sequential pattern mining, structured pattern mining, correlation mining, and associative classification, even frequent pattern-based clustering. This article explores several intriguing research paths and gives a quick summary of the present state of frequent pattern mining.

Frequent pattern mining study is thought to have greatly increased the range of data analysis and, in the long term, will have a significant influence on data mining approaches and applications. In this work, a high utility itemset mining method called HUI-Miner is proposed. Utility-list is a special structure used to contain utility data about an itemset as well as heuristic data for HUI-Miner's search space trimming. It performs better than current algorithms in terms of memory use and running time. The goal of utility mining is to determine which itemsets have the greatest utilities out of a group of items with varying values that make up a utility itemset. Finding highly useful itemsets with negative item values is crucial for mining intriguing patterns, such as association rules, as an itemset may be connected with negative item values in some applications.

2. Literature Survey

A developing field of study in High-utility itemset mining is data mining. According to this study, a histogram of item quantities is maintained at each tree node of the UP-Hist tree. Modern algorithms fall short of UP-Hist tree in terms of the quantity of candidate high utility itemsets produced and overall execution time, according to extensive studies on actual and simulated datasets. The UP-Hist Growth algorithm leverages the UP-Hist tree data structure to find high-utility patterns. The root node of the tree is the starting point for processing each transaction and matching it with nodes inside the tree in order to insert reorganised transactions into the UP-Hist tree. The utility-value of the transaction-prefix also updates the node utility value, hence the support is increased by one. In terms of overall execution period as well as the quantity of candidates produced in the first phase, the suggested method performs better than the most recent techniques. The list of item-quantity values is represented by the histograms in the picture [1].

The Intrusion Detection System (IDS), which can identify many forms of assaults, is essential to network security. SVM (Support Vector Machine) is used in this suggested system to categorise data and assess the system's efficacy. 4,900,000 single connection instances with 42 characteristics, including assaults or normal, are included in the NSL-KDD Cup'99 training dataset. In order to make the nominal features sufficient input for classification using SVM, the data set transformation method is used. This study suggests a technique for intrusion detection using SVM that, when combined with a Gaussian RBF kernel, can speed up the classification model building process and improve intrusion detection precision. The experimental findings demonstrate that the disadvantage of SVM, namely the lengthy time necessary to develop a model, may be addressed provided data sets are adequately handled and the appropriate SVM kernel is selected. The time needed to create the model was 77.07 seconds, and the attack detection accuracy attained was 94.1857% when the data sets were processed correctly and the appropriate SVM kernel was used [2].

The detection of common subgraphs within graph data sets is the focus of the major study field of graph mining, which is included in the category of data mining. Frequent subgraph mining (FSM), which is the foundation of graph mining, attempts to extract all regular subgraphs with occurrence counts surpassing a certain threshold from a particular set of data. During the recent years, a large number of major FSM algorithms have been suggested, showing the area is

becoming more developed. FSM's wide range of applications demonstrate how important it is. The three application fields of chemistry, web, and biology all employ FSM algorithms. The creation of candidates, navigating the search space, and occurrence counting are the main topics of this paper's "cutting-edge" study of FSM. Candidate generation and support computation are the two components that need the greatest computing, with candidate generation being the most expensive. According on candidate generation, search methodology, and approach to frequency counting, FSM algorithms are classed. Graphs can be used with FTM algorithms, but trees and graphs can both be used with FGM algorithms. The final compilation of frequent subgraphs has to be reduced, and the mining work needs to be improved, among other things. Combinatorial complexity and the usefulness of frequent subgraphs are trade-offs [3].

Finding new and practical information from a graph representation of data is known as mining graph data. A crucial component of graph mining is frequent pattern mining (FPM), which aids in finding patterns that theoretically reflect relationships between discrete entities. It is extremely difficult and computationally demanding to develop algorithms that find every frequently occurring subgraph in a big graph dataset since graph and sub graph isomorphism are crucial to the calculations. In order to identify common patterns, this research compares graph mining methods and methodologies. In the case of graph mining, the primary distinction is that the method for figuring out the support is very different. Graph isomorphism, which is the most expensive stage since it is an NP-complete issue, is the key obstacle in the creation of the algorithm for attaining high performance to improve the graph mining process. The necessity for effective graph mining algorithms is growing as a result of the amount and computational complexity of patterns in computer sciences increasing [4].

This study formalises the concept of aggregate movement behaviour by introducing a unique type of spatio-temporal pattern termed a trajectory pattern. It depicts a collection of independent trajectories that all have in common that they pass through the same order of locations and take around the same amount of time. The areas of interest in the given space and the average travel time for moving things between regions are the two key concepts. A trajectory pattern is a series of regularly visited geographical places in the order that the sequence specifies. Two approaches are put forth to extract trajectory patterns from the source trajectory data: pre-conceived areas of interest, which employ arbitrary prior information to identify a collection of locations of interest. T-patterns, which are utilised to choose the zones in origin-destination matrices, are the fundamental building blocks of spatiotemporal data mining. They are dynamically entwined with the extraction of temporal information from sequences, enabling more accurate trajectory patterns. An empirical study produced encouraging findings [5].

3. Proposed System

The approach in this study proposes a unique framework for mining closed high utility itemsets, which serves as a lossless, compact representation of HUIs. A method called DAHU (Derive All High Utility itemsets) is also recommended in order to obtain all high utility itemsets from the set of closed + high utility itemsets while accessing the original database. CHUD and DAHU have proven to be highly successful with a wide variety of high utility itemsets in

experiments on real and synthetic datasets. Moreover, the CHUD plus DAHU strategy outperforms contemporary algorithms in mining high utility itemsets once all of them have been retrieved by DAHU.

Additionally, The CHUD + DAHU technique outperforms cutting-edge algorithms while mining high utility itemsets after DAHU has retrieved all of them. High utility itemset mining is combined with the concept of a closed itemset to create Closed+ High Utility Itemsets (CHUIs), are the system's suggested concise and understandable description of HUIs. This representation is discovered using the three effective algorithms CHUD (Closed+ High Utility itemset Discovery), AprioriHC (Apriori-based technique for mining High utility Closed+ itemset), and AprioriHC-D (AprioriHC method with Discarding unpromising along with isolated items). The AprioriHC and AprioriHC-D algorithms employ breadth first search as well as inherit several desirable characteristics from the widely recognised Apriori algorithm.

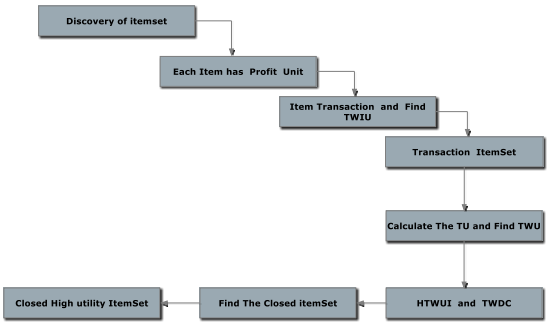


Fig 1: System Architecture

Three unique strategies—REG, RML, and DCM—included in the CHUD algorithm significantly improve its performance. Due to its capacity to take into account nonbinary frequency values of goods in transactions and various profit values for each item, mining high utility itemsets is one of the most significant research challenges in data mining. Such itemsets can be extracted from a transaction database by looking for itemsets whose usefulness is greater than a user-specified threshold. In this research, we present an effective concurrent approach for mining high utility itemsets by dynamically pruning the tree structure, named CHUI (Concurrent High Utility Itemsets). To record the crucial utility data of the potential itemsets, the CHUI-Tree tree structure is introduced.

We build dynamic CHUI-Tree pruning and talk about its logic by keeping track of shifts in support counts of potential high utility items while the tree is being built. By using a concurrent technique, the CHUI-Mine algorithm makes it possible to build a CHUI-Tree as well as find high utility itemsets at the same time.

The issue of massive memory consumption for tree generation along with traversal in tree-based algorithms for mining high utility itemsets is solved by our technique. Numerous experimental findings demonstrate that the CHUI. Mine algforithm is scalable and effective. A unique arrangement for mining closed high utility itemsets, which function as a small and lossless representation of HUIs, was proposed by the system. In order to identify this representation, this study offered three effective algorithms. Authors suggested a technique

dubbed DAHU (Derive All High Utility Itemsets), which recovers all HUIs from the set of CHUIs without using the original database. According to authors, this approach significantly lowers the quantity of HUIs.

The issue of a high number of candidates prevents AprioriHC-D and AprioriHC from functioning properly on dense databases when the minimum utility is low. Each transaction in this module has a Transaction ItemSet, which is classified using the Transaction id and a finite item set because it is known how many items the client has sold. The non-purchased item is designated with a '0'. And a certain number of purchased things happens in each transaction. We are aware of the transactions in this module that have a high transaction volume. Utility determines from all of the transactions what Absolute Utility (AU) is.

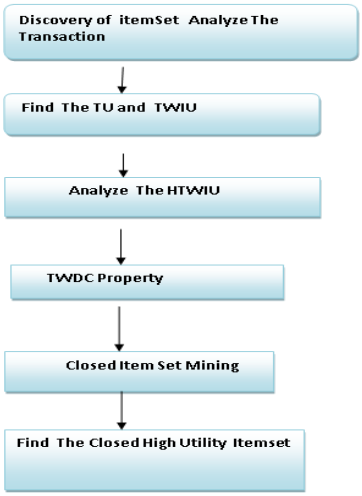


Fig 2: Flow Diagram

AU is determined by multiplying. Likewise, examine the weighted utilisation. We take into account that The High Transaction Weighted Utility ItemSet is generated from TWU, and we take into account that TR is considered HTWUI if the transaction's high utility is larger than its minimum utility. According to the transaction-weighted downward closure property (TWDC), all of an itemset X's supersets that aren't HTWUIs are low utility itemsets. In this research, we offer a unique framework for mining closed+ high utility itemsets that functions as a compact and lossless representation of high utility itemsets.

This framework aims to be highly effective for the mining process and deliver a succinct mining outcomes to users. For mining closed + high utility itemsets, we introduce CHUD (Closed + High Utility Itemset Discovery), an effective approach. A technique named DAHU (Derive All High Utility itemsets) is also suggested without using the original database, retrieve every high utility itemset from the collection of closed + high utility itemsets. Outcomes of trials on actual and artificial datasets demonstrate the great efficiency of CHUD and DAHU and the significant decrease in the quantity of high-useful item pairings (up to 800 times in our tests) that our technique delivers. Additionally, CHUD and DAHU outperform the most advanced algorithms for mining high utility itemsets when all high utility itemsets can be retrieved by DAHU. The following benefits of the suggested strategy are listed:

- The suggested representation is lossless because to a novel structure called a utility unit array that makes it possible to effectively recover all HUIs and their utilities.Numerous orders of magnitude less itemsets result from it.
- According to the results, CHUD can mine all HU Is significantly more quickly than the most recent algorithms.
- UP-Growth, one of the top strategies for mining HUIs right now, is outperformed by the CHUD and DAHU combo, which offers a new technique to acquire all HUIs.

4. Results

The practise of extracting new, non-trivial, and possibly useful information from huge data stores is known as data mining.This research provides a special plan for mining closed high utility itemsets (CHUIs), which is the usefulness or profit of an item in a transaction. To locate this representation, three effective algorithms—AprioriCH, AprioriHC-D, and CHUD—are presented. Authors suggested DAHU (Derive All High Utility Itemsets) to retrieve all HUIs from the set of CHUIs without utilising the original database. Experiments on actual along with artificial datasets have revealed that CHUD and DAHU are quite effective and significantly reduce the amount of high utility itemsets. Further suggestions for improving CHUD's performance included the use of the three effective methods REG, RML, and DCM.

Fig 3: Complete CHUI

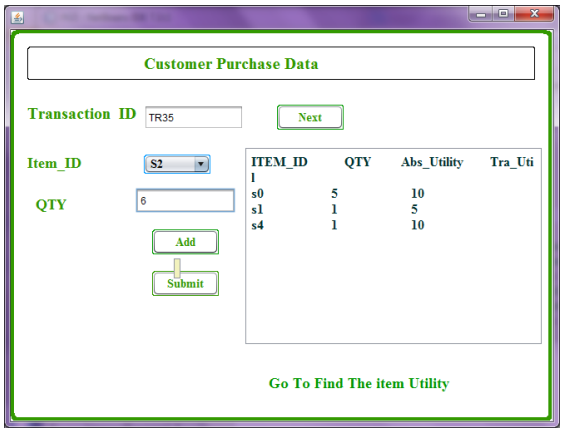
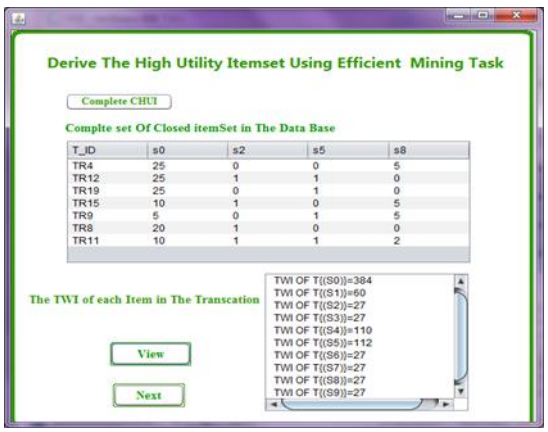


Fig 4: Customer Purchase Data



5. Conclusion

The critical task of mining High Utility Itemsets (HUIs) has several applications. The list of HUIs, however, can be rather vast, which causes HUI mining algorithms to experience slow execution and high memory usage. Concise representations of HUIs have been offered as a solution to this problem. Despite the fact that the idea of generator offers a number of advantages in a variety of applications, no succinct representation of HUIs based on it has yet been developed. In this work, the system proposed a compact representation of all high utility itemsets called closed + high utility itemsets to solve the issue of redundancy in high utility itemset mining. This is the first investigation into the lossless and compact encoding of high utility itemsets, to our knowledge. We suggested an effective method called CHUD to mine this new type of itemset. Further suggestions for improving CHUD's performance included the use of the three effective methods REG, RML, and DCM. Information extraction from new sources that may be beneficial is known as data mining. from vast data bases that is not straightforward. Market basket analysis, a kind of data mining used in retail research, is used to analyse client transactions. The association between the things that occur in transactions more frequently was the focus of earlier data mining techniques. They don't take an item's significance or utility into account while often mining an itemset.

Due to the restrictions of frequently used mining goods, utility mining is now a growing industry. When mining, the profitability or utility of an object is taken into account. In a transaction, an item's utility refers to its significance or financial gain. Finding the item set with utility values over the specified threshold is the main goal of mining high utility goods. We give a literature review on several mining methods in this work.

Reference

1. R. Agrawal and R. Srikant, “Fast algorithms for mining associa-tion rules,” in Proc. 20th Int. Conf. Very Large Data Bases, 1994, pp. 487–499.
2. C. F. Ahmed, S. K. Tanbeer, B.-S. Jeong, and Y.-K. Lee, “Efficient tree structures for high utility pattern mining in incremental data-bases,”IEEE Trans. Knowl. Data Eng., vol. 21, no. 12, pp. 1708– 1721, Dec. 2009.
3. J.-F. Boulicaut, A. Bykowski, and C. Rigotti, “Free-sets: A con-densed representation of Boolean data for the approximation of frequency queries,” Data Mining Knowl. Discovery, vol. 7, no. 1, pp. 5–22, 2003.
4. T. Calders and B. Goethals, “Mining all non-derivable frequent itemsets,” in Proc. Int. Conf. Eur. Conf. Principles Data Mining Knowl. Discovery, 2002, pp. 74–85.
5. K. Chuang, J. Huang, and M. Chen, “Mining top-k frequent pat-terns in the presence of the memory constraint,” VLDB J., vol. 17, pp. 1321–1344, 2008.
6. R. Chan, Q. Yang, and Y. Shen, “Mining high utility itemsets,” in Proc. IEEE Int. Conf. Data Min., 2003, pp. 19–26.
7. A. Erwin, R. P. Gopalan, and N. R. Achuthan, “Efficient mining of high utility itemsets from large datasets,” in Proc. Int. Conf. Pacific-Asia Conf. Knowl. Discovery Data Mining , 2008, pp. 554–561.

8. K. Gouda and M. J. Zaki, “Efficiently mining maximal frequent itemsets,” in Proc. IEEE Int. Conf. Data Mining, 2001, pp. 163–170.
9. T. Hamrouni, “Key roles of closed sets and minimal generators in concise representations of frequent patterns,” Intell. Data Anal. , vol. 16, no. 4, pp. 581–631, 2012.
10. J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without can-didate generation,” in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000, pp. 1–12.
11. T. Hamrouni, S. Yahia, and E. M. Nguifo, “Sweeping the disjunc-tive search space towards mining new exact concise representa-tions of frequent itemsets,” Data Knowl. Eng. , vol. 68, no. 10, pp. 1091–1111, 2009.
12. H.-F. Li, H.-Y. Huang, Y.-C. Chen, Y.-J. Liu, and S.-Y. Lee, “Fast and memory efficient mining of high utility itemsets in data streams,” in Proc. IEEE Int. Conf. Data Mining, 2008, pp. 881–886.
13. C.-W. Lin, T.-P. Hong, and W.-H. Lu, “An effective tree structure for mining high utility itemsets,” Expert Syst. Appl., vol. 38, no. 6, pp. 7419–7424, 2011.
14. G.-C. Lan, T.-P. Hong, and V. S. Tseng, “An efficient projection-based indexing approach for mining high utility itemsets,” Knowl. Inf. Syst , vol. 38, no. 1, pp. 85–107, 2014.
15. H. Li, J. Li, L. Wong, M. Feng, and Y. Tan, “Relative risk and odds ratio: A data mining perspective,” in Proc. ACM SIGACT-SIG-MOD-SIGART Symp. Principles Database Syst., 2005, pp. 368–377.
16. B. Le, H. Nguyen, T. A. Cao, and B. Vo, “A novel algorithm for mining high utility itemsets,” in Proc. 1st Asian Conf. Intell. Inf. Database Syst. , 2009, pp. 13–17.