OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

HIGHLY EFFICIENT ALGORITHM FOR HIGH-UTILITY ITEM SET MINING

Mrunal S. Surve[1], Prof. Rahul L. Paikrao[2]
PG Student, Department of Computer Engineering, Amrutvahini College of Engineering, Pune[1]
Assistant Professor, Department of Computer Engineering, Amrutvahini College of Engineering, Pune[2]
survemrunal@gmail.com[1], rahul.paikrao@avcoe.org[2]

Abstract: High utility item set mining (HUIM) has emerged as an important research topic in data mining, with applications to retail-market data analysis, stock market prediction, and recommender systems, etc. However, there are very few empirical studies that systematically compare the performance of state-of-the-art HUIM algorithms. In this paper, system presents an experimental evaluation on HUIM algorithms. Our experiments show that EFIM are generally the top performers in running time, while EFIM also consumes the least memory in most cases. In order to compare this algorithm in depth, we use another synthetic datasets with varying parameters so as to study the influence of the related parameters, in particular the number of transactions, the number of distinct items and average transaction length, on the running time and memory consumption of EFIM. In this work, we demonstrate that, EFIM is more efficient under low minimum utility values and with large sparse datasets, in terms of running time; although EFIM is the fastest in dense real datasets, it is among the slowest algorithms in sparse datasets. We suggest that, when a dataset is very sparse or the average transaction length is large and running time is favoured over memory consumption. This work has reference value for researchers and practitioners when choosing the most appropriate HUIM algorithm for their specific applications.

Keywords: High-Utility Mining, Item Set Mining, Pattern Mining.

---∴∴∴---

I INTRODUCTION

EFIM (Efficient High Utility Item Set Mining) which introduces several new ideas to more efficiently discovers high -utility item sets both in terms of execution time and memory. EFIM relies on two upper-bounds named sub-tree utility and local utility to more effectively prune the search space. It also introduces a novel array-based utility counting technique called Fast Utility Counting to calculate these upper-bounds in linear time and space. Transaction merging is obviously desirable. However, a key problem is to implement it efficiently. To find identical transactions in O(n) time, sort the original database according to a new total order T on transactions. Sorting is achieved in time, and is performed only once. Projected databases generated by EFIM are often very small due to transaction merging.

EFIM is in general two to three orders of magnitude faster and consumes up to eight times less memory than the rest. Scalability tests were run with regard to dataset size by varying the number of transactions of the real datasets. EFIM has excellent scalability on both sparse and dense datasets.

II LITERATURE SURVEY

[1] Bharti Ahuja, Rupali Bhartiya, "Novel Technique for Mining a High Utility Item Sets from Transactional Database.

Data mining, the pulling out of unseen predictive information from huge databases, is a powerful innovative technology with large potential to facilitate companies focus on the most important information in their data warehouses. Data mining tools guess future trends moreover behaviors, granting businesses to create proactive, knowledge-driven decisions. The term utility means the value or profit of an item in a transaction. The major purpose of high utility items at mining is to discover the item set having utility values above the given threshold. Although a number of Algorithm's have been proposed but there is dilemma like it produce large number of candidate Item sets for High Utility Item sets.

Large number of Item sets degrades the performance of mining in terms of implementation time and space requirement. This situation may not as good as when database contains huge number of transaction. This paper presents a high utility item set mining method. In this technique, the futile patterns are detached at the initial stage of mining. So it will be helping for getting less time consumption for finding high utility of item sets in the transactional database. [1]

[2] Sarode Nutan S., Kothavle Suhas R., "An Efficient Algorithm for Finding High Utility Item Sets From Online Sell.

In the databases data mining and knowledge discovery is a new interdisciplinary field. In data mining there is merging ideas from statistics, machine learning, databases .Data Mining is defined as an activity that extracts some new information contained in large databases. In traditional data mining techniques detecting the statistical correlations between the items which are more frequent in the transaction databases termed as frequent item set mining, that item sets which appear more frequently must be of more importance to the user from the business perspective. The term utility refers to the importance or the usefulness of the item set in transactions quantified in terms like profit, sales or any other user preferences. By considering user preferences such as profit, quantity and cost from online transaction processing systems, utility mining extracts item sets with high utilities. Utility Mining which not only considers the frequency of the item sets but also considers the utility associated with the item sets. In High Utility Item set Mining the objective is to identify item sets that have utility values above a given utility threshold. [2]

[3] V. G. Vijilesh, Dr. S. Hari Ganesh, "An Efficient Algorithm for Mining Closed High Utility Item Set.

Mining of High utility item sets refers to discovering sets of data items that have high utilities. In recent years the high utility item sets mining has extensive attentions due to the wide applications in various domains like biomedicine and commerce. Extraction of high utility item sets from database is very problematic task. The formulated high utility item set degrades the efficiency of the mining process. The existing algorithms suffer the problem of producing a large amount of candidates, which degrades the mining performance in terms of time and space. The Closed high utility item set proposed to achieve this goal. In this paper, a novel algorithm has been proposed to discover the high utility item set without candidate's generation. Experimental results show that the proposed algorithm is faster than the state-of the-art algorithms. [3]

[4] Anu Augustin, Dr. Vince Paul, "High Utility Item Sets Mining – A Brief Explanation with a Proposal.

High utility item sets mining is relevant for business vendors. So that they can give more offers to high utility item sets. To understand the above sentence we need to know what high utility item sets is. High utility item sets are those ones that yield high profit when sold together or alone that meets a user-specified minimum utility threshold from a transactional database. This high utility item set mining is not a new topic, but it is an emerging area. The basis of high utility mining is frequent item set mining. The various problems in frequent item set mining are purchase quantity not taken into account; all items have same importance etc. So the number of items generated will be more. These limitations are overcome by high utility item set mining. For that in HUI mining a utility value (weight) is assigned to each item. Also a threshold applied to remove unwanted item sets. Setting the threshold externally is a tedious work. Too low threshold will generate many HUI's and too high may cause no HUI's to found. In Top-K only top hui's will be found. Here the minimum threshold is set internally. It is zero initially. Performance degrades when there are many hui's in the database. So the concept of closed item set mining is introduced for memory and space efficiency. Also for proper utilization of resources. [4]

III PROPOSED SYSTEM

EFIM (Efficient high-utility Item set Mining) is a one-phase HUIM algorithm that proposes efficient database projection and transaction merging techniques for reducing the cost of database scans. EFIM efficiently merges transactions that are identical in each projected database through a linear time and space implementation. As larger item sets are explored, both projection and merging reduce the size of the database. Furthermore, the authors propose a new array-based utility counting technique, named Fast Utility Counting (FUC), which helps calculate the new upper-bounds (i.e., sub-tree utility and local utility) in linear time and space. In general, EFIM's complexity is roughly linear with the number of item sets in the search space. In their reported results, EFIM was found to be up to 3 times faster than previous state of-the-art algorithms, such as FHM, HUI-Miner, d2HUP, UP-Growth, while consuming up to 8 times less memory.

EFIM Algorithm: We next present our proposal, the EFIM algorithm. It is a one-phase algorithm, which introduces several novel ideas to reduce the time and memory required for HUIM.

IV CONCLUSION

We have presented a novel algorithm for high-utility item set mining named EFIM. It relies on two new upper-bounds named sub-tree utility and local utility. It also introduces a novel array-based utility counting approach named Fast Utility Counting to calculate these upper-bounds in linear time and space. Moreover, to reduce the cost of database scans, EFIM introduces techniques for database

projection and transaction merging, also performed in linear time and space. An extensive experimental study on various datasets shows that EFIM is in general two to three orders of magnitude faster and consumes up to eight times less memory than the state-of-art algorithms UP-Growth+, HUP-Miner, d2HUP, HUI-Miner and FHM. The source code of all algorithms and datasets used in the experiments can be downloaded as part of the SPMF data mining library goo.gl/rIKIub. For future work, we will extend EFIM for popular variations of the HUIM problem such as mining closed+ high-utility item set, generators of high-utility item sets, and on-shelf high-utility item sets.

REFERENCES

[1] Bharti Ahuja, Rupali Bhartiya, "Novel Technique for Mining a High Utility Item Sets from Transactional Database", International Journal of Mechanical and Production Engineering, Volume 5, Issue 2, 2017.

[2] Sarode Nutan S., Kothavle Suhas R., "An Efficient Algorithm for Finding High Utility Item Sets From Online Sell", International Research Journal of Engineering and Technology (IRJET), 2015.

[3] V. G. Vijilesh, Dr. S. Hari Ganesh, "An Efficient Algorithm for Mining Closed High Utility Item Set", International Journal of Computer Science and Mobile Computing, Volume 7, Issue 2, 2018.

[4] Anu Augustin, Dr. Vince Paul, "High Utility Item Sets Mining – A Brief Explanation with a Proposal", International Journal of Science and Research, 2015.

[5] Kirti Santoki, Priyanka Raval, "A Survey on High Utility Item Set Mining with Various Techniques", International Journal of Engineering Development and Research, 2017.

[6] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Ted Gueniche, Prashant Barhate, "Efficient Incremental High Utility Itemset Mining".

[7] Quang-Huy Duong, Philippe Fournier-Viger, Heri Ramampiaro, Kjetil Norvag, Thu-Lan Dam, "Efficient High Utility Itemset Mining using Buffered Utility-Lists".