IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

Survey on Algorithms for High Utility Itemset Generation

Prof Jyoti B. Kulkarni[1], Aishwarya Dingre[2], Sonal Bhosale[3], Deepali Mehroliya[4], Shivani Mudgal[5]

Assistant Professor, Department of Computer Engineering, Sinhgad College of Engineering, Pune, India [1]
Student, Department of Computer Engineering, Sinhgad College of Engineering, Pune, India [2]

Abstract: Data mining is becoming a popular research topic with its frequent applications in online e-business, web click stream analysis and cross marketing. Mining high utility itemsets from a transactional database is concerned with the discovery of itemsets with high utilities like profits or gains. Efficient discovery of the frequent and useful itemsets in huge datasets is a crucial task in data mining. In the recent years, many methods have been proposed for generating high utility patterns. Owing to this, there are few problems such as, if the minimum utility threshold value is set too low, huge amount of itemsets are generated. But, if the minimum utility threshold is set too high, very few or no high utility itemsets will be generated. In high utility itemset mining, the profit values or utility value for every item and the number of units of each item is taken into consideration. We hereby present the study of issues related to the different structures used and algorithms for mining the high utility itemsets.

Keywords: Data mining; frequent itemset; high utility itemset; transactional database.

I. INTRODUCTION

Data mining is becoming a popular and growing area of research in today's era. Data mining helps to produce profitable and useful conclusions from structured and unstructured data. It is concerned with examining of huge volumes of data to find interesting similarities or relations which are proportional to better understanding of the underlying processes. Data mining actions use combination of techniques like artificial intelligence, statistics, and technologies based machine learning. Data mining is regarded as knowledge mining from data.

High utility itemsets mining, is an extension to the problem of frequent pattern mining. In data mining, frequent pattern mining is a popular problem, which consists of finding frequent patterns in the transaction databases. Frequent itemset mining is popular, but it has some important limitations when it comes to analyzing the customer transactions. A significant limitation is that purchase quantities are not taken into consideration. Thus, an item may only appear once or zero time in a transaction. That means whether you buy 5, 10 or 100 units of an item, they all are viewed as the same. Second major drawback of frequent itemset mining is that all items are considered to have equal importance, utility of weight. For example, a car and a packet of bread are considered to be equally important in frequent itemset mining. Thus, frequent pattern mining may find many frequent patterns that are not of much interest. For example, one may find that {bread, milk} is a frequent pattern. However, from a business point of view or perspective, this pattern may not be of much interest because it does not generate profit as expected. Moreover, frequent pattern mining algorithms may miss the rare patterns that generate a high profit. To address these limitations, the problem of frequent itemset mining has been put forward or redefined as the problem of high utility itemset mining. In high utility itemset mining, a transaction database contains transactions where number of purchased items or purchase quantities are taken into account along with the unit profit of each item. The minimum utility threshold value needs to be stated for high utility itemset mining. If the minimum utility threshold is set too low, huge number of itemsets are generated and if the threshold value is set too high, there is a possibility that very few or no itemset will be generated. High utility itemsets mining is growing rapidly as more innovative mining techniques and wider applications are currently being developed. Mining high utility itemsets from the transactional databases is important and has a wide range of applications like business promotion in chain hypermarkets, online e-commerce management, mobile commerce environment planning, website click stream analysis, cross marketing in retail stores etc.

II. RELATED WORK

Vincent S. Tseng, Senior Member, Cheng-Wei Wu, Philippe Fournier-Viger, Philip S. Yu[1]

DOI 10.17148/IJARCCE.2017.6322

proposed a new framework for top-k high utility itemset mining, where ‘k’ is the desired number of high utility itemsets to be mined. Since setting the appropriate minimum utility threshold can be a difficult task for the user, two types of efficient algorithms named TKO (mining Top-K utility itemsets in one phase) and TKU (mining Top-K Utility itemsets) are proposed for mining such itemsets without the need to set minimum utility threshold. A structural comparison of the two algorithms with discussions on their advantages and limitations is stated in the paper. Empirical evaluations on both real and synthetic datasets show that the performance of the proposed algorithms is near to that of the optimal case of state-of-the-art algorithms of utility mining, where k is the desired number of high utility itemsets to be mined.

Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li, and Fangchun Ya[2] proposed a differentially PFP-growth algorithm ,i.e. a private FIM algorithm which is based on the FP-growth algorithm. The PFP-growth algorithm mainly consists of two phases : pre-processing phase and mining phase. In order to improve the utility and the privacy tradeoff in pre-processing phase, a new smart splitting method is put forward to transform the database. The pre-processing phase needs to be performed only once for a given database. To offset the information loss caused by transaction splitting in the mining phase , a runtime estimation method is devised to estimate the actual support of itemsets in the original database. Additionally, by leveraging the downward closure property, a dynamic reduction method is put forward to dynamically reduce the noise which gives privacy during the mining process.

Vincent S. Tseng, Cheng-Wei Wu, Philippe Fournier Viger, and Philip S. Yu[3] have proposed a novel framework for mining closed high utility itemsets (CHUIs), which serves as a compact and lossless representation of HUIs. This paper represents proposed three efficient algorithms named AprioriCH (Apriori logic based algorithm for mining High utility closed itemsets), AprioriHC-D (AprioriHC algorithm enabling the Discarding unpromising and the isolated items) and CHUD (Closed High Utility Itemset Discovery) to search this representation. To recover all high utility itemsets (HUIs) from the set of CHUIs, authors have proposed a method called DAHU (Derive All High Utility Itemsets) and that too, without accessing the original database. Authors claimed that this technique achieves great reduction in the number of HUIs. AprioriHC-D and AprioriHC both algorithms cannot perform well on the dense databases when min_utility is low since they suffer from the problem of a large amount of candidates.

Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu[4] proposed two algorithms for mining high utility itemsets with a set of effective strategies for pruning candidate itemsets. They are utility pattern growth (UP-Growth) and UP-Growth+. The information of high utility itemsets is maintained in a tree-based data structure named as utility pattern tree (UP-Tree) such that, with only two scans of database,efficient candidate itemsets can be generated. The performance of UP-Growth and UP-Growth+ is compared with the state-of-the-art algorithms on both real and synthetic data sets. Empirical results show that the proposed algorithms, particularly UPGrowth+, reduces the number of candidates effectively and also outperforms other algorithms quite substantially on basis of runtime, especially when the databases contain lots of long transactions.

Hua-Fu Li, Hsin-Yun Huang, Suh-Yin Lee proposed two efficient one pass algorithms namely MHUI-BIT and MHUITID for mining high utility itemsets from data streams within a transaction sensitive sliding window. Two effective representations of a lexicographical tree-based summary data structure and itemset information were developed for improving the efficiency of mining high utility itemsets.

DOI 10.17148/IJARCCE.2017.6322

We present in the following table some prominent works in short along with our observations and inferences.

No.	Title Paper	Author Name	Claims by Author	Observations and Inferences
1	Efficient Algorithms for Mining Top-K High Utility Itemsets (2016)	Vincent S. Tseng, Cheng-Wei Wu, Viger, Philip S. Yu, 2016	Two types of efficient algorithms named TKU and TKO are proposed.	Empirical evaluation on both real and synthetic datasets show that the performance of the proposed algorithms is close to that of the optimal case of state-of-the-art utility mining algorithms, where 'k' is the desired number of high utility itemsets to be mined.
2	Differentially Private Frequent Itemsets Mining via Transaction Splitting (2015)	Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li, and Fangchun Ya,2015	PFP growth algorithm consists of preprocessing phase and the mining phase	A novel smart splitting method is proposed to transform the database. For a given database, the preprocessing phase needs to be performed only once.
3	Efficient Algorithms for Mining the Concise and Lossless Representation of Closed+ High Utility Itemsets (2015)	Vincent S. Tseng, Cheng-Wei Wu, Philippe Fournier Viger, and Philip S. Yu, 2015	High utility Itemsets can be compacted after pruning the database.	AprioriHC-D and AprioriHC both algorithms cannot perform well on dense databases when min_util is low since they suffer from the problem of large amount of candidates
4	Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases (2013)	Vincent S. Tseng, BaiEn Shie, Cheng-Wei Wu, and Philip S. Yu, 2013	Tree based data structures (UP tree, UP-Growth and UP-Growth+) can be used to store the candidate itemsets.	Improvement in the runtime, especially when the database contains lots of long transaction.
5	Fast and memory efficient mining of high-utility itemsets from data streams: with and without negative item profits	Hua-Fu Li, Hsin-Yun Huang, Suh-Yin Lee, 2014	Adapted approaches of algorithms MHUI-BIT and MHUI-TID are developed to discover high utility itemsets with negative item profits from data streams.	To improve the efficiency of mining high utility itemsets two effective representations of an extended lexicographical tree-based summary data structure and itemsets information were developed.

DOI 10.17148/IJARCCE.2017.6322

III. CONCLUSION

High Utility Itemset mining is becoming an increasingly popular topic in the field of Data Mining. This paper represents a survey on many different High Utility Itemset mining algorithms that were proposed and implemented by researchers earlier, for the better development in the field of Data Mining. The various algorithms discussed above will be of great use for developing a new improved technique for mining high utility itemsets which is efficient and effective. In future, we are going to develop a project by modifying the TKU and TKO algorithms which will mine the high utility itemsets and will produce more accurate results than the state-of-the-art algorithms.

ACKNOWLEDGMENT

We would like to express our sincere gratitude towards our guide Prof. J. B. Kulkarni for her invaluable guidance and supervision that helped us in our research. She has always encouraged us to explore new concepts and pursue newer research problems. I credit our project contribution to her. Collectively, we would also like to thank our project committee members Prof. J. R. Prasad and Prof. D. P. Salapurkar for their time, suggestions, and for graciously agreeing to be on our committee, and always making themselves available. We cannot thank them enough.

REFERENCES

[1] Vincent S. Tseng, Cheng-Wei Wu, Viger, Philip S. Yu,, " Efficient Algorithms for Mining Top-K High Utility Itemsets", IEEE Transactions on Knowledge and Data Engineering, DOI 10.1109/TKDE.2015.

[2] Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li, and Fangchun Ya, "Differentially Private Frequent Itemset Mining via Transaction Splitting", IEEE Transactions on Knowledge and Data Engineering, Vol.27, No 7, July 2015.

[3] Vincent S. Tseng, Cheng-Wei Wu, Philippe Fournier-Viger, and Philip S. Yu, "Efficient Algorithms for Mining the Concise and Lossless Representation of High Utility Itemsets", IEEE Transactins on Knowledge and Data Engineering, Vol. 27, No. 3, 2015.

[4] Vincent S. Tseng, Bai-En Shie, Cheng-Wei Wu, and Philip S. Yu, "Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases", IEEE Transactions on Knowledge and Data Engineering, Vol.25, No. 8, AUGUST 2013, pp 1772-1786.

[5] Hua-Fu Li, Hsin-Yun Huang, Suh-Yin Lee, "Fast and memory efficient mining of high-utility itemsets from data streams: with and without negative item profits", Springer, 2010. DOI 10.1007.

DOI 10.17148/IJARCCE.2017.6322