ISSN:2348-2079

Volume-6 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Enhanced Efficient High Average Utility Pattern Mining For Shopping Package

D.Lavanya*[1],A.Pavithra[1],M.Abinaya[1],M.Nandhini[1],Ms.J.Sakunthala[2]

[1]FINAL YEAR, B.Tech-IT, Nandha Engineering College-Erode 52.
[2]Associate Professor, Department Of IT, Nandha Engineering College-Erode 52.
E-MAIL: *lavanyaduraisami@gmail.com

ABSTRACT: **High-utility itemset mining (HUIM) is a critical issue in recent years since it can be used to reveal the profitable products by considering both the quantity and profit factors instead of frequent itemset mining (FIM) or association-rule mining (ARM). Several algorithms have been presented to mine high-utility itemsets (HUIs) and most of the designed algorithms have to handle the exponential search space for discovering HUIs when the number of distinct items and the size of database are very large. The proposed algorithm first adopts the TWU model to find the number of high-transaction-weighted utilization 1-itemsets (1-HTWUIs) as the particle size, which can greatly reduce the combinational problem in the evolution process Frequent weighted itemsets represent correlations frequently holding in data in which items may weight differently.**

Keywords: **High average-utility pattern, tighter upper bounds, utility mining, pruning strategy, data mining.**

1. INTRODUCTION

Data mining is about finding new information in a lot of data. Data mining, *the extraction of hidden predictive information from large databases*, it is a powerful new technology

This paper tackles the issue of discovering rare and weighted item sets, i.e., the infrequent weighted item set (IWI) mining problem. UP-Tree is used, to maintain the information of transactions and high utility item sets. Two strategies are applied to minimize the overestimated utilities stored in the nodes of global UP-Tree. In following sections, the elements of UP-Tree are first defined By applying strategy DGN (Discarding Global Node), the utilities of the nodes that are closer to the root of a global UP-Tree are further reduced. DGN is especially suitable for the databases containing lots of long transactions. In other words, the more items a transaction contains, the more utilities can be discarded by DGN. On the contrary, traditional TWU (Total Weight Utility) mining model is not suitable for such databases since the more items a transaction contains, the higher TWUwith great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

It helps organization to make full use of the data stored in their databases and when it comes to decision making, this is true in all fields, and is also true in all different type

Data cleaning is a process of cleaning the data by filling the missing values, smoothing noisy data, identifying or removing outliers and resolving Data Mining often requires data integration, the merging of data from multiples data stores into coherent data store, as in data warehousing. These sources may include multiple data bases, data cubes or flat files.

In data transformation, data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance.

Data Reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume. But maintains the integrity of the original data, mining on the reduced data set should be more efficient and produces the same analytical result. In data mining, association rules are useful for analyzing and predicting customer behaviour. It plays an important role in shopping basket data analysis, product clustering, and catalogue design and store layout. An example of an association rule would be if a customer buys a dozen eggs, and also 80% likely to purchase milk.

Classification is a data mining technique used to predict group membership for data instances. For example, classification used to predict whether the weather on a particular day will be sunny, rainy or cloudy. Popular classification techniques include decision trees and neural networks**.**

It is used to predict unknown or missing values. Predicting the identity of one thing based purely on the description of another, related thing not necessarily future events, just unknowns based on the relationship between a things that are know and a thing need to predict.

Sequential Pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity.

Security and privacy are not very new concepts in data mining, but there is too much that can be done in this area with data mining. Analysis of social networks and group dynamics from electronic communication give a thorough analysis of impact of social networks and group dynamics. Specifying the need to understand cognitive networks, it also models knowledge network using the Enron E-mail corpus. Recording of electronic communication like email logs, and web logs have captured human process. Analysis of this can present an opportunity to understand sociological and psychological process. Privacy preserving data analysis on graph and social networks provides various types of privacy breach and present an analysis using k-candidate anonymity, K-degree anonymity and k-neighbourhood anonymity.

There are many motivating factors for the study of this area. Biggest is profit. Everyone wants profit. It presents models of assets prices, and presents the model of relative changes of stock prices. Market-Based Profile Infrastructure: Giving back to the user present a global solution for distributed recommendations in an adaptive decentralized network.

This is another promising area. It comprises of many areas such as remote sensing, earth-science, biosphere, oceans and predicts the ecosystem. There are also issues in mining the earth science like high dimensionality because long term series data are common in data mining. Study of this area is important due to radical changes in ecosystem has led to floods, drought, ice-storms, hurricanes, tsunami and other disasters. Land Cover Change detections also one of the areas, in a press release by NASA. It shows the history of natural disasters.Conventional data mining is thought to be as containing a large repository, and then mine knowledge. But there is an eminent need for mining knowledge from distributed resources. Typical algorithms which are available to user are based on assumption that the data is memory resident, which makes them unable to cope with the increasing complexity of distributed algorithms. Similar issues also rise while mining data in sensor network, and grid data mining. It needs distribution classification algorithms.

The limitations of frequent or rare item set mining motivated the utility based mining approach, which allows a user to conveniently express the usefulness of item sets as utility values and then find item sets with high utility values higher than a threshold. In utility based mining the term utility refers to the quantitative representation of user preference i.e. the utility value of an item set is the measurement of the importance of that item set in the user's perspective. For e.g. if a sales analyst involved in some retail research needs to find out which item sets in the stores earn the maximum sales revenue. The utility value of an item set can be profit, popularity, and page rank, measure of some aesthetic aspect such as beauty or design or some other measures of user's preference.

Two types of utility measures for any item set, transaction utility and external utility. The transaction utility of an item in a transaction is defined according to the information stored in the transaction. For e.g. the quantity of an item sold in the super market transaction database. The external utility of an item set is based on the information provided by the user and is not available in the transactions. For e.g. in case of sales database the external utility may be the profit associated with the sale of item sets.

Frequent item set mining is based on the rationale that the item sets which appear more frequently in the transaction databases are of more importance to the user .However the practical usefulness of mining the frequent item set by considering only the frequency of appearance of the item sets is challenged in many application domains such as Retail research. It has been that in many real applications that the item sets that contribute the most in terms of some user defined utility function (for e.g. profit) are not necessarily frequent item sets. Utility mining attempts to bridge this gap by using item utilities as an indicative measurement of the importance of that item in the user's perspective. Utility mining is a comparatively new area of research and most of the literature work is focused towards reducing the search space while searching for the high utility item sets.

Algorithm 1 : EHAUPM Algorithm

Input: *D*, a transactional database; *ptable*, a pro_t table ;a minimum high average-utility threshold.

Output: The set of high average-utility itemsets, HAUIs.

```
// X:AUL, the average-utility list of
an itemset(X);

// I _, the set of items in D;

// I _:AULs, the set of average-utility
lists of items in D;

1 calculate the auub of each item in D;
2 calculate the total utility TU in D;
3 for each item ij such that auub(ij) < TU _ _ do
4 remove ij from D;
5 recalculate the auub of each remaining item in D;
6 sort items in transactions in auub-ascending order;
7 construct I _:AULs and EAUCM;
8 Search(;, I _:AULs, EAUCM, _, TU);
9 return HAUIs;
```

The proposed algorithm _rst calculates the auub of each item (Line 1) and the total utility (TU) by scanning the database once (Line 2). For each item in the database, its auub value is then checked against the minimum high average utility count (Line 3), and items not satisfying the threshold are removed (Line 4). After that, the auub value is then calculated again for each item to obtain its lower upper bound value (Line 5). The remaining items in the transactions are sorted by their auub-ascending order (line 6), and the MAU-list of each remaining item is then built (Line 7), as well as the EAUCM of 2-itemsets (Line 7).After that, the search algorithm is then applied to recursively mine HAUIs using I _:AULs by performing a depth-_rst Search (Line 8), and the resulting HAUIs are ret urned(Line 9). The pseudocode of the Search algorithm is given in Algorithm 2. As shown in Algorithm 2, the Search algorithm takes an itemset and a set of its 1-extensions as arguments and then sequentially processes each itemset (Xa) in the set of1-

extensions of itemset (X) (Lines 1 to 12). For each itemset (Xa), the utility of each entry in the MAU-list structure of Xa is then summed (Line 2), and if its value is no less than that.

VII. CONCLUSION

In this paper, we present two efficient upper-bounds to further reduce the upper-bound values compared to the traditional *auub* model. The MAU-list structure is also developed to keep the necessary information for the mining process, and avoid performing multiple database scans. Three pruning strategies are respectively developed to prune unpromising itemsets early. The existing system faces the issue of discovering infrequent itemsets by using weights for differentiating between relevant items only.Discovering infrequent itemsets is not carried out within each transaction.The usefulness of the discovered patterns has not been validated on data coming from a real-life context.Thus, the computational time and search space can be greatly reduced. Experiments on both real-life and synthetic datasets showed that the proposed algorithm significantly outperforms the state-of-the-art HAUI-Miner algorithm and the developed pruning strategies are efficient to speed up the mining performance, and can also be used with the traditional *auub* model.

REFERENCES

[1] R. Agrawal and R. Srikant, ``Fast algorithms for mining association rules,''
in *Proc. Int. Conf. Very Large Data Bases*, 1994, pp. 487_499.

[2] R. Agrawal and R. Srikant, ``Mining sequential patterns,'' in *Proc. Int.*
Conf. Data Eng., 1995, pp. 3_14.
J. C.-W. Lin *et al.*: EHAUPM With Tighter Upper Bounds

[3] C. F. Ahmed, S. K. Tanbeer, B. S. Jeong, and Y. K. Lee, ``Ef_cient
tree structures for high utility pattern mining in incremental databases,''
IEEE Trans. Knowl. Data Eng., vol. 21, no. 12, pp. 1708_1721,
Dec. 2009.

[4] M.-S. Chen, J. S. Park, and P. S. Yu, ``Ef_cient data mining for path
traversal patterns,'' *IEEE Trans. Knowl. Data Eng.*, vol. 10, no. 2,
pp. 209_221, Mar. 1998.

[5] C. Creighton and S. Hanash, ``Mining gene expression databases for
association rules,'' *Bioinformatics*, vol. 19, no. 1, pp. 79_86, 2003.

[6] Y.-C. Li and C.-C. Chang, ``A new FP-tree algorithm for mining frequent
itemsets,'' in *Proc. Adv. Workshop Content Comput.*, 2004, pp. 266_277.

[7] A. Erwin, R. P. Gopalan, and N. R. Achuthan, ``Ef_cient mining of high
utility itemsets from large datasets,'' in *Proc. Paci_c Asia Conf. Knowl.*
Discovery Data Mining, 2008, pp. 554_561.

[8] P. Fournier-Viger, C.-W.Wu, S. Zida, and V. S. Tseng, ``FHM: Faster highutility
itemset mining using estimated utility co-occurrence pruning,'' in
Proc. Int. Symp. Found. Intell. Syst., 2014, pp. 83_92.

[9] P. Fournier-Viger *et al.*, ``The SPMF open-source data mining library
version 2 and beyond,'' in *Proc. Eur. Conf. Mach. Learn. Principles Pract.*
Knowl. Discovery, 2016, pp. 36_40.