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 10, October 2017

Effective and Efficient High Utility Frequent Item Set Mining Using Modified LP-Tree Algorithm

P. S. Saranya[1] M. Narmatha M.Sc., M.Phil.,[2]

Scholar, Sri Jayendra Saraswathy Maha Vidyalaya College of Art and Science, Coimbatore[1]

Assistant Professor, Department of Computer Science, Sri Jayendra Saraswathy Maha Vidyalaya College of Art and Science, Coimbatore[2]

Abstract: Mining high utility item set from large database refers to the discovery of item sets with high utility like profits. Although a number of relevant approaches have been proposed in recent years, they incur the problem of producing large number of candidate item set for high utility item sets. Such a large number of candidate item set degrades the mining performance in terms of execution time and space requirement. The situation may become worse when the database contains large number of long transactions or long high utility item sets (HUIs). Utility mining is the best solution for the above problems explained. In utility mining, each item is associated with a utility (e.g. unit profit) and an occurrence count in each transaction (e.g. quantity). The utility of an item set represents its importance, which can be measured in terms of weight, value, quantity or other information depending on the user specification. The algorithm used here is Modified LP-Tree algorithm (Modified Linear Prefix Tree) for mining high utility item sets with a set of techniques for pruning candidate item sets. The information of high utility item sets is maintained in a special data structure named Modified LP-Tree such that the candidate item sets can be generated efficiently with only two scans of the database. This method not only reduces the number of candidates effectively but also out performs other algorithms substantially in terms of execution time, especially when the database contains lots of long transactions.

Keyterms: Data mining, high utility item sets, Modified LP-Tree algorithm.

I INTRODUCTION

DATA mining is the process of revealing nontrivial, previously unknown and potentially useful information from large databases. Discovering useful patterns hidden in a database plays an essential role in several data mining tasks, such as frequent pattern mining, weighted frequent pattern mining, and high utility pattern mining. Among them, frequent pattern mining is a fundamental research topic that has been applied to different kinds of databases, such as transactional databases, streaming databases and time series databases and various application domains, such as bioinformatics, Web click-stream and mobile environments. Nevertheless, relative importance of each item is not considered in frequent pattern mining. To address this problem, weighted association rule mining was proposed. In this framework, weights of items, such as unit profits of items in transaction databases, are considered. With this concept, even if some items appear infrequently, they might still be found if they have high weights.

However, in this framework, the quantities of items are not considered yet. Therefore, it cannot satisfy the requirements of users who are interested in discovering the item sets with high sales profits, since the profits are composed of unit profits, i.e., weights, and purchased quantities In view of this; utility mining emerges as an important topic in data mining field. Mining high utility item sets from databases refers to finding the item sets with high profits. Here, the meaning of item set utility is interestingness, importance, or profitability of an item to users. Utility of items in a transaction database consists of two aspects: 1) The importance of distinct items, which is called external utility, 2) The importance of items in transactions, which is called internal utility. Utility of an item set is defined as the product of its external utility and its internal utility. An item set is called a high utility its utility is no less than a user-specified minimum utility threshold; otherwise, it is called a low-utility item set. Mining high utility item sets from databases is an important task has a wide range of applications such as website click stream analysis, business promotion in chain hypermarkets, cross marketing in retail stores, online e-commerce management, mobile commerce environment planning, and even finding important patterns in biomedical applications. It is widely recognized that FP-Growth achieves a better performance than Apriority-based algorithms since it finds frequent item sets without generating any candidate item set and scans database just twice. In the framework of frequent item set mining, the importance of items to users is not considered. Thus, the topic called weighted association rule mining was brought to attention and

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 10, October 2017

proposed the concept of weighted items and weighted association rules. However, since the framework of weighted association rules does not have downward closure property, mining performance cannot be improved.

II LITERATURE REVIEW

Fast Algorithms for Mining Association Rules. [1] We consider the problem of discovering association rules between items in a large database of sales transactions. We present two new algorithms for solving this problem that are fundamentally different from the known algorithms. Efficient Tree Structures for High Utility Pattern Mining in Incremental Databases. [2] Recently, high utility pattern (HUP) mining is one of the most important research issues in data mining due to its ability to consider the non-binary frequency values of items in transactions and different profit values for every item.

Mining Top-k Frequent Patterns in the Presence of the Memory Constraint. [3] We explore in this paper a practicably interesting mining task to retrieve top-k (closed) item sets in the presence of the memory constraint. Specifically, as opposed to most previous works that concentrate on improving the mining efficiency or on reducing the memory size by best effort, we first attempt to specify the available upper memory size that can be utilized by mining frequent item sets.

Centralized Class Specific Dictionary Learning for wearable sensors based physical activity recognition proposed by Sherin M Mathews, et al. In [13] this process address a novel technique for a sensor platform that performs physical activity recognition by leveraging a class specific regularizer term into the dictionary pair learning objective function. This algorithm jointly learns a synthesis dictionary and an analysis dictionary in order to simultaneously perform signal representation and classification once the time-domain features have been extracted. Specifically, the class specific regularizer term ensures that the sparse codes belonging to the same class will be concentrated thereby proving beneficial for the classification stage.

Mining Top-K Sequential Rules. [4] Mining sequential rules requires specifying parameters that are often difficult to set (the minimal confidence and minimal support). Depending on the choice of these parameters, current algorithms can become very slow and generate an extremely large amount of results or generate too few results, omitting valuable information.

Mining Top-K Association Rules. [5] Mining association rules is a fundamental data mining task. However, depending on the choice of the parameters (the minimum confidence and minimum support), current algorithms can become very slow and generate an extremely large amount of results or generate too few results, omitting valuable information. Dictionary and deep learning algorithms with applications to remote health monitoring system proposed by Sherin Mary Mathews. In [14] In this dissertation, present three dictionary learning approaches and a deep learning framework for classification tasks related to remote health monitoring systems. This dissertation presents a more robust class specific centralized dictionary learning method to solve the wearable sensor-based physical activity classification problem.

Novel Concise Representations of High Utility Item sets Using Generator Patterns. [6] Mining High Utility Item sets (HUIs) is an important task with many applications. However, the set of HUIs can be very large, which makes HUI mining algorithms suffer from long execution times and huge memory consumption.

Mining T op-K Frequent Closed Patterns without Minimum Support. [7] We propose a new mining task: mining top-k frequent closed patterns of length no less than min `, where k is the desired number of frequent closed patterns to be mined, and min ` is the minimal length of each pattern.

Leveraging discriminative dictionary learning algorithms for single lead ECG classification proposed by Sherin Mary Mathews. In [11] Detecting and classifying cardiovascular diseases and their underlying ethology are necessary in critical-care patient monitoring. In this work, explore the effectiveness of discriminative dictionary learning algorithms for electrocardiogram (ECG) classification task and exhibit that they can achieve very competitive performance compared to traditional methods with lower computational cost.

Pruning strategies for mining high utility item sets. [8] High utility item set mining problem involves the use of internal and external utilities of items (such as profits, margins) to discover interesting patterns from a given transactional database.
Efficient updating of discovered high-utility item sets for transaction deletion in dynamic databases. [9] Most algorithms related to association rule mining are designed to discover frequent item sets from a binary database.

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 10, October 2017

Applying the maximum utility measure in high utility sequential pattern mining. [10] Recently, high utility sequential pattern mining has been an emerging popular issue due to the consideration of quantities, profits and time orders of items. Sequences in the existing.

Maximum Correntropy Based Dictionary Learning Framework for Physical Activity Recognition Using Wearable Sensors proposed by Sherin M Mathews, et al. In [12] Physical activity recognition is difficult due to the inherent complexity involved with different walking styles and human body movements.

III PROPOSED METHODOLOGY

Many of state-of-the-art mining algorithms use tree structures, and they create nodes independently and connect them as pointers when constructing their own trees. Accordingly, the methods have pointers for each node in the trees, which is an inefficient way since they should manage and maintain numerous pointers. In this project, a novel tree structure is proposed to solve the limitation. Our new structure, M-LP-tree (Modified Linear Prefix – Tree) is composed of array forms and minimizes pointers between nodes. In addition, M-LP-tree uses minimum information required in mining process and linearly accesses corresponding nodes. It also suggests an algorithm applying M-LP-tree to the mining process. M-LP-tree which can conduct mining operations more quickly and efficiently than previous algorithms.

Our LP-tree can solve the existing system limitation due to its special structure based on the linear form. It can obtain advantages by converting tree's nodes as array forms. It can increase memory efficiency through arrayed nodes since they can reduce connection information. It can also speed up item traversal times since M-LP-tree does not use pointers in most cases and generates a large number of nodes at once due to its linear structure. By applying the features of LP-tree to mining process, we can obtain the following benefits: Tree generation rate of our approach becomes faster than that of FP-growth since ours can create multiple nodes at once by a series of array operations. Meanwhile, FP-growth makes nodes one by one.

We can access parent or child nodes without corresponding pointers when searching trees since the nodes are stored as an array form, Memory usage for each node becomes relatively small since LP-tree does not require internal node pointers, It is possible to traverse trees more quickly compared to searching for them with pointers since our approach directly accesses corresponding memories due to the feature of the array structure. The main goal of the proposed algorithm was to reduce not only memory usage needed for building trees but also time to traverse them by applying a linear structure instead of the previous form used in FP growth, M-LP-tree contributed to improving performance of frequent pattern mining since it spent less memory generating nodes compared to FP-tree and accessed them without any pointers in many cases, M-LP-tree will present outstanding performance in terms of runtime, memory usage, and scalability.

3.1 METHODOLOGIES

3.1.1 Extracting data items

In the first module, the user can choose the dataset to get high utility item sets, after choosing the dataset, the data items should be extracted from the transactions dataset, the data items are individual items which occur in the transaction.

3.1.2 Calculating TU (Transaction Utility):

After constructing parse tree for the XML database. The Transaction Utility (TU) of each item is calculated. The transaction Utility is calculated by using the value of profit and its quantity. Transaction Utility=Profit*quantity

3.1.3 Finding TWU (Transaction Weighted Utility):

After calculating Transaction Utility (TU) of each item from the parse tree. We need to find Transaction Weighted Utility(TWU) for each items from the calculated Transaction Utility(TU).The TWU of item set whose value is less than the given threshold value are discarded by pruning items .The discarded item sets are called unpromising item set they don't yield more utility to the user. The item set whose TWU value is less than minimum utility is called unpromising item set, otherwise promising item set.

3.1.4 Mining frequent patterns based on M-LP-tree

Mining frequent patterns based on M-LP-tree (LP-growth) LP-growth searches LP-tree and create a conditional LP-tree for mining frequent patterns. To do that, our algorithm first selects the bottom item from the header list and traverses nodes connected to corresponding node links. Then, supports of the visited nodes are stored, and nodes from each linked node to a root are searched.

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 10, October 2017

Each node can be accessed directly if the search is conducted within one LPN. In other words, given a current node, the algorithm immediately accesses N (k1)to approach a parent node of . Iterating the traversal regarding one LPN, the algorithm reaches a header of the LPN, where the header refers to its parent node, i.e. the other LPN.

IV EXPERIMENTAL RESULT AND ANALYSIS

In this section, we present experimental results by comparing our algorithm, with the state-of-art algorithms. In order to show that the experiments are reasonable, we evaluate their performances based on three important criteria: runtime, memory usage, and scalability.

i. Memory Usage (kb):

In this section, we evaluate memory usage for each algorithm with the same datasets as the runtime tests. Our algorithm, it guarantees memory consumption as good as that of the state-of-the-art algorithm. Moreover, our algorithm presents the most outstanding results in many cases

No of Web Documents	FP Growth	UP - Tree	Modified LP - Tree
100	156	110	96
200	184	159	140
300	245	211	187
400	317	294	221
500	389	320	244

ii. Processing Time (Ms):

We can observe that our proposed outperforms the others in almost all of the cases. Modified LP-Tree uses the proposed linear structure to its trees instead of the previous tree form in order to minimize access times to search nodes. As a result, its advantages have a positive effect on reducing runtime in whole experiments. Especially as the minimum support threshold becomes lower, the difference of runtime between our algorithm and the others is bigger.

No of Web Documents	FP Growth	UP - Tree	Modified LP - Tree
100	963	856	676
200	1050	948	812
300	1120	1075	948
400	1148	1098	1056
500	1250	1192	1173

iii. Scalability (%):

Proposed algorithm shows the best memory scalability while the others have relatively poor performance, which indicates that our Modified LP-tree can store these increasing attributes more efficiently than the other structures of the competitor algorithms. Through the above experimental results, we know that the proposed algorithm outperforms the others with respect to increasing transactions and items in terms of scalability as well as runtime and memory usage for the real datasets.

No of Web Documents	FP Growth	UP - Tree	Modified LP - Tree
100	93.6	95.5	96.8
200	93.1	94.8	96.4
300	92.6	94.3	95.9
400	92.4	93.8	95.6
500	92.0	93.5	95.1

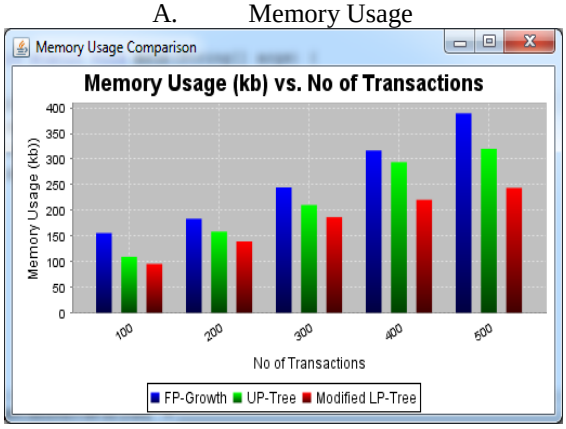
IJARCCE

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

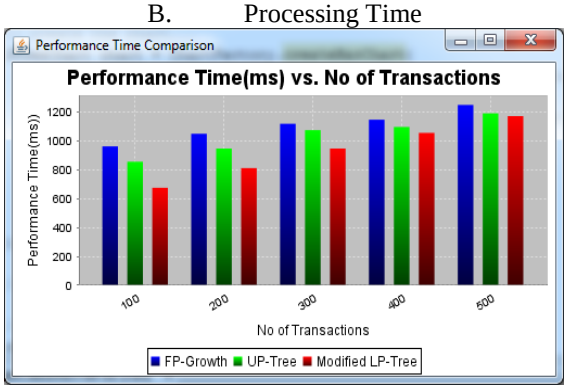
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

GRAPH RESULTS

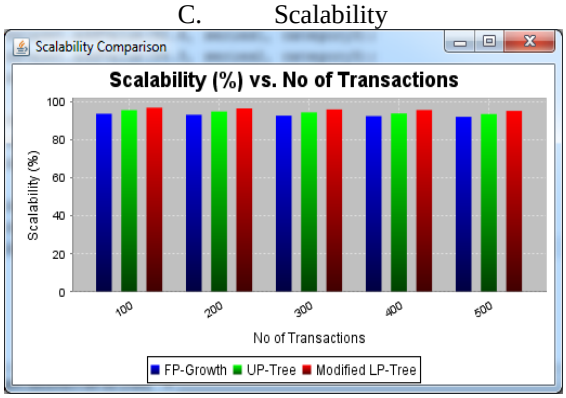
A. Memory Usage



B. Processing Time



C. Scalability



Our experimental results showed that Modified LP-Tree presented outstanding performance in terms of runtime, memory usage, and scalability. The techniques and strategies described in this paper can be applied to not only general frequent pattern mining but also a variety of pattern mining fields such as closed/maximal pattern mining, top-k pattern mining, and graph mining

V CONCLUSION

In this paper, we proposed a new tree structure, Modified LP-tree, and an algorithm, applying it to the mining process. The main goal of the proposed algorithm was to reduce not only memory usage needed for building trees but also time to traverse them by applying a linear structure instead of the previous form used in FP growth. Modified LP-tree contributed to improving performance of frequent pattern mining since it spent less memory generating nodes compared to Modified LP-tree and accessed them without any pointers in many cases. Our experimental results showed that presented outstanding performance in terms of runtime, memory usage, and scalability. We could also observe that our algorithm outperformed the previous algorithms especially in the runtime experiments due to the reduced pointer accesses. The techniques and strategies described in this paper can be applied to not only general frequent pattern

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 10, October 2017

mining but also a variety of pattern mining fields such as closed/maximal pattern mining, top-k pattern mining, and graph mining. We expect that these future researches lead to improvement of mining performance in various areas.

REFERENCES

[1] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," inProc. Int. Conf. Very Large Data Bases, 1994, pp. 487–499.
[2] C. Ahmed, S. Tanbeer, B. Jeong, and Y. Lee, "Efficient tree structures for high-utility pattern mining in incremental databases," IEEE Trans. Knowl. Data Eng., vol. 21, no. 12, pp. 1708–1721, Dec.2009.
[3] K. Chuang, J. Huang, and M. Chen, "Mining top-kfrequent patterns in the presence of the memory constraint,"VLDB J., vol. 17,pp. 1321–1344, 2008.
[4] R. Chan, Q. Yang, and Y. Shen, "Mining high-utility itemsets," in Proc. IEEE Int. Conf. Data Mining, 2003, pp. 19–26.
[5] P. Fournier-Viger and V. S. Tseng, "Mining top-k sequential rules," inProc. Int. Conf. Adv. Data Mining Appl., 2011, pp. 180–194.
[6] P. Fournier-Viger, C. Wu, and V. S. Tseng, "Mining top-kassociation rules," inProc. Int. Conf. Can. Conf. Adv. Artif. Intell., 2012, pp. 61–73.
[7] P. Fournier-Viger, C. Wu, and V. S. Tseng, "Novel concise representations of high utility itemsets using generator patterns," in Proc. Int. Conf. Adv. Data Mining Appl. Lecture Notes Comput. Sci., 2014, vol. 8933, pp. 30–43.
[8] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," inProc. ACM SIGMOD Int. Conf. Manag. Data, 2000, pp. 1–12.
[9] J. Han, J. Wang, Y. Lu, and P. Tzvetkov, "Mining top-kfrequent closed patterns without minimum support," in Proc. IEEE Int. Conf. Data Mining, 2002, pp. 211–218.
[10] S. Krishnamoorthy, "Pruning strategies for mining high utility itemsets,"Expert Syst. Appl., vol. 42, no. 5, pp. 2371–2381, 2015.
[11] Mathews, Sherin M., Luisa F. Polanía, and Kenneth E. Barner. "Leveraging a discriminative dictionary learning algorithm for single-lead ECG classification." Biomedical Engineering Conference (NEBEC), 2015 41st Annual Northeast. IEEE, 2015.
[12] Mathews, Sherin M., Chandra Kambhamettu, and Kenneth E. Barner. "Maximum correntropy based dictionary learning framework for physical activity recognition using wearable sensors." International Symposium on Visual Computing. Springer International Publishing, 2016.
[13] Mathews, Sherin M., Chandra Kambhamettu, and Kenneth E. Barner. "Centralized class specific dictionary learning for wearable sensors based physical activity recognition." Information Sciences and Systems (CISS), 2017 51st Annual Conference on. IEEE, 2017.
[14] Mathews, Sherin Mary. Dictionary and deep learning algorithms with applications to remote health monitoring systems. Diss. University of Delaware, 2017.

DOI 10.17148/IJARCCE.2017.61060