Journal of Theoretical and Applied Information Technology
31st March 2015. Vol.73 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

UDFP-TREE: AN EFFICIENT TREE FOR INTERACTIVE MINING OF FREQUENT PATTERNS FROM UNCERTAIN DATA

MOHAMMAD H. NADIMI-SHAHRAKI, FATEMEH HABIBOLLAHI, HAMID RASTEGARI
Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Isfahan, Iran
E-mails: nadimi@iaun.ac.ir, habibolahi@sco.iaun.ac.ir, rastegari@iaun.ac.ir

ABSTRACT

An increasing number of efficient methods have been proposed to mine frequent patterns from uncertain data obtained from real applications such as social networks and life-sciences. Since these data are constantly being updated, needs of the users are changed and they adjust a new minimum support threshold to find new and proper frequent patterns. Obviously, finding the new frequent patterns by running the algorithm from scratch is very costly especially when the database is very large. In this paper, an efficient tree called UDFP-tree is proposed for interactive mining from uncertain data. The proposed tree aims to construct the mining model separated from the mining process. The experimental results show that by using UDFP-tree, there is no need to reconstruct the mining model when user changes the minimum support threshold.

Keywords: *Data Mining, Interactive Mining, Uncertain Data, Frequent Pattern Mining*

1. INTRODUCTION

There are many applications such as social networks and life-sciences in which the content is uncertain where an item is not present or absent, but rather an existence probability of being in the transaction is given. Recently, data mining of uncertain data especially frequent pattern mining has become an interesting research topic. Usually, uncertain databases are large and dynamic, then scalability, efficiency and interactive mining are the problems concerning extraction of frequent pattern from these data. There have been proposed many algorithms [7-16] for frequent pattern mining from uncertain data based on two main methods Apriori and FP-growth. The most of proposed algorithms have been developed for static mining. However, uncertain data are constantly being updated and users frequently change new minimum support threshold to find new and proper frequent patterns. Obviously, running the algorithm from scratch to find new frequent patterns is very costly especially when the database is very large. To solve this problem, researchers propose interactive mining which provides a user with an appropriate opportunity to adjust minimum support threshold. For such a mining, database remains unchanged and only minimum support threshold is changed. Frequent patterns are extracted as per the adjusted minimum support threshold. This principle is called "build once, mine many" [9]. There have been proposed many algorithms [7-16] for frequent pattern mining from uncertain data based on two main methods Apriori and FP-growth. The most of proposed algorithms have been developed for static mining and they are not easily adoptable for interactive mining. Thus, the research question is that "can a tree be designed for interactive mining from uncertain data?" Nadimi-Shahraki et al. [9] discussed this problem generally and proposed a two layer model for interactive mining. In this paper, based on their proposed model, an efficient tree called UDFP-tree is proposed for interactive mining from uncertain data. UDFP-tree is an extension of DFP-tree [19] which is mainly developed for incremental frequent pattern mining from certain data. The proposed tree aims to construct the mining model separated from the mining process. The experimental results show that by using UDFP-tree, there is no need to reconstruct the mining model when user changes the minimum support threshold. The rest of the paper is organized as follows. In the next section, related work is described. The two-layer model is discussed in Section 3. Then, the proposed tree is introduced in Section 4. The experimental results and conclusions are presented in Sections 5 and 6, respectively.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. RELATED WORKS

There have been proposed many algorithms for frequent pattern mining. The first efficient algorithm was proposed by Aggarwal et al. which is called Apriori [10]. The most weaknesses of this algorithm are multi scan of database and generating many candidate patterns. To solve these weaknesses, J. Han et al. proposed FP-growth [11] in 2000. It aims to enhance the efficiency of frequent pattern mining by using a novel tree structure called FP-tree. Efficiently, data is captured and kept by FP-tree by two database scans. Then, to mine frequent patterns, FP-growth algorithm explores the tree in a divide-and-conquer strategy without candidate pattern generation. Although the experimental results show that the performance of FP-growth is very higher than Apriori, FP-growth is not adoptable to interactive mining of frequent patterns because FP-tree kept only frequent patterns.

In 2002, R.P. Gopalan introduced DB-tree algorithm where both frequent and infrequent are kept in the tree. Consequently, this method is easily adoptable to interactive mining of frequent patterns [12]. In 2003, a tree structure called CATS was proposed by W. Cheung [13] for incremental mining of frequent patterns. It improves FP-tree idea for data compression and provided conditions for mining frequent patterns without candidate pattern generation. In this method, data is captured by only one database scan and in spite of FP-tree both frequent and infrequent items are kept which makes this method adoptable to interactive mining.

In 2007, C.K.S Leung [14] proposed a new tree structure called Can-Tree for incremental mining. It captures contents of a transactional database by one database scan. The items in a Can-tree are sorted based on a standard and convectional order and inserted in a tree. Since all items are inserted in the tree, it can be used for interactive mining too.

In 2009, a new tree structure called CP-tree was introduced by S.K Tanbeer et al. [15]. It keeps the content of a database by scanning once and then uses an FP-growth algorithm for extracting frequent patterns. In this method, all items are inserted into a tree based on a logical order and they are inserted into a conventional list called l-list. While items are inserting into a tree, their duplication is registered on the l-list and then the tree is reconstructed. Since all items are kept by CP-tree, this tree is easily adoptable to interactive mining.

Recently, researchers introduced some efficient methods for frequent pattern mining from uncertain data. In 2007, U-Apriori algorithm was proposed by Chui et al. [7] in which it uses the earlier knowledge related to the features of a set of frequent items. It uses a set of k-member items for creating a set of k+1-member items. Consequently, UCP-Apriori algorithm was proposed by C.-K. Chui, B. Kao et al. [16]. This algorithm is based on Apriori property and major steps of its creation are such as U-Apriori algorithm that uses Decremental Pruning (DP) for pruning. This technique applies the statistical features of probabilities of items for reaching fewer candidate patterns during extraction process. Inherently, the efficiency of both algorithms U-Apriori and UCP-Apriori suffer from generating many candidate patterns and multi database scan.

UH-mine algorithm was proposed for frequent pattern mining from uncertain data by Aggarwal et al, in 2009 [17], which is similar to FP-growth, this algorithm follows a direct divide-and-conquer method based on partitioning. UH-mine firstly scans an input database to find frequent items, it removes infrequent items from the database, and the frequent items in each input transaction are stored as per a specific global sorting. The converted database is stored in an array structure in which each row corresponds to a transaction.

Aggarwal et al, in 2009 [17] introduced one of the best algorithms for frequent pattern mining from uncertain data. It adopts the tree structure of UFP-tree for exhibiting a database to find frequent items. This algorithm follows an integrated model by which if the minimum support threshold is changed, then the algorithm must be executed from the scratch.

The above algorithms are mostly developed based on an integrated and single-layer model which cannot be easily used for interactive mining. Thus, in the next sections a suitable two layer model for interactive mining is firstly explained, and then our proposed tree is developed based on this architecture.

3. TWO-LAYER MODEL FOR INTERACTIVE MINING

The most existing algorithms for frequent pattern mining from uncertain data are based on integrated single-layer model and this integrity makes these algorithms unable to be used for interactive mining. In these methods, the tree is constructed regarding to an adjusted minimum support threshold by which only frequent items are kept by the tree. We discussed this problem [9] and show that to support

Journal of Theoretical and Applied Information Technology
31st March 2015. Vol.73 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the principle of "build once, mine many" [8] of interactive mining, the mining model construction must be separated from the mining process. Consequently, we proposed a two-layer model for interactive mining and the experimental results show that by using this model there is no need to reconstruct the mining model when user changes the minimum support threshold. As shown in figure 1, the first layer of proposed model is to construct the mining model and form the mining materials.

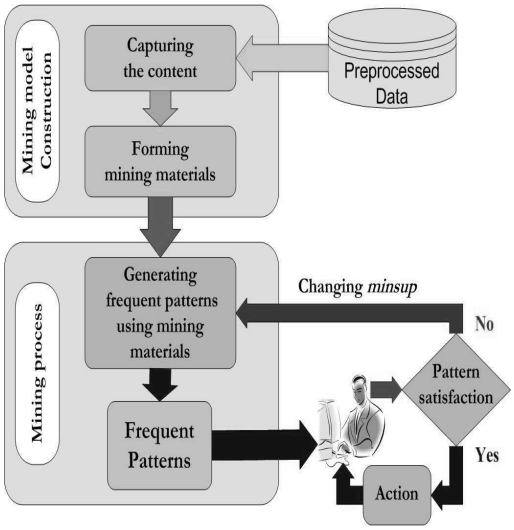

Figure 1: Two-Layer Model For Interactive Mining [9].

Since the mining model is constructed without considering any minimum support threshold or interestingness measurements, it is separated from the mining process. In the second layer, the mining model is explored and frequent patterns are mined regarding to an adjusted minimum support threshold. Thus, the two-layer model is suitable for interactive mining.

4. THE PROPOSED TREE

In this section, based on our previous two-layer model an efficient tree called UDFP-tree is proposed for interactive mining from uncertain data. UDFP-tree is an extension of DFP-tree [18] which introduced for frequent pattern mining from certain data. It aims to separate the mining model from the mining process and there is no need to reconstruct the mining model when user changes the minimum support threshold. The tree and mining model are constructed by one database scan in which all frequent and infrequent item sets are considered. It is possible to extract frequent patterns in the mining process with different thresholds without need to mining model reconstruction. Figure 2 shows the pseudo code of algorithm for UDFP-tree construction.

Algorithm UDFP-tree: *Uncertain Dynamic Frequent Pattern tree*
Input: *A transaction Database DB*
Output: *UDFP-tree*

1. Create Table Transactions;
2. Probability (RandomGenerator);
3. Create the root of UDFP-tree and label it as "Null"
4. for each transaction t_i in Table Transactions
5. {
6. Transaction Sort(ListItems);
7. Clustering Probability;
8. Addchild;
9. Add probability;
10. Items t_i insert into UDFP-tree;
11. }

Figure 2: A Pseudo Code Of The Proposed Tree.

5. EVALUATION

To evaluate the performance of proposed tree for interactive mining, it is experimentally compared with UFP-tree algorithm [17] which is the best algorithms for frequent pattern mining from uncertain data. Since UFP-tree is based an integrated model, it is executed from scratch when the minimum support threshold is changed. Thus, it is expected the performance of proposed tree is more than UFP-tree for interactive mining. Consistently, in each experiment, the algorithms are separately run in the same experimental environment. The experimental evaluation is conducted by the popular synthetic dataset T10I4D100K and real dataset Connect-4. A descending sequence of minimum support threshold for each dataset is considered and the total runtime of mining model construction by using UDFP-tree and UFP-tree for these sequences is observed. Although this experiment was run over several descending sequences, according to the space limitation and similar results observation, the experimental result of only one descending sequence for each dataset is presented. The descending sequence for real dataset Connect-4 consists of five values {90%, 80%, 70%, 60%, 50%} and for synthetic dataset T10I4D100K consists of four descending values {2%, 1.5%, 1%, 0.5%}. The graphs plotted in Figure 3 shows the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

total runtime required for constructing the mining model by different methods over UDFP-tree and UFP-tree for above descending scenarios. UDFP-tree constructs the mining model independent of the minimum support threshold, thus it fits "build once, mine many" principle and there is no need to reconstruct the mining model when the minimum support threshold is changed.

Therefore, the cost of mining model construction by UDFP-tree for the descending sequence of minimum support threshold is fixed. However, the UFP-tree must be reconstructed for each minimum support threshold and the total runtime of mining model construction is increased.

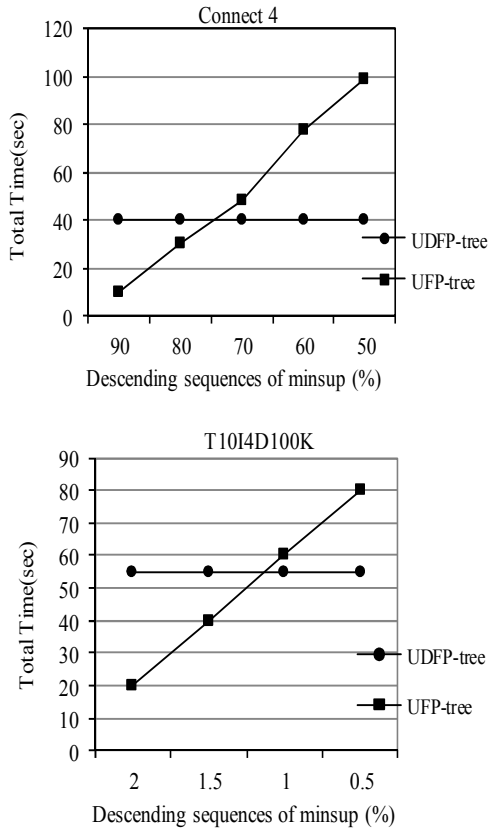

Figure 3: Total Runtime Of Mining Model Construction For Interactive Mining.

6. CONCLUSION

Recently, data mining of uncertain data especially frequent pattern mining has become an interesting research topic. On the other hand, in many applications, interactive mining is a need where user changes minimum support threshold to find proper frequent patterns. Thus, in this paper, an efficient tree is proposed called UDFP-tree based on our previous model [9] for interactive mining from uncertain data. By using the proposed tree, the mining model construction is separated from the mining process, and it is possible to build mining model once and use it many times. To evaluate the proposed tree, it was compared with UFP-tree algorithm which is the best algorithms for static frequent pattern mining from uncertain data. The experiments results show that although the runtime of UFP-tree for static mining was less than UDFP-tree, in interactive mining, the UFP-tree must be reconstructed and the number of changes in minimum support threshold increases its cost. Therefore, after UDFP-tree is constructed it can be frequently mined by different minimum support threshold.

REFRENCES:

[1] Bernecker, Thomas, et al. "Probabilistic frequent itemset mining in uncertain databases ", *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009.

[2] Calders, Toon, Calin Garboni, and Bart Goethals. "Approximation of frequentness probability of itemsets in uncertain data", *Data Mining (ICDM)*, 2010 IEEE 10th International Conference on. IEEE, 2010.

[3] Calders, Toon, Calin Garboni, and Bart Goethals. "Efficient pattern mining of uncertain data with sampling", *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2010.pp. 480-487.

[4] Leung, CK-S., and Boyu Hao. "Mining of frequent itemsets from streams of uncertain data", *Data Engineering, 2009.ICDE'09. IEEE 25th International Conference on*. IEEE, 2009.

[5] Leung, Carson Kai-Sang, and Syed K. Tanbeer. "Fast tree-based mining of frequent itemsets from uncertain data", *Database Systems for Advanced Applications*. Springer Berlin Heidelberg, 2012.

[6] Zhang, Qin, Feifei Li, and Ke Yi. "Finding frequent items in probabilistic data", *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008.

[7] Chui, Chun-Kit, Ben Kao, and Edward Hung. "Mining frequent itemsets from uncertain data", *Advances in knowledge discovery and data mining.Springer Berlin Heidelberg*, 2007.pp, 47-58.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[8] Leung, Carson Kai-Sang, Mark Anthony F. Mateo, and Dale A. Brajczuk. "A tree-based approach for frequent pattern mining from uncertain data", *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2008.pp. 653-661.

[9] Nadimi-Shahraki, M. H., and Norwati Mustapha. "A Two-layer Model for Interactive Mining of Frequent Patterns", *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Vol. 1. 2011.

[10] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules", Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.

[11] Han, Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation", *ACM SIGMOD Record*. Vol. 29. No. 2. ACM, 2000.

[12] Ezeife, Christie I., and Yue Su. "Mining incremental association rules with generalized FP-tree", *Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 2002.pp. 147-160.

[13] Cheung, William, and Osmar R. Zaiane. " Incremental mining of frequent patterns without candidate generation or support constraint", *Database Engineering and Applications Symposium, 2003. Proceedings. Seventh International*. IEEE, 2003.

[14] Leung, Carson Kai-Sang, et al. "CanTree: a canonical-order tree for incremental frequent-pattern mining", Knowledge and Information Systems 11.3 (2007). pp. 287-311.

[15] Tanbeer, Syed Khairuzzaman, et al. "Efficient single-pass frequent pattern mining using a prefix-tree", Information Sciences 179.5 (2009): pp. 559-583.

[16] Chui, Chun-Kit, and Ben Kao. "A decremental approach for mining frequent itemsets from uncertain data", *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2008.pp. 64-75.

[17] Aggarwal, Charu C., et al. "Frequent pattern mining with uncertain data", *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009.

[18] Hamedanian, Mohadeseh, Mohammad Nadimi, and Mohammad Naderi. "An Efficient Prefix Tree for Incremental Frequent Pattern Mining", International Journal of Information 3.2 (2013).