# A Survey Paper on Frequent Pattern Mining for Uncertain Database

**Jigisha V. Patel**[1]
[1]LJIET, Ahmedabad

*Abstract*— There are number of existing algorithms proposed that mines frequent patterns from certain or precise data. But know a day's demand of uncertain data mining is increased. There are many situations in which data are uncertain. For frequent pattern mining from uncertain data mainly two approaches are proposed that are level-wise approach and pattern-growth approach. Level-wise approach use the generate and test strategy. U-Apriori algorithm is the example of Level-wise approach. Pattern-growth approach uses tree like structure. UF-growth algorithm, UFP-growth algorithm, CUFP-mine algorithm, PUF-growth algorithm are the examples of pattern-growth approach. Here we are taking the survey of algorithms that are used to mine frequent patterns from uncertain data.

***Key words:*** Frequent Pattern, Uncertain Database, U-Apriori algorithm

## I. INTRODUCTION

Most of the mining approaches for frequent patterns are based on precise or certain data. But sometimes data is associated with uncertainty because of measurement inaccuracy, missing and approximate values, outdated data sources and other errors. Reason behind uncertainty is that user is unknown about the presence and absence of item or event in transaction. For example in medical transaction, physician may highly suspect about patient suffers from flu or patient suffers from several possible diseases with different possibility which shows the existential probability about patients diseases. Sensor databases and satellite images are the other example of uncertain data. In sensor network user is not sure about the presence or absence of noise from sensor databases while in satellite images user is uncertain about the object is present or absent in image. It is useful to mine the position of object so we can see proper image resolution.

Many algorithms are implemented to mine frequent pattern from uncertain data, from which U-Apriori algorithm is most popular which is based on candidate generation. But multiple scanning is required in U-Apriori so other tree based algorithms are proposed.

## II. FREQUENT PATTERN MINING FOR UNCERTAIN DATABASE

### A. U-Apriori algorithm

Chau et al. Proposed an Apriori based algorithm known as U-Apriori [1].The algorithm uses generate and test strategy to mine frequent items from uncertain database. According to existential probability it finds expected support of each Itemset and then finds all expected support based frequent itemset using generate and test strategy approach. Now it joins expected support based items and then generates candidates. These generated candidates are tested based on expected support and user specified minimum support. The process is repeated until we generate frequent itemset. Using trimming strategy we can further improves the efficiency of algorithm.

Limitation of U-Apriori algorithm is that it requires multiple scan and it does not scale well when dealing with large amount of data.

### B. UF-Growth algorithm

UF-growth algorithm does not require candidate generation. It is a tree based algorithm. There are mainly two operations required to find frequent items from uncertain data, which are as follows,

*1) UF-tree construction*

It captures uncertain data from a transaction. For that it deals with existential probability that defines the presence of item in the transaction.

UF-tree stores an item, its expected support and occurrence count. And according to the expected support it puts item one by one and makes tree. Item that contain the occurrence count, less then support count are removed.

The node which contains same item and same existential probability are merged and occurrence count is incremented by one. It continues the process until all the items are added in the tree. The process of generating tree is shown below.

| TID | Items |
|---|---|
| t1 | { a:0.2 , b:0.2 , c:0.7 , f:0.8 } |
| t2 | {a:0.5 , c:0.9 , e:0.5} |
| t3 | { a:0.3 , e:0.4 , f:0.5 , d:0.5} |
| t4 | { a:0.9 , b:0.2 , e:0.5 , d:0.1} |

Table. I: Uncertain database [2]

| Item | *ExpSup* |
|---|---|
| a | 1.90 |
| b | 0.40 |
| c | 1.60 |
| d | 0.60 |
| e | 1.40 |
| f | 1.30 |

→

| Item | *ExpSup* |
|---|---|
| a | 1.90 |
| c | 1.60 |
| e | 1.40 |
| f | 1.30 |
| d | 0.60 |

Table II: Expected support count [2]

Here we have taken minimum support is 0.5. So from above table item b is removed because it has a expected count 0.40,which is less then minimum support count.

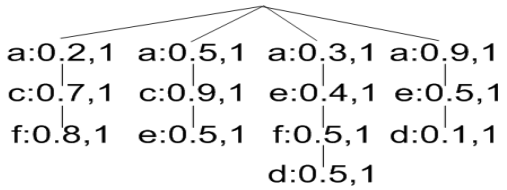


Fig. 1: UF-tree [2]

*2) Mining or growing frequent patterns*

After UF-tree is generated UF-growth algorithm is applied to mine frequent patterns. The expected count of an itemset is calculated by the multiplication of the expected supports of all items in the corresponding path [3].

Limitation of UF-growth algorithm is that, it may not be too compact because path in the corresponding UF-tree are shared only if tree nodes on the path have the same item and same existential probability [2].

 

### *C. UFP-Growth Algorithm*

UFP-growth algorithm is extended from the FP-growth algorithm which is one of the most well-known pattern mining algorithms in deterministic databases [4]. To make the UF-tree more compact UFP-growth algorithm is proposed. First it builds UFP-tree and then mines frequent pattern using UFP-growth algorithm.

It works similar to the UF-growth but the difference is that we partition the probabilities into set of k clusters. Here we store the maximum probability value in each cluster instead of center of the cluster because to make sure that the support computed from the summary is no less than the true support [5]. It stores an item and its maximum existential probability. After generating UFP-tree we mine frequent patterns from the generated tree using UFP-growth algorithm.

### *D. CUFP-Growth algorithm*

It is the compressed structure of UF-tree. In construction of UF-tree only same item with same existential probabilities are merged together in the tree which creates some redundant node. So for removing the redundant node CUFP-tree algorithm is proposed. In CUFP-tree algorithm same item can be merged even their existential probabilities may differs.

After tree construction CUFP-mine algorithm is applied to mine frequent patterns. CUFP-tee algorithm runs faster than UF-tree algorithm [3].

### *E. PUF-Growth algorithm*

Tree based algorithm require larger memory to store the tree so it is needful to make tree as compact as possible. PUF-tree algorithm provides compact tree structure for frequent pattern mining of uncertain data. To make the UF-tree and UFP-tree compact PUF-tree algorithm is proposed [2].

Key feature of the algorithm is that it is a prefix – capped uncertain frequent pattern tree (PUF-tree), which is as compact as original FP-tree [2]. The algorithm guarantees that it only produces the frequent patterns with no false positives or no false negatives. PUF-tree stores an item and its prefixed item cap. Prefixed item cap is defined as the upper bound of an item in transaction.

PUF-tree is compact then UF-tree and UFP-tree because UF-tree contains multiple node for the same item and UFP-tree stores extra cluster information. PUF-growth algorithm provides scalability and in shorter amount of time it mines frequent patterns from large database.

### *F. Comparison*

| ALGORITHMS | Comments |
|---|---|
| U-Apriori | It uses generate and test approach. So multiple scanning is required. |
| UF-growth | It is a tree based algorithm that overcomes limitation of U-Apriori because it does not require candidate generation. |
| UFP-growth | To make the UF-tree compact the algorithm is proposed that partition the probabilities into set of k clusters. |
| CUFP-Mine | It is the compressed structure of UF-tree that runs faster than UF-growth algorithm. |
| PUF-tree | To make UF-tree and UFP-tree compact PUF-tree is proposed that is scalable and guarantees that only frequent patterns are generated. |

## III. CONCLUSION

Mining frequent patterns from the uncertain database is a very crucial task. There are many approaches that have been discussed; nearly all of the previous studies were using Level-wise approach and Pattern-growth approach for mining the frequent patterns from uncertain database. Thus the goal of this research was to find frequent patterns from uncertain database using the approach that requires minimum time and minimum memory.

## REFERENCES

[1] Carson Kai-Sang Lenug, Boyu Hao,"Mining of Frequent Pattern from Streams of Uncertain Data", IEEE-2009.
[2] Carson Kai-Sang Lenug and Syed Khairuzzaman Tanbeer, "PUF-Tree: A Compact Tree Structure for Frequent Pattern Mining of Uncertain Data", Springer, Part I, pp 13-25, 2013.
[3] Chun-Wei Lin, Tzung -Pei Hong, "A new mining approach for uncertain databases using CUFP trees", Science Direct, 2011.
[4] Yongxin Tong, Yurong Cheng, Philip S.Yu,"Mining Frequent Itemsets over Uncertain Databases" ,Vol. 5,No. 11,2012
[5] Charu C.Aggarwal,Yan Li,"Frequent Pattern Mining with Uncertain Data" ,2009
[6] Feng Gao, Chengrong Wu, "Mining Frequent Itemset from Uncertain Data", IEEE-2011.
[7] Carson Kai-Sang Lenug, Christopher L. Carmichael," Efficient Mining of Frequent Patterns from Uncertain Data", IEEE-2007.
[8] Liang Wang, David Wai-Lok, Cheung, Reynold Cheng," Efficient Mining of Frequent Item Sets on Large Uncertain Databases", IEEE, VOL.24, NO.12, December 2012.
[9] Le Wang, Lin Feng, Mingfei Wu,"AT-Mine: An Efficient Algorithm of Frequent Itemset Mining on Uncertain Dataset", journal of computers, VOL. 8, NO. 6, june-2013.
[10] Michel Chau, Dr. Reynold Cheng, Ben Kao, "Uncertain Data Mining: A New Research Direction", December 2005.

 