UFIMT: An Uncertain Frequent Itemset Mining Toolbox

Yongxin Tong † Lei Chen † Philip S. Yu §
†Hong Kong University of Science & Technology, Hong Kong, China
§University of Illinois at Chicago, USA
†{yxtong, leichen}@cse.ust.hk, §psyu@cs.uic.edu

ABSTRACT

In recent years, mining frequent itemsets over uncertain data has attracted much attention in the data mining community. Unlike the corresponding problem in deterministic data, the frequent itemset under uncertain data has two different definitions: the expected support-based frequent itemset and the probabilistic frequent itemset. Most existing works only focus on one of the definitions and no comprehensive study is conducted to compare the two different definitions. Moreover, due to lacking the uniform implementation platform, existing solutions for the same definition even generate inconsistent results. In this demo, we present a demonstration called as *UFIMT* (Uncertain Frequent Itemset Mining Toolbox) which not only discovers frequent itemsets over uncertain data but also compares the performance of different algorithms and demonstrates the relationship between different definitions. In this demo, we firstly present important techniques and implementation skills of the mining problem, secondly, we show the system architecture of *UFIMT*, thirdly, we report an empirical analysis on extensive both real and synthetic benchmark data sets, which are used to compare different algorithms and to show the close relationship between two different frequent itemset definitions, and finally we discuss some existing challenges and new findings.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms, Experimentation

Keywords

Frequent Itemset Mining, Uncertain Database, UFIMT

1. INTRODUCTION

With the emerging of many new applications, such as sensor network monitoring, protein-protein interaction (PPI) network analysis, etc., uncertain data mining has become a new challenge in data mining communities. Due to the fundamental status of frequent itemset mining in the data mining field, mining frequent itemsets over uncertain databases has also attracted much attention [1, 2, 3, 4, 5, 6, 8, 9] recently. For instance, with the popularization of wireless sensor networks, wireless sensor network systems collect huge amount of data. However, due to the inherent uncertainty of sensors, the collected data is often inaccurate. For the probability-included uncertain data, how can we discover frequent patterns (itemsets) so that the users can find the hidden rules in data? The inherent probability property of data is ignored if we simply employ the traditional method of frequent itemset mining in deterministic data to uncertain data. Hence, it is necessary to develop specialized algorithms for mining frequent itemsets over uncertain databases.

Unlike the research of frequent itemset mining in deterministic data, the corresponding problem in uncertain data presents different semantic explanations and problem definitions. In deterministic data, frequent itemset has a unique definition, that is, an itemset is frequent if and only if its support (frequency) is no less than a specified minimum support, *min_sup*. However, the definition of a frequent itemset over uncertain data has two different semantic explanations: expected support-based frequent itemset [4] and probabilistic frequent itemset [2], both of which consider the support of an itemset as a discrete random variable. However, the two definitions employ different semantic explanations for the random variable. The former defines the expectation of the support of an itemset as the measurement, denoted as the expected support of this itemset. Thus, an itemset is frequent if and only if its expected support is no less than a specified minimum expected support threshold, *min_esup*. The latter uses the probability that an itemset appears at least the specified minimum support (*min_sup*) times as the measurement, denoted as the frequent probability of an itemset, therefore, an itemset is frequent if and only if its frequent probability is larger than a given probabilistic threshold. Therefore, most prior works study the above two definitions independently [2, 6].

In this work, through experimental verification and theoretical analysis, we find that the two definitions have a rather close connection. Generally speaking, this finding is reasonable since both definitions consider the support of an itemset as a random variable following a Poisson Binomial distribution, the expected support of an itemset equals to the expectation of the random variable. As a result, calculating the frequent probability of an itemset is equivalent to calculating the cumulative distribution function of this random variable. In addition, the existing mathematical theory shows that a Poisson distribution or a Normal distribution can approximate a Poisson Binomial distribution under high

KDD'12, August 12–16, 2012, Beijing, China.

confidence [3, 7, 9]. Thus, an interesting result is that the frequent probability of an itemset can be directly computed as long as we know the expected and variance of the support of the itemset when the uncertain database is enough large [7]. So, the efficiency of mining probabilistic frequent itemsets can be greatly improved by employing many efficient expected support-based frequent itemset mining algorithms. In this demo, based on our system, *UFIMT*, we verify this conclusion through extensive experimental comparisons.

Besides ignoring the hidden relationship between two above definitions, existing research on the same definition also shows contradictory conclusions. For example, in the research of mining expected support-based frequent itemsets, [5] shows that UFP-growth algorithm always outperforms UApriori algorithm with respect to the running time. However, [1] reports that UFP-growth algorithm is always slower than UApriori algorithm. These inconsistent conclusions make later researchers confused about which result is correct.

In addition, different experimental results also originate from the discrepancy among many implementation skills, blurring what the contributions of algorithms are. Therefore, a uniform baseline implementation is necessary, which is one of our important motivations to develop *UFIMT*.

Except verifying the relationship between two above definitions, clarifying inconsistent conclusions in current related researches, and providing uniform baseline implementations, we also study another variant problem, mining probabilistic frequent closed itemsets over uncertain data recently [8], which aims to compress the size of all frequent itemsets in order to help user better understand the result set. Therefore, the prototype system, UFIMT, provides solutions of some related problems of frequent itemset mining over uncertain data as well. For example, UFIMT can discover the probabilistic frequent closed itemsets.

Thus, in this demo, we try to achieve the following goals:

- Present a novel prototype system, $UFIMT$, for mining frequent itemsets over uncertain databases. The prototype system seamlessly integrates the uncertain data preprocessing module, the algorithm library of mining frequent itemsets over uncertain data, and the comparison and analysis of algorithms.
- Clarify the relationship of the two definitions of frequent itemsets over uncertain data. In fact, the two definitions can be integrated together due to the mathematical correlation between them. Based on this relationship, instead of paying expensive computation cost to mine probabilistic frequent itemsets, we can directly use the solutions for mining expected support-based itemsets when the size of data is large enough.
- Provide uniform baseline implementations for nine existing representative algorithms for mining frequent itemsets and mining frequent closed itemsets under uncertain databases. These implementations adopt common basic operations and offer a base for comparing with the future work in this area.
- Propose an objective and sufficient experimental evaluation platform. Meanwhile, we test the performances of the existing representative algorithms over extensive benchmarks and verify the contradictory conclusions in the existing research.

The rest of the demo is organized as follows. In Section 2, we introduce the technical specification of our prototype system, *UFIMT* (Uncertain Frequent Itemset Mining Toolbox). A system demonstration plan is proposed in Section 3. We conclude in Section 4.

2. TECHNICAL SPECIFICATION

In this section, we firstly introduce basic definitions about mining frequent itemsets over uncertain databases. Then, existing representative algorithms under the above two definitions are briefly reviewed. Finally, the system architecture of *UFIMT* is described and illustrated.

2.1 Frequent Itemsets over Uncertain Data

In this subsection, we generally introduce several basic definitions about mining frequent itemsets over uncertain databases.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of distinct items. We name a non-empty subset, X, of I as an itemset. For brevity, we use $X = x_1x_2 \ldots x_n$ to denote itemset $X = \{x_1, x_2, \ldots, x_n\}$. X is a $l - itemset$ if it has l items. Given an uncertain transaction database UDB, each transaction is denoted as a tuple $< tid, Y >$ where tid is the transaction identifier, and $Y = \{y_1(p_1), y_2(p_2), \ldots, y_m(p_m)\}$. Y contains m units. Each unit has an item y_i and a probability, p_i, denoting the possibility of item y_i appearing in the tid tuple. The number of transactions containing X in UDB is a random variable following the Poisson Binomial distribution, denoted as $sup(X)$. Given an UDB, the expected support-based frequent itemset and the probabilistic frequent itemset are defined as follows.

Definition 1. (Expected Support) Given an uncertain transaction database UDB which includes N transactions, and an itemset X, the expected support of X is shows as follows:

$$esup(X) = \sum_{i=1}^{N} p_i(X)$$

Definition 2. (Expected-Support-based Frequent Itemset) Given an uncertain transaction database UDB which includes N transactions, and a minimum expected support ratio, min_esup, an itemset X is an expected support-based frequent itemset if and only if $esup(X) \geq N \times min\_esup$

Definition 3. (Frequent Probability) Given an uncertain transaction database UDB which includes N transactions, a minimum support ratio min_sup, and an itemset X, X's frequent probability, denoted as $Pr(X)$, is shown as follows:

$$Pr(X) = Pr\{sup(X) \geq N \times min\_sup\}$$

Definition 4. (Probabilistic Frequent Itemset) Given an uncertain transaction database UDB which includes N transactions, a minimum support ratio min_sup, and a probabilistic frequent threshold pft, an itemset X is a probabilistic frequent itemset if X's frequent probability is larger than the probabilistic frequent threshold, namely,

$$Pr(X) = Pr\{sup(X) \geq N \times min\_sup\} > pft$$

2.2 Uncertain Frequent Itemset Mining

In this subsection, we briefly review eight existing representative algorithms of mining frequent itemsets over uncertain data. These algorithms are also seamlessly integrated into our prototype system, *UFIMT*. We categorize the eight representative algorithms into three groups: expected support-based frequent itemset mining algorithms, exact

probabilistic frequent itemset mining algorithms, and approximate probabilistic frequent itemset mining algorithms.

Expected support-based frequent itemset mining algorithms. These algorithms aim to find all expected support-based frequent itemsets. The complexity of computing the expected support of an itemset is $O(N)$, where N is the number of transactions.

In this categorization, *UFIMT* contains three representative algorithms: UApriori [4], UFP-growth [5], and UH-Mine [1]. UApriori is the first expected support-based frequent itemset mining algorithm which extends the well-known Apriori algorithm to the uncertain environment and employs the generate-and-test framework to find all expected support-based frequent itemsets. In our system, we also modify the Trie-Tree-based Apriori framework to speed up our algorithm. UFP-growth algorithm extends the well-known FP-growth algorithm. Similar to the traditional FP-growth algorithm, UFP-growth algorithm also firstly builds an index tree, called UFP-tree to store all information of the uncertain database. Then, based on the UFP-tree, the algorithm recursively builds conditional subtrees and finds expected support-based frequent itemsets. UH-Mine is also based on the divide-and-conquer framework. The algorithm is extended from the H-Mine algorithm which is a classical algorithm in deterministic frequent itemset mining. Similar to H-Mine, UH-Mine first builds the special data structure, UH-Struct, and then recursively discovers the expected support-based frequent itemsets based on the DFS strategy.

Exact probabilistic frequent itemset mining algorithms. These algorithms discover all probabilistic frequent itemsets and report exact frequent probability for each itemset. Due to computing the exact frequent probability instead of the simple expectation, these algorithms need to spend at least $O(Nlog^2N)$ computation cost for each itemset. Moreover, in order to avoid redundant processing, the Chernoff bound-based pruning is the main factor that affects the running time of the algorithms in this.

In this categorization, *UFIMT* contains two algorithms: DP (Dynamic Programming-based Apriori algorithm) [2] and DC (Divide-and-Conquer-based Apriori Algorithm) [6]. Both of them are based on the Apriori framework. The main difference between the two algorithms is the method for calculating the frequent probability of each itemset. The time complexity of the DP algorithm calculating the frequent probability is $O(N^2 \times min\_sup)$, however, that of DC algorithm is $O(Nlog^2N)$.

Approximate probabilistic frequent itemset mining algorithms. These algorithms can obtain the approximate frequent probability with high quality by only acquiring the first moment (expectation) and the second moment (variance). Due to the sound properties of the Poisson Binomial distribution, the time complexities of calculating the expectation and the variance of each itemset are $O(N)$ (The first type of algorithms have the same time complexity). Therefore, these algorithms are much faster than the exact probabilistic frequent itemset mining algorithms but obtain the similar probability information when the uncertain databases are large enough. To sum up, the third type of algorithms actually build a bridge between two different definitions of frequent itemsets over uncertain databases.

In this categorization, *UFIMT* contains three algorithms: PDUApriori (Poisson Distribution-based UApriori) [9], NDUApriori (Normal Distribution-based UApriori) [3], and NDUH-Mine (Normal Distribution-based UH-Mine) [7]. Because the support of an itemset follows a Poisson Binomial distribution that can be approximated by a Poisson distribution, based on cumulative distribution function (CDF) of the Poisson distribution, PDUApriori solves the corresponding expected support λ of the given pft and calls UApriori to find all approximate probabilistic frequent itemsets. Moreover, according to the *Lyapunov Central Limit Theory*, NDUApriori algorithm uses the cumulative distribution function of a standard Normal distribution to approximately calculate the frequent probability.

Although the above two algorithms can obtain good approximation, it is impractical to apply them on very large sparse uncertain databases due to their Apriori framework. Therefore, we propose a novel algorithm, NDUH-Mine which integrates the framework of UH-Mine and the Normal distribution approximation in order to achieve a win-win partnership in sparse uncertain databases. In the other words, we calculate the variance of each itemset when we obtain the expected support of each itemset. Through extensive experiments over *UFIMT*, we can observe that NDUH-Mine has a better performance than PDUApriori and NDUApriori on very large sparse uncertain databases, which confirms the goal of our design.

Therefore, the Normal distribution-based approximation algorithms build a bridge between the definition of expected support-based frequent itemset mining and definition of the probabilistic frequent itemset mining. In particular, existing efficient expected support-based mining algorithms can be directly reused in the problem of mining probabilistic frequent itemsets and retain their intrinsic properties. In addition, more experimental comparisons and analysis about above algorithms can be found in [7].

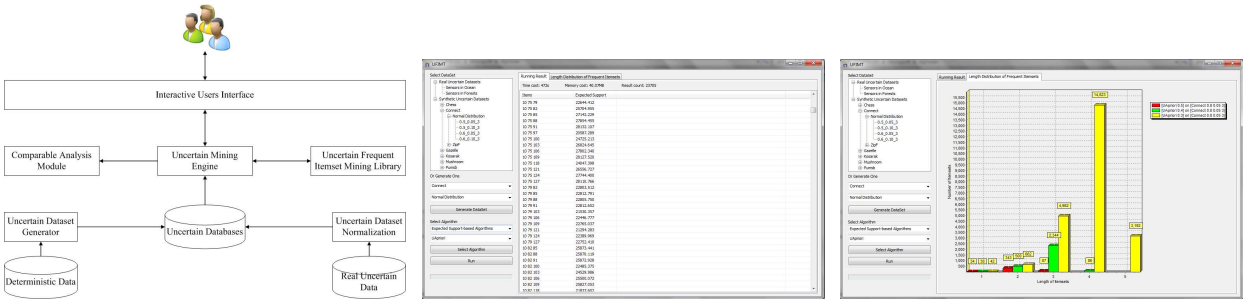
2.3 System Architecture of UFIMT

In this subsection, we focus on our system architecture of *UFIMT*. Figure 1(a) shows the software system architecture of *UFIMT*. It consists of the following modules.

Data Preprocessing Module: it provides the normalized uncertain data. In our prototype system, uncertain data originates from two kinds of data sources. The first kind of data comes from a real sensor monitoring network. Because different kinds of sensors generate different formal data, our system needs normalizes different kinds of uncertain data into the unique data format. Moreover, the second kind of data is generated from some classical deterministic benchmarks, e.g. mushroom, connect, etc. In our data preprocessing module, it contains an uncertain data generator which assigns probabilities to deterministic data following a Normal distribution or a Zipf distribution.

Uncertain Mining Engine and Algorithm Library: *UFIMT* includes an uncertain mining engine which can compare the performance of mining algorithms and analyze the mining results according to users' requirements. In addition, an algorithm library is an important module in *UFIMT* as well. The algorithm library is consisted of the eight representative uncertain frequent itemset mining algorithms above and the novel variant algorithm developed by us, mining probabilistic frequent closed itemsets [8], respectively. All algorithms are based on the same basic functions, so that the comparable results are fair.

Comparable Analysis and User Visualization: Different from the traditional frequent pattern mining tools, *UFIMT* not only shows the result of single mining algorith-



(a) System Architecture of *UFIMT* (b) Displaying A Running result in *UFIMT* (c) Length distribution of frequent itemsets

Figure 1: System Architecture and Screenshots of *UFIMT*

m, but also supports the performance comparison of different algorithms. For the visualization of the result of the single algorithm, *UFIMT* can show the running time, the memory usage, all frequent itemsets and their probability information. Moreover, *UFIMT* also uses the histograms to show all kinds of distributions of frequent itemsets. In particular, *UFIMT* can also demonstrate the effect of mining frequent closed itemsets. On the other hand, to visualize the comparison of multi-algorithms, *UFIMT* uses the visual methods to report the differences among the performances of the algorithms. Two screenshots of *UFIMT* are shown in Figure 1(b) and Figure 1(c), respectively. Figure 1(b) shows a running result of UApriori algorithm. Figure 1(c) reports the length distribution of all frequent itemsets under different minimum expected supports.

3. DEMONSTRATION PLAN

In the design and implementations of our prototype, *UFIMT*, we mainly provide a practical platform of mining frequent itemsets over uncertain databases. In addition, *UFIMT* seamlessly integrates eight existing representative algorithms of mining frequent itemsets over uncertain data [7] and a few novel variant solutions, such mining probabilistic frequent closed itemsets over uncertain databases [8].

We will introduce our prototype system thoroughly in our demo. In particular, we will focus on the following aspects.

Firstly, we will show the overview of *UFIMT*, including the motivations and the problems handled by this prototype system. We will briefly introduce the difference between frequent itemset mining over deterministic data and uncertain data to the audience. In particular, two different definitions of frequent itemsets over uncertain data will be explained, and several challenges about this topic will be reviewed, such as the contradictory conclusions in current researches.

Secondly, we will report the system architecture and the technical details in *UFIMT*, including the implementation details. We will discuss implementation details of all mining algorithms. In particular, we will show the audience how *UFIMT* implements the common basic functions in an efficient and effective way.

Thirdly, we will demonstrate our prototype system over both real and synthetic uncertain databases which includes on two real uncertain databases and six synthetic uncertain databases. We will show our visualization of *UFIMT* through the live presentation as well.

Finally, we will show our prototype system and invite audiences to use *UFIMT*. Audiences are encouraged to use the system on all kinds of datasets in order to further understand how to find frequent itemsets over uncertain databases.

4. CONCLUSIONS

In this demo, we present a novel prototype system, *UFIMT*, which provides a practical platform for mining frequent itemsets over uncertain databases. Since there are two definitions of frequent itemsets over uncertain data, most existing researches are categorized into two directions. However, through our explorations, *UFIMT* firstly clarifies that there is a close relationship between the two different definitions. Secondly, *UFIMT* provides baseline implementations of eight existing representative algorithms and tests their performances. Finally, through extensive experiments, *UFIMT* verifies several existing inconsistent conclusions and finds some new rules on this topic.

5. ACKNOWLEDGMENTS

This work is supported in part by the Hong Kong RGC GRF Project No.611411, National Grand Fundamental Research 973 Program of China under Grant 2012-CB316200, HP IRP Project 2011, Microsoft Research Asia Grant, MRA11EG05, US NSF grants DBI-0960443, CNS-1115234, and IIS-0914934, and Google Mobile 2014 Program.

6. REFERENCES

[1] C. Aggarwal, Y. Li, J. Wang, and J. Wang. Frequent pattern mining with uncertain data. In *KDD'09*.

[2] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Züfle. Probabilistic frequent itemset mining in uncertain databases. In *KDD'09*.

[3] T. Calders, C. Garboni, and B. Goethals. Approximation of frequentness probability of itemsets in uncertain data. In *ICDM'10*.

[4] C. K. Chui, B. Kao, and E. Hung. Mining frequent itemsets from uncertain data. In *PAKDD'07*.

[5] C. K.-S. Leung, M. A. F. Mateo, and D. A. Brajczuk. A tree-based approach for frequent pattern mining from uncertain data. In *PAKDD'08*.

[6] L. Sun, R. Cheng, D. W. Cheung, and J. Cheng. Mining uncertain data with probabilistic guarantees. In *KDD'10*.

[7] Y. Tong, L. Chen, Y. Cheng, and P. S. Yu. Mining frequent itemsets over uncertain databases. In *VLDB'12*.

[8] Y. Tong, L. Chen, and B. Ding. Discovering threshold-based frequent closed itemsets over probabilistic data. In *ICDE'12*.

[9] L. Wang, R. Cheng, S. D. Lee, and D. W.-L. Cheung. Accelerating probabilistic frequent itemset mining: a model-based approach. In *CIKM'10*.