A Novel Approach on Mining Frequent Item Sets on Large Uncertain Databases

R. Manimegalai[1], D. Dhanabagyam[2]

[1]M. Phil Scholar, Department of Computer Science, Sri Jeyendra Saraswathi College of Arts and Science, Coimbatore – 641 005, India

[2]Assistant Professor, Department of Computer Applications, Sri Jeyendra Saraswathi College of Arts and Science, Coimbatore – 641 005, India

Abstract: ***The data handled in emerging applications like location-based services, sensor monitoring systems and data integration, are often inexact in nature. In this paper, we study the important problem of extracting frequent item sets from a large uncertain database, interpreted under the Possible World Semantics (PWS). This issue is technically challenging, since an uncertain database contains an exponential number of possible worlds. By observing that the mining process can be modeled as a Poisson binomial distribution, we develop an approximate algorithm, which can efficiently and accurately discover frequent item sets in a large uncertain database. The important issue of maintaining the mining result for a database that is evolving was discussed. Specifically, we propose incremental mining algorithm, which enable Probabilistic Frequent Item Set (PFI) results to be refreshed. This reduces the need of re-executing the whole mining algorithm on the new database, which is often more expensive and unnecessary. All our approaches support both tuple and attribute uncertainty, which are two common uncertain database models. We also perform extensive evaluation on real and synthetic data sets to validate our approaches.***

Keywords: Sensor, Poisson binominal distribution, probabilistic Frequent Item set, incremental mining

1. Introduction

The databases used in many important and novel applications are often uncertain. For example, the locations of users obtained through RFID and GPS systems are not precise due to measurement errors [1], [2]. As another example, data collected from sensors in habitat monitoring systems (e.g., temperature and humidity) are noisy [3]. Customer purchase behaviors, as captured in supermarket basket databases, contain statistical information for predicting what a customer will buy in the future [4, 5]. Integration and record linkage tools also associate confidence values to the output tuples according to the quality of matching [6]. In structured information extractors, confidence values are appended to rules for extracting patterns from unstructured data [7]. To meet the increasing application needs of handling a large amount of uncertain data, uncertain databases have been recently developed [8-12].

2. Problem Definition

Mining frequent item sets is an important problem in data mining, and is also the first step of deriving association rules [13]. Hence, many efficient item set mining algorithms (e.g., Apriori [13] and FP-growth [14]) have been proposed. While these algorithms work well for databases with precise values, it is not clear how they can be used to mine probabilistic data. Here we develop algorithms for extracting frequent item sets from uncertain databases. Although our algorithms are developed based on the Apriori framework, they can be considered for supporting other algorithms (e.g., FP-growth) for handling uncertain data.

3. Related Work

For uncertain databases, Aggarwal et al. [15] and Chui et al. [16] developed efficient frequent pattern mining algorithms based on the expected support counts of the patterns. However, Bernecker et al. [5], Sun et al. [18], and Yiu et al. [12] found that the use of expected support may render important patterns missing. Hence, they proposed to compute the probability that a pattern is frequent, and introduced the notion of PFI. In [5], dynamic-programming based solutions were developed to retrieve PFIs from attribute-uncertain databases. However, their algorithms compute exact probabilities, and verify that an item set is a PFI in Oðn2Þ time. Our model-based algorithms avoid the use of dynamic programming, and are able to verify a PFI much faster (in OðnÞ time). In [16], approximate algorithms for deriving threshold-based PFIs from tuple-uncertain data streams were developed. While in [4] only considered the extraction of singletons (i.e., sets of single items), our solution discovers patterns with more than one item. Recently, Sun et al. [17] developed an exact thresholdbased PFI mining algorithm. However, it does not support attribute-uncertain data considered in this paper. In a preliminary version of this paper [33], we examined a model-based approach for mining PFIs. Here, we study how this algorithm can be extended to support the mining of evolving data. To our best knowledge, maintaining frequent item sets in evolving uncertain databases has not been examined before. We propose novel incremental mining algorithms for both exact and approximate PFI discovery. Our algorithms can also support attribute and tuple uncertainty models

4. Proposed Methodology

Many efficient item set mining algorithms (e.g., Apriori and FP-growth) have been proposed. While these algorithms work well for databases with precise values, it is not clear how they can be used to mine Probabilistic data. We develop algorithms for extracting frequent item sets from uncertain databases. Although our algorithms are developed based on the Apriori framework, they can be considered for

supporting other algorithms for handling uncertain data. The efficient frequent pattern mining algorithms based on the expected support counts of the patterns is used for uncertain databases. The use of expected support may render important patterns missing. Hence, they proposed to compute the probability that a pattern is frequent, and introduced the notion of PFI. Dynamic-programming based Solutions were developed to retrieve PFIs from attribute-uncertain databases

4.1 Standard statistical properties of s-pmf

An interesting observation about s(I) is that it is essentially the number of successful poisson trials [29]. To explain, we let XjI be a random variable, which is equal to one if I is a subset of the items associated with transaction tj, or zero otherwise. Notice that Pr(I C tj) can be easily calculated in our uncertainty models.

- For attribute-uncertainty,
  Pr(I tj) = II Pr(v Σ tj).
- For tupel-uncertainty,

$$Pr(I \subseteq t_j) = \begin{cases} Pr(t_j), & \text{if } I \subseteq t_j, \\ 0, & \text{otherwise.} \end{cases}$$

Given a database of size n, each I is associated with random variables X1, X2…….. Xn. In both uncertainty models considered in this paper, all tuples are independent. Therefore, these n variables are independent, and they represent n poisson Trials. Moreover, XI = ∑nj=1 XI follows a Poisson binomial distribution. Next we observe an important relationship between

XI and PrI(i)
PrI(i) = Pr (XI = i).

This is simply because XI is the number of items that I exists in the database. Hence The s-pmf of I, i.e., PrI(i) is the pmf of XI, a Poisson binomial distribution. Using the above formula we can rewrite the formula as which computes the Frequentness probability of I,as

Prfreq (I) = ∑ Pr ((XI = i).
=Pr(XI > msc(D)).

Therefore, if the cumulative distribution function(cdf) of XI is known, Prfreq (I) can also be evaluated. Next, we discuss an approach to approximate this cdf, in order to compute Prfreq(I) efficiently.

4.2 Proposed Approach: Approximating S-Pmf

We can express Prfreq (I) = 1 – Pr (XI < msc (D) – 1). For notational convenience, let PI be Pr (I C tj). Then the expected value of XI in denoted by μI , can be computed by μI = ∑ PjI. since a Poisson distribution can be well approximated by a Poisson distribution.

$$F(i,\mu) = \frac{\Gamma(i+1,\mu)}{i!} = \frac{\int_{\mu}^{\infty} t^{(i+1)-1}e^{-t}dt}{i!}$$

Since minsup is fixed and independent of μ, let us examine the partial derivative w.r.t μ

$$\begin{aligned} \frac{\partial F(i,\mu)}{\partial \mu} &= \frac{\partial}{\partial \mu}\left(\frac{\int_{\mu}^{\infty} t^{(i+1)-1}e^{-t}dt}{i!}\right) \\ &= \frac{1}{i!}\frac{\partial}{\partial \mu}\left(\int_{\mu}^{\infty} t^{i}e^{-t}dt\right) \\ &= \frac{1}{i!}(-\mu^{i}e^{-\mu}) \\ &= -f(i,\mu) \leq 0. \end{aligned}$$

Thus, the cdf of the poisson distribution F(I, μ) is monotonically decreasing w.r.t. μ, When i is fixed. Consequently, 1 – F (i – 1, μ) increases monotonically with μ.

4.3 PFI Testing

Given the values of minsup and minprob, we can test whether I is a threshold-based PFI, in three steps.

Step 1
Find a real number μm satisfying the equation:

Minprob = 1 – F (msc (D) – 1 , μm).

The above equation can be solved efficiently by employing numerical methods.

Step 2
Use the above formula compute μI. Notice that the database D has to be scanned once.

Step 3
If μI > μm we conclude that I is a PFI. Otherwise I must not be a PFI.

To understand why this works, first notice that the minprob and the freuentness probability is the same. Essentially, step 1 finds out the value of μm that corresponds to the frequentness probability threshold. Hence, these steps together can test whether an item set is a PFI. In order to verify whether I is a PFI, once μm is found, we do not have to evaluate Prfreq(I). Instead, we compute μI in step 2, which can be done in O (n) time. This is a more scalable method compared with solutions which evaluate Prfreq(I) in O (n 2) time. Next, we study how this method can be further improved.

5. Experimental Results

5.1 Simulation Model and Methods

Our experiments were carried out on the Windows XP operating system, on a machine with a 2.66 GHz Intel Core 2 Duo processor and 16 GB memory. The implementation can be preceded through JSP in J2EE but it will be considered as web communication .For proactive routing we need dynamic web. So java will be more suitable for platform independence and dynamic web concepts. For maintaining route information we go for MySQL Server as database back end.

We now present the experimental results on two data sets. The first one, called accidents, comes from the Frequent Item set Mining (FIMI) Data Set Repository.1 This data set is obtained from the National Institute of Statistics (NIS) for the region of Flanders (Belgium), for the period of 1991-2000. The data are obtained from the "Belgian Analysis Form for Traffic Accidents," which is filled out by a police officer for each traffic accident occurring on a public road in Belgium. The data set contains 3, 40,184 accident records, with a total of 572 attribute values. On average, each record has 45 attributes. We use the first 10k tuples as our default data set. The default value of minsup is 20 percent. To test the incremental mining algorithms, we use the first 10k tuples as the old database D, and the subsequent tuples as the delta database d. The default size of d is 5 percent of D. For both data sets, we consider both attribute and tuple uncertainty models. For attribute uncertainty, the existential probability of each attribute is drawn from a Gaussian distribution with mean 0.5 and standard deviation 0.125. This same distribution is also used to characterize the existential probability of each tuple, for the tuple uncertainty model. The default value of minprob is 0.4. In the results presented, minsup is shown as a percentage of the data set size n. Notice that when the values of minsup or minprob are large, no PFIs can be returned; we do not show the results for these values.

5.2 Results on Threshold-Based Pfi Mining

We now compare the performance of three PFI mining algorithms mentioned in this paper: 1) DP, the Apriori algorithm used in [5]; 2) MB, the modified Apriori algorithm that employs the PFI testing method and 3) MBP, the algorithm that uses the improved version of the PFI testing method.

Accuracy. Since MB approximates s-pmf by a Poisson distribution, we first examine its accuracy with respect to DP, which yields PFIs based on exact frequentness probabilities. Here, we use the standard recall and precision measures [7], which quantify the number of negatives and false positives. Specifically, let FDP be the set of PFIs generated by DP, and FMB be the set of PFIs produced by MB. The recall and the precision of MB, relative to DP, are defined as follows

$$recall = \frac{|F_{DP} \cap F_{MB}|}{|F_{DP}|} \quad (1)$$

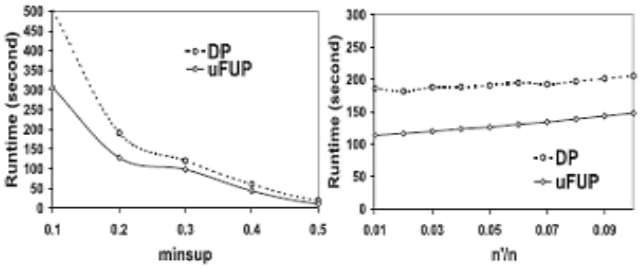
$$precision = \frac{|F_{DP} \cap F_{MB}|}{|F_{MB}|}$$

Table 1: Recall and Precision of MB

Minsup	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*
Recall	1	1	1	1	1
Precision	0.9997	1	1	1	1

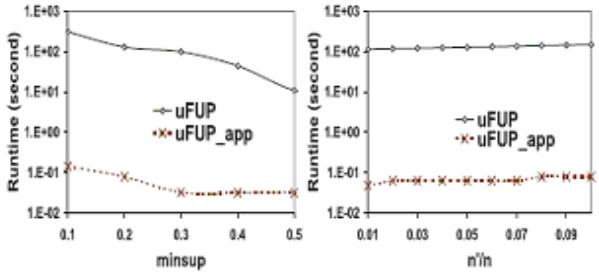
Table 2: Existential probability

Distribution	Mean	Standard Deviation
G_0	0.8	0.125
G_1(default)	0.5	0.125
G_2	0.5	0.25
G_3	0.5	$\sqrt{1/12} \approx 0.289$
G_4	0.5	0.5
G_5	0.5	1.0
Un	0.5	$\sqrt{1/12} \approx 0.289$



(a) Runtime vs. *minsup* (b) Runtime vs. n'/n

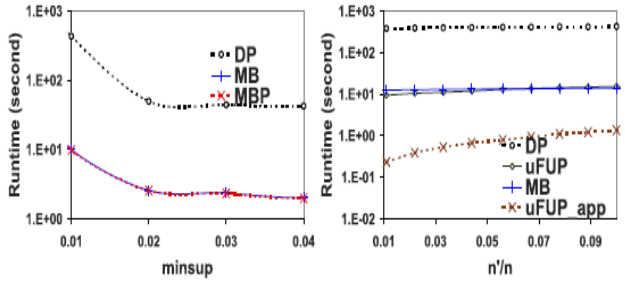
Figure 1: Efficiency of uFUP Versus DP



(a) Runtime vs. *minsup* (b) Runtime vs. n'/n

Figure 2: Efficiency of uFUPappVersus uFUP

Synthetic data set. Finally, we test our algorithms on a synthetic data set. Fig. 13a compares the performance of MB, MBP, and DP, for the attribute uncertainty model. We found that MB and MBP outperform DP



(a) Runtime vs. *minsup* (b) Runtime vs. n'/n

6. Conclusion

We propose a model-based approach to extract threshold-based PFIs from large uncertain databases. Its main idea is to approximate the s-pmf of a PFI by some common probability model, so that a PFI can be verified quickly. We also study two incremental mining algorithms for retrieving PFIs from evolving databases. Our experimental results show that these algorithms are highly efficient and accurate. They support both attribute- and tuple uncertain data. We will examine how to use the model based approach to

develop other mining algorithms (e.g., clustering and classification) on uncertain data. It is also interesting to study efficient mining algorithms for handling tuples updates and deletion. Another interesting work is to investigate PFI mining algorithms for probability models that capture correlation among attributes and tuples. This work gives rise to several interesting directions for future research. In particular, additional important item ranking criteria should be explored for potential diversity improvements. This may include consumer-oriented or manufacturer oriented ranking mechanisms, depending on the given application domain, as well as external factors, such as social networks

References

[1] N. Khoussainova, M. Balazinska, and D. Suciu, "Towards Correcting Input Data Errors Probabilistically Using Integrity Constraints," Proc. Fifth ACM Int'l Workshop Data Eng. for Wireless and Mobile Access (MobiDE), 2006.
[2] P. Sistla, O. Wolfson, S. Chamberlain, and S. Dao, "Querying the Uncertain Position of Moving Objects," Temporal Databases: Research and Practice, Springer Verlag, 1998.
[3] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong, "Model-Driven Data Acquisition in Sensor Networks," Proc. 13th Int'l Conf. Very Large Data Bases (VLDB), 2004.
[4] C. Aggarwal and P. Yu, "A Survey of Uncertain Data Algorithms and Applications," IEEE Trans Knowledge and Data Eng., vol. 21, no. 5, pp. 609-623, May 2009.
[5] T. Bernecker, H. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic Frequent Itemset Mining in Uncertain Databases," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2009.
[6] N. Dalvi and D. Suciu, "Efficient Query Evaluation on Probabilistic Databases," Proc. 13th Int'l Conf. Very Large Data Bases (VLDB), 2004
[7] T. Jayram et al., "Avatar Information Extraction System," IEEE Data Eng. Bull., vol. 29, no. 1, pp. 40-48, Mar. 2006.
[8] R. Cheng, D. Kalashnikov, and S. Prabhakar, "Evaluating Probabilistic Queries over Imprecise Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2003.
[9] N. Dalvi and D. Suciu, "Efficient Query Evaluation on Probabilistic Databases," Proc. 13th Int'l Conf. Very Large Data Bases (VLDB), 2004.
[10] J. Huang, "MayBMS: A Probabilistic Database Management System," Proc. 35th ACM SIGMOD Int'l Conf. Management of Data, 2009.
[11] R. Jampani, L. Perez, M. Wu, F. Xu, C. Jermaine, and P. Haas, "MCDB: A Monte Carlo Approach to Managing Uncertain Data," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2008
[12] M. Mutsuzaki, "Trio-One: Layering Uncertainty and Lineage on a Conventional DBMS," Proc. Third Biennial Conf. Innovative Data Systems Research (CIDR), 2007.
[13] R. Agrawal, T. Imieli_nski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data, 1993.
[14] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2000.
[15] C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent Pattern Mining with Uncertain Data," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2009
[16] C.K. Chui, B. Kao, and E. Hung, "Mining Frequent Itemsets from Uncertain Data," Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD), 2007.
[17] L. Sun, R. Cheng, D.W. Cheung, and J. Cheng, "Mining Uncertain Data with Probabilistic Guarantees," Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2010
[18] C.J. van Rijsbergen, Information Retrieval. Butterworth, 1979.
[19] C.J. van Rijsbergen, Information Retrieval. Butterworth, 1979.

Author Profile

R. Manimegalai received the M.Sc IT degree in Annamalai University in 2009. Working as Assistant Professor in Computer Science department for 3 years. Area of Interest Data Mining and Java.