Skip Search Approach for Mining Probabilistic Frequent Itemsets from Uncertain Data

Takahiko Shintani, Tadashi Ohmori and Hideyuki Fujita
Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan

Keywords: Frequent Itemset, Probabilistic Data, Uncertain Data.

Abstract: Due to wider applications of data mining, data uncertainty came to be considered. In this paper, we study mining probabilistic frequent itemsets from uncertain data under the Possible World Semantics. For each tuple has existential probability in probabilistic data, the support of an itemset is a probability mass function (pmf). In this paper, we propose skip search approach to reduce evaluating support pmf for redundant itemsets. Our skip search approach starts evaluating support pmf from the average length of candidate itemsets. When an evaluated itemset is not probabilistic frequent, all its superset of itemsets are deleted from candidate itemsets and its subset of itemset is selected as a candidate itemset to evaluate next. When an evaluated itemset is probabilistic frequent, its superset of itemset is selected as a candidate itemset to evaluate next. Furthermore, our approach evaluates the support pmf by difference calculus using evaluated itemsets. Thus, our approach can reduce the number of candidate itemsets to evaluate their support pmf and the cost of evaluating support pmf. Finally, we show the effectiveness of our approach through experiments.

1 INTRODUCTION

One of important problems in data mining is a discovery of frequent itemsets within a large database(Agrawal and R.Srikant, 1994; Han et al., 2000). Due to wide applications in frequent itemset mining, data uncertainty came to be considered(Aggarwal and Yu, 2009). For example, data collected by sensor device are noisy. The locations of users obtained through GPS systems are not precise. The user activities that were estimated from the aceleraion sensor data are underspecified. The uncertain data is the probabilistic database which each item and/or transaction has a probability value, called attribute uncertainty model and tuple uncertainty model respectively. On the probabilistic database, frequent itemsets are probabilistic. The support of an itemset is a random variable. Several algorithms for mining frequent itemsets from uncertain data have been proposed. In (Chuim et al., 2007; Leung et al., 2007; Leung et al., 2008; Aggarwal et al., 2009; Wang et al., 2013; MacKinnon et al., 2014; Cuzzocrea et al., 2015), the frequent itemsets are detected by their expected support count. However, it is reported that many important itemsets are missed by using expected support(Zhang et al., 2008).

By using the Possible Worlds Semantics (PWS), we can interpret probabilistic databases(Dalvi and Suciu, 2004). A possible world means the case where a set of transactions occurs. We can find frequent itemsets under PWS by counting their support counts from every possible world. Since the enormous number of possible worlds have to be considered, this is impractical. The approximate algorithms for finding frequent itemsets from the attribute uncertain model and tuple uncertain model were proposed(Wang et al., 2012; Leung and Tanbeer, 2013), but these algorithms cannot find exact solutions. In (Zhang et al., 2008), the algorithm for finding exact solutions of frequent items were proposed, but this algorithm cannot handle itemsets. The algorithms for finding exact solution of probabilistic frequent itemsets were proposed in (Bernecker et al., 2009; Sun et al., 2010). The bottom-up manner algorithm finds frequent itemsets in ascending order of length like Apriori algorithm(Agrawal and R.Srikant, 1994). This algorithm can prune candidate itemsets by Apriori down-closed property. The dynamic programing (DP)(Bernecker et al., 2009) and divide-and-conquer (DC)(Sun et al., 2010) for evaluating the support probability were proposed. DP and DC evaluate the support probability from scratch, so its cost is high. In (Sun et al., 2010), the top-down manner algorithm, TODIS, was proposed. TODIS can evaluate the support probabil-

Shintani, T., Ohmori, T. and Fujita, H.
Skip Search Approach for Mining Probabilistic Frequent Itemsets from Uncertain Data.
DOI: 10.5220/0006035401740180
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR*, pages 174-180
ISBN: 978-989-758-203-5

Table 1: Example of PDB.

TID	Transaction	Existential Prob.
T_1	$\{a,e,f,g\}$	0.7
T_2	$\{a,b,e,g\}$	1.0
T_3	$\{b,c,h,i,j\}$	0.5
T_4	$\{b,d,f,h,j\}$	0.8

ity function by inheriting a superset of the itemset, but it involves evaluating many redundant itemsets.

In this paper, we study mining probabilistic frequent itemsets from uncertain data in the tuple uncertainty model under PWS by extending our previous study(Tateshima et al., 2015). We propose skip search approach to avoid evaluating redundant itemsets which becomes probabilistic infrequent. Our skip search approach starts to evaluate the support probability from the average length of itemsets. When the evaluated itemset is not probabilistic frequent, it's super-itemset is evaluated next. When the evaluated itemset is probabilistic frequent, its sub-itemset is evaluated next. Moreover, our skip search approach evaluates the support probability function efficiently by difference calculus. We show the effectiveness of our skip search approach by experiments.

This paper is organized as follows. In the next section, we explain the problem of mining probabilistic frequent itemsets. In section 3, we propose our skip search approach. Performance evaluations are given in section 4. Section 5 concludes this paper.

2 MINING PROBABILISTIC FREQUENT ITEMSETS

First, we introduce basic concepts of frequent itemsets on exact databases. Let $\mathcal{L} = \{i_1, i_2, \ldots, i_m\}$ be a set of literals, called items. Let $\mathcal{D} = \{t_1, t_2, \ldots, t_n\}$ be a set of transactions, where each transaction t is a set of items such that $t \subseteq I$. A transaction has an associated unique identifier called TID. A set of items $X \subseteq I$ is called an itemset. Itemset X is a sub-itemset of itemset Y if and only if X is a subset of Y. Y is called a super-itemset of X. We say each transaction t *contains* an itemset X, if $X \subseteq t$. The itemset X has *support* s in $\mathcal{D}$ if s transactions contain X, here we denote $s = sup(X)$.

In the tuple uncertainty model, each transaction t_i has an existential probability p_i. Here $0 < p_i \leq 1$. Existential probability indicates the probability of the transaction occurs. Example of probabilistic database (PDB) in tuple uncertainty model is shown in Table 1. Table 1 consists of ten items with four transactions. For example, T_1 denotes that the probability of a transaction $\{a,e,f,g\}$ occurs is 0.7.

Table 2: Possible worlds for Table 1.

$\mathcal{W}$	Occurred transactions	Prob. $P(W)$
W_1	T_2	0.03
W_2	T_1, T_2	0.07
W_3	T_2, T_3	0.03
W_4	T_2, T_4	0.12
W_5	T_1, T_2, T_3	0.07
W_6	T_1, T_2, T_4	0.28
W_7	T_2, T_3, T_4	0.12
W_8	T_1, T_2, T_3, T_4	0.28

We can interpret PDB by using PWS. Table 2 shows all possible worlds for PDB in Table 1. Here, the probability of a possible world W_i is denoted as $P(W_i)$, the sum of them is $1 (= \sum_i P(W_i))$. For example, W_2 denotes the case where T_1 and T_2 occur, T_3 and T_4 do not occur. The probability of W_2 is calculated as follows: $P(W_2) = p_{T_1} * p_{T_2} * (1 - p_{T_3}) * (1 - p_{T_4}) = 0.7 * 1.0 * (1 - 0.5) * (1 - 0.8) = 0.07$.

Since there are many possible worlds and each possible world has a probability, the support of an itemset becomes a random variable. We denote $f_X(k)$ as the probability mass function(pmf) of an itemset X at $sup(X) = k (k \geq 0)$. For example, itemset $\{b,h\}$ is contained in T_3 and T_4. When both T_3 and T_4 occur, $sup(\{b,h\})$ becomes 2. The possible worlds where both T_3 and T_4 occur are W_7 and W_8, $f_{\{b,h\}}(2) = P(W_7) + P(W_8) = 0.4$. When either T_3 or T_4 occurs, $sup(\{b,h\})$ becomes 1. The possible worlds where either T_3 or T_4 occurs are W_3,W_4,W_5 and W_6, $f_{\{b,h\}}(1) = P(W_3) + P(W_4) + P(W_5) + P(W_6) = 0.5$. When neither T_3 nor T_4 occurs, $sup(\{b,h\})$ becomes 0. The possible worlds where neither T_3 nor T_4 occurs are W_1 and W_2, $f_{\{b,h\}}(0) = P(W_1) + P(W_2) = 0.1$.

In PDB, we define that an itemset X is probabilistic frequent if the following equation is satisfied.

$$Pr(sup(X) \geq minsup) \geq minprob \qquad (1)$$

Here, $Pr(sup(X) \geq minsup)$ is the sum of the probability that $sup(X)$ is $minsup$ or more. $minsup$ and $minprob$ are user-specified minimum thresholds of the support and the probability. For example, $Pr(sup(\{b,h\}) = f_{\{b,h\}}(2) + f_{\{b,h\}}(1) = 0.9$ when $minsup = 1$. If $minprob = 0.7$, $\{b,h\}$ is probability frequent.

The problem of mining probabilistic frequent itemsets is to find all itemsets that satisfy equation 1 on the assumption that we are given $minsup$, $minprob$ and PDB.

In (Sun et al., 2010), the bottom-up manner algorithm, a-Apriori, was proposed. The p-Apriori finds probabilistic frequent itemsets in ascending order of length like Apriori. In the first pass (pass

1), the support pmf(spmf) for each item are evaluated. All the items which satisfy equation 1 are picked out. These items are called probabilistic frequent 1-itemsets. Here after, k-itemset is defined a set of k items. The second pass, the 2-itemsets are generated using probabilistic frequent 1-itemsets, which are called candidate 2-itemsets. Then spmf for each candidate 2-itemsets are evaluated, the probabilistic frequent 2-itemsets which satisfy equation 1 are determined. In k-th pass, the candidate k-itemsets are generated by using probabilistic frequent $(k-1)$-itemsets, spmf for each candidates are evaluated, and the probabilistic frequent k-itemsets are determined. The candidate generation is same as Apriori. This iteration terminates when the probabilistic frequent itemset becomes empty. The dynamic programing(DP) algorithm(Bernecker et al., 2009) and divide-and-conquer(DC) algorithm(Sun et al., 2010) for evaluating spmf have been proposed. These algorithms do not examine possible worlds. By examining all transactions in PDB for each candidate itemset, DC evaluated spmf. The DC divides $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$, spmf of X is evaluated by the convolution of $f_{X,1}$ and $f_{X,2}$.

$$f_X(k) = \sum_{i=0}^{k} f_{X,1}(i) * f_{X,2}(k-i) \qquad (2)$$

Here, $f_{X,j}$ (here, $j = 1$ or 2) is the pmf of $sup(X)$ in $D_{X,j}$. The p-Apriori reduces the number of candidate itemsets by pruning with the Apriori downclosed property. However, all transactions in PDB have to be examined to evaluate spmf for each itemset.

Next, we explain the top-down manner algorithm, TODIS(Sun et al., 2010). The TODIS evaluates spmf of itemsets efficiently by inheriting spmf from their super-itemsets. Since TODIS starts spmf evaluations from the longest itemset, all potentially probabilistic frequent itemsets have to be identified on ahead. These itemsets are called candidate itemsets. It cannot be probabilistic frequent that an itemset does not satisfy *minsup* without considering existential probability. By ordinary frequent itemset mining algorithm such as Apriori, TODIS generates all candidate itemsets which satisfy minsup without considering existential probability. Hereafter, an itemset which satisfy *minsup* without considering existential probability is called a count frequent itemset. For each candidate itemset, TODIS also generates id-list, which is utilized to evaluate spmf. An id-list of itemset X is a set of TIDs which contains X. We denote id-list of itemset X as L^X. Then, TODIS evaluates spmf of every candidate itemsets in a top-down manner. First, the spmf of the longest candidate itemsets is evaluated by DC algorithm. Afterward, the spmf of candidate itemsets are evaluated in descending order of their length. The spmf of a candidate itemset is evaluated by inheriting spmf of its super-itemset. Here, we show evaluating the spmf of a candidate itemset X. The spmf of an itemset Z is denoted as f_Z. Assume that an itemset Y is a super-itemset of X, and f_Y is known. f_X is evaluated from transactions L^X. Since X is a sub-itemset of Y, $L^X \supseteq L^Y$. f_X can be evaluated from transactions in L^Y and $L^X \setminus L^Y$. Let $T_0, \ldots, T_{n-1}$ be a set of transactions in $L^X \setminus L^Y$. f_X can be evaluated by convolving $T_0, \ldots, T_{n-1}$ to f_Y. Here, let p_i be probability that T_i occurs, $q_i(= 1-p_i)$ be probability that T_i does not occur. Let $f_Z(k)$ be f_Z at $sup(Z) = k$. Let $f_Z^i(k)$ be spmf that convolved f_Z from T_0 to T_i. If any transactions does not occur, $k = 0$. $f_X(0)$ is calculated as

$$f_X(0) = f_Y^{n-1}(0) = f_Y(0) * \Pi_{m=0}^{n-1} q_m \qquad (3)$$

If $k \geq 1$, $f_Y^{i+1}(k)$ is the sum of the case that T_{i+1} occurs at $f_Y^i(k-1)$ and T_{i+1} does not occur at $f_Y^i(k)$. $f_Y^{i+1}(k)$ is calculated as

$$f_Y^{i+1}(k) = f_Y^i(k-1) * p_{i+1} + f_Y^i(k) * q_{i+1} \qquad (4)$$

$f_X(k)$ is the spmf of $sup(X) = k$ that convolved $f_Y(k)$ from T_0 to T_{n-1}, so we can calculate $f_X(k)$ by repeating Equation 4 from $i = 0$ to $n-1$. Thus, f_x is evaluated by inheriting f_Y using Equation 3 and 4. The top-down manner algorithm efficiently evaluates spmf of an itemset by inheriting the spmf of its super-itemsets. However, the spmf of all candidate itemsets has to be evaluated. Even if the evaluated candidate itemset is not probabilistic frequent, this algorithm cannot avoid to evaluate of spmf for its super-itemsets. Because all super-itemsets have already been evaluated.

3 SKIP SEARCH APPROACH

In this section, we describe the way to evaluating spmf by inheriting the spmf of sub-itemsets. Then, we explain the skip search approach for avoiding to evaluate spmf for redundant candidate itemsets.

3.1 Evaluating Spmf by Inheriting Sub-itemset

TODIS evaluates spmf of candidate itemsets in descending order, so all candidate itemsets can be evaluated spmf by inheriting spmf of its super-itemset. When the order to evaluate spmf is not one way, it is not enough. We can evaluate spmf of a candidate itemset by deconvolving spmf of its sub-itemset. Here, we propose the way to evaluate spmf by inheriting spmf of sub-itemset. Assume that an itemset Y

is a sub-itemset of a candidate itemset X, and f_Y is known. f_X is evaluated from transactions in L^X. Since X is a super-itemset of Y, $L^X \subseteq L^Y$. Let $T_0, \dots, T_{n-1}$ be a set of transactions in $L^Y \setminus L^X$. The spmf of X can be evaluated by deconvolving $T_0, \dots, T_{n-1}$ to the spmf of Y. Here, let p_i be the probability that T_i occurs, $q_i (= 1 - p_i)$ be the probability that T_i does not occur. Let $f_X(k)$ be the spmf of X at $sup(X) = k$. Let $f_Y^i(k)$ be the spmf of $sup(Y) = k$ that deconvolved f_Y from T_0 to T_i. If any transactions does not occur, $k = 0$. $f_Y(0)$ is calculated as

$$f_Y(0) = f_Y^0(0) * q_0 \tag{5}$$

$f_Y^0(0)$ is calculated as

$$f_Y^0(0) = \frac{f_Y(0)}{q_0} \tag{6}$$

We can calculate $f_Y^{i+1}(0)$ by deconvolving $f_Y^i(0)$ with T_{i+1}

$$f_Y^{i+1}(0) = \frac{f_Y^i(0)}{q_{i+1}} \tag{7}$$

$f_X(0)$ can be calculated by deconvolving $f_Y(0)$ from $T_0$to T_{n-1}.

$$f_X(0) = f_Y^{n-1}(0) = f_Y(0) * \Pi_{m=0}^{n-1} \frac{1}{q_m} \tag{8}$$

If $k \geq 1$, k transactions occur. When f_Y is deconvolved with T_0, $f_Y(k)$ is calculated as

$$f_Y(k) = f_Y^0(k) * q_0 + f_Y^0(k-1) * p_0 \tag{9}$$

Thus,

$$f_Y^0(k) = \frac{f_Y(k) - f_Y^0(k-1) * p_0}{q_0} \tag{10}$$

When f_Y is deconvolved with $T_0, \dots, T_i$, $f_Y^i(k)$ is calculated as

$$f_Y^i(k) = f_Y^{i+1}(k) * q_{i+1} + f_Y^{i+1}(k-1) * p_{i+1} \tag{11}$$

Thus,

$$f_Y^{i+1}(k) = \frac{f_Y^i(k) - f_Y^{i+1}(k-1) * p_{i+1}}{q_{i+1}} \tag{12}$$

$f_X(k)$ can be calculated by deconvolving $f_Y(k)$ from T_0 to T_{n-1}, so we can calculate $f_X(k)$ by repeating Equation 8 to 12 from $i = 0$ to $n - 1$.

This difference calculus by inheriting sub-itemset can be applied to p-Apriori. When pass $k \geq 2$, spmf of candidate itemsets can be evaluated by inheriting their sub-itemsets. Since p-Apriori finds probabilistic frequent itemsets in ascending order of the length of itemsets, all sub-itemsets have already been evaluated. We can evaluate spmf of a candidate k-itemset by inheriting its sub-itemsets.

3.2 Order to Evaluate Spmf of Candidate Itemsets

The skip search approach evaluates spmf in a bidirectional way so that we can avoid to evaluate spmf of probabilistic infrequent itemsets by Apriori downclosed property. Assume that the spmf of candidate itemset X was evaluated. If X is not probabilistic frequent, we can omit to evaluate spmf of all super-itemset of X. Because an itemset whose sub-itemsets are not probabilistic frequent cannot be probabilistic frequent. Since our skip search approach starts evaluating spmf from the average length of candidate itemsets, it's super-itemsets that have not been evaluated spmf are remaining. When X is probabilistic infrequent, a sub-itemset of X in candidate itemsets is selected as a candidate itemset to evaluate spmf next. The spmf of this itemset is evaluated by difference calculus in section 3.1. When X is probabilistic frequent, spmf of all sub-itemsets of X in candidate itemsets are evaluated. A super-itemset of X in candidate itemsets is selected as the candidate itemset to evaluate spmf next, since it has a potential to be probabilistic frequent. The spmf of this itemset is evaluated same as TODIS. Hereby, our skip search approach can omit evaluations for redundant candidate itemsets.

Here, we show the procedure to select a candidate itemset evaluating spmf next.

```
An itemset $X$ in candidate itemsets is selected at random. Here, the length of $X$ is close to the average length of candidate itemsets. Then, the spmf of $X$ is evaluated.
  If $X$ is probabilistic frequent:
    An itemset $Y$ which is super-itemset of $X$ in candidate itemsets is selected. The length of $Y$ is close to the median length of $X$ and the longest super-itemset of $X$.
  If $X$ is not probabilistic frequent:
    An itemset $Y'$ which is sub-itemset of $X$ in candidate itemsets is selected. The length of $Y'$ is close to the median length of $X$ and the shortest unevaluated sub-itemset of $X$.
  If all sub-itemset and super-itemset have been already evaluated:
    An itemset is selected at random.
```

3.3 Procedure of Skip Search Approach

Here, we describe the procedure of skip search approach.

Step1. Set of candidate itemsets $\mathcal{C}$, that of evaluating itemsets $\mathcal{E}$ and that of probabilistic frequent itemsets $\mathcal{F}$ are initialized to the empty set.

Step2. All count frequent itemsets are inserted to $\mathcal{C}$, and their id-lists are generated.

Step3. A candidate itemset $c \in C$ is selected from candidate itemset set at random.

Step4. f_c is evaluated from L^c. If a super-itemsets or sub-itemsets of c exists in $\mathcal{E}$, f_c is evaluated by inheriting it. Then, c is deleted from $\mathcal{C}$.

- If c is probabilistic frequent, c is inserted into $\mathcal{F}$. When any sub-itemset c' of c has not been evaluated yet, $f_{c'}$ is evaluated, inserted into $\mathcal{F}$ and deleted from $\mathcal{C}$. A super-itemset of c in $\mathcal{C}$ is selected as the candidate itemsets evaluating spmf next.
- If c is not probabilistic frequent, all super-itemset of c are deleted from $\mathcal{C}$. A sub-itemset of c in $\mathcal{C}$ is selected as the candidate itemsets evaluating spmf next.
- If super-itemset and sub-itemset do not exist in $\mathcal{C}$, a candidate itemset in $\mathcal{C}$ is selected from candidate itemset set at random.

Step5. If the candidate itemset evaluating spmf next is empty, this procedure terminates. Otherwise, step4 is repeated.

In skip search approach, candidate itemsets which are selected at random cannot be evaluated by inheriting their super/sub-itemsets. The spmf of these itemsets are evaluated by algorithm DP or DC, it is costly. To solve this problem, we propose another approach, "skip search approach from maximal". A maximal candidate itemsets(Bayardo, 1998) is selected in Step 3. In Step 4, candidate itemsets which are sub-itemset of selected maximal candidate itemset are evaluated. This procedure can avoid using DC in Step 4. The spmf of maximal candidate itemsets have to be evaluated by DC, and most of these itemsets are not probabilistic frequent. However, the cost of evaluating them by DP or DC is relatively small. Because the support of these itemset is small. Example of the order to evaluate spmf in skip search approach from maximal is shown in Figure 1. First, a maximal candidate itemset $\{a,b,c,d,e\}$ is selected and evaluated. Then a candidate itemset $\{a,c,d\}$ which is a sub-itemset of $\{a,b,c,d,e\}$ is selected and evaluated. Since $\{a,c,d\}$ is not probabilistic frequent, $\{a,b,c,d\}$ and $\{a,c,d,e\}$ which are super-itemsets of $\{a,c,d\}$ are deleted. We can omit to evaluate spmf of $\{a,b,c,d\}$ and $\{a,c,d,e\}$. Next, $\{a,d\}$ which is a sub-itemset of $\{a,c,d\}$ is selected and evaluated. $\{a,d\}$ is probabilistic frequent, so spmf of all its sub-itemsets, $\{a\}$ and $\{d\}$, are evaluated. Then, $\{a,d,e\}$ which is a super-itemset of $\{a,d\}$ is selected as a candidate itemset to evaluate next. When all sub-itemsets of $\{a,b,c,d,e\}$ were evaluated, other maximal candidate itemset, for example $\{d,e,f\}$, is selected as a candidate itemset to evaluate next.

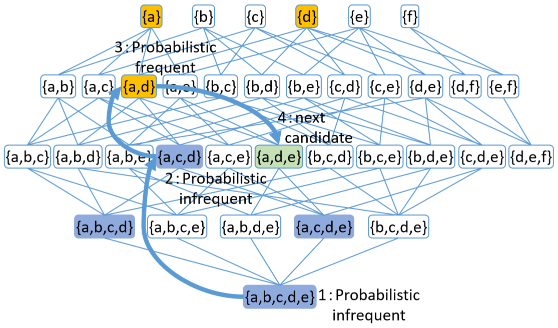

Figure 1: Example of the order to evaluate spmf in skip search approach from maximal.

This enhance is effective when all candidate itemsets and their id-lists cannot fit in memory. Since the information of a maximal candidate itemset and its sub-itemsets is required in Step 4, we can reduce the size of memory usage.

4 PERFORMANCE EVALUATION

We evaluated the performance of skip search approaches by comparing with the top-down manner algorithm, TODIS, and the bottom-up manner algorithm, p-Apriori. In experiments, p-Apriori evaluates spmf by ihneriting sub-itemsets described in section 3.1, that is more efficient than using DC. This algorithm is denoted as "p-Apriori w diffcalc" in experimantal results. In skip search approaches and TODIS, the count frequent itemsets are found by Apriori algorithm. In experimental results, the naive skip search approach is denoted as "skip". The skip search approach described in section 3.3 is denoted as "skip f max".

To evaluate the performance of our approach, synthetic data emulating retail transactions are used, where the generation procedure is based on the method described in (Agrawal and R.Srikant, 1994). The average length of a transaction is 40, the average length of a frequent itemset is 10, and the dataset size N is $500k$. For each transaction, we set the existential probability with a Gaussian distribution.

Figure 2 shows the execution time varying *minsup*. Here, *minprob* was set to 0.3. When *minsup* is small, the difference between the execution time of skip search approaches and TODIS. Since the average length of probabilistic frequent itemsets becomes long for small *minsup*, the number of candidate itemsets which skip search approaches can omit to evalu-

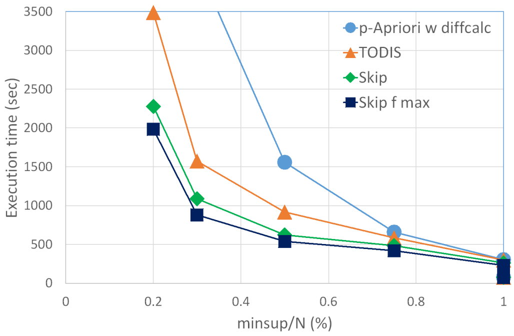

Figure 2: Execution time varying *minsup*.

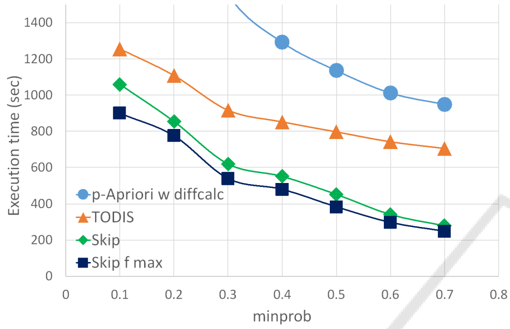

Figure 3: Execution time varying *minprob*.

ate spmf increases.

Figure 3 shows the execution time varying *minprob*. Here, *minsup* was set to 0.003. Our skip search approaches significantly outperform other algorithms. As the minimum probability decreases, the difference of the execution time between skip search approaches and TODIS shrinks. In this experiment, *minsup* is fixed, so the number of candidate itemsets is constant (Figure 4). When *minprob* is small, the ratio of the number of omitted candidate itemsets becomes small. Thus, the difference of the execution time becomes small. The difference of the execution time between skip and skip f max becomes small for large *minprob*. The number of probabilistic frequent itemsets becomes smaller as *minprob* increases. Since the number of probabilistic infrequent itemsets is large, the difference between skip and skip f max becomes small.

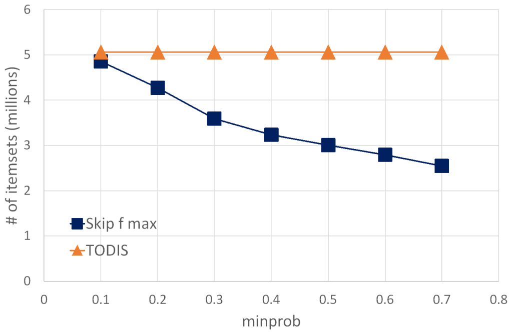

Figure 4: Number of evaluated itemsets varying *minprob*.

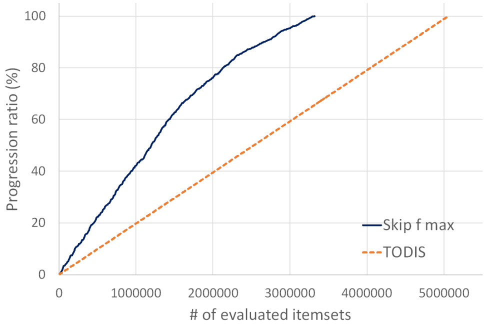

Figure 5: Progression ratio.

Figure 5 shows the progression ratio of skip f max and TODIS. The horizontal axis is the number of itemsets which have been evaluated their spmf, the vertical axis is the progression ratio. The progression ratio means the ratio of itemsets either probabilistic frequent or not have been confirmed. Here, the minimum support threshold and the minimum probability threshold are set to 0.003 and 0.3 respectively. TODIS evaluates the spmf of candidate itemsets one by one in descending order, so its progression ratio becomes linear. In skip f max, its progression ratio becomes linear when an evaluated candidate itemset is probabilistic frequent. However, when an evaluated candidate itemset is not probabilistic frequent, its super-itemsets are deleted at a time, so the gradient of progression ratio significantly increases.

5 CONCLUSIONS

In this paper, we proposed skip search approach for mining probabilistic frequent itemsets under the Possible Worlds Semantics. By starting spmf evaluation from the average length of candidate itemsets for each maximal itemset and its sub-itemsets, our skip search approach can omit to evaluate redundant itemsets which become probabilistic infrequent. It can evaluate spmf efficiently by inheriting spmf from its sub/super-itemsets. Performance evaluations show our skip search approach from maximal can attain good performance.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP25280022 and JP26330129.

This work was supported by CREST, Japan Science and Technology Agency.

REFERENCES

Aggarwal, C., Li, Y., and Wang, J. (2009). Frequent pattern mining with uncertain data. In *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Aggarwal, C. and Yu, P. (2009). A survey of uncertain data algorithms and applications. In *IEEE Transactions on Knowledge and Data Enginerring*.

Agrawal, R. and R.Srikant (1994). Fast algorithm for mining association rules. In *20th International Conference on Very Large Data Bases*.

Bayardo, R. (1998). Efficiently mining long patterns from databases. In *1998 ACM SIGMOD International Conference on Management of Data*.

Bernecker, T., Kriegel, H., Renz, M., Verhein, F., and Zuefle, A. (2009). Probabilistic frequent itemset mining in uncertain databases. In *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chuim, C., Kao, B., and Hung, E. (2007). Mining frequent itemsets from uncertain data. In *11th Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Cuzzocrea, A., Leung, C., and MacKinnon, R. (2015). Approcimation to expected support of frequent itemsets in mining probabilistic sets of uncertain data. In *19th Annual Conference in Knowledge-Based and Intelligent Information and Engineering Systems*.

Dalvi, N. and Suciu, D. (2004). Efficient query evaluation on probabilistic databases. In *13th International Conference on Very Large Data Bases*.

Han, J., Pei, J., and Y.Yin (2000). Mining frequent patterns without candidate generation. In *2000 ACM SIGMOD International Conference on Management of Data*.

Leung, C., Carmichael, C., and Hao, B. (2007). Efficient mining of frequent patterns from uncertain data. In *Workshops of 7th IEEE International Conference on Data Mining*.

Leung, C., Mateo, M., and Brajczuk, D. (2008). A tree-based approach for frequent pattern mining from uncertain data. In *12th Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Leung, C. and Tanbeer, S. (2013). Puf-tree: a compact tree structure for frequent pattern mining of uncertain data. In *17th Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

MacKinnon, R., Strauss, T., and Leung, C. (2014). Disc: efficient uncertain frequent pattern mining with tightened upper bound. In *Workshops of 14th IEEE International Conference on Data Mining*.

Sun, L., Cheng, R., Cheung, D., and Cheng, J. (2010). Mining uncertain data with probabilistic gurantees. In *16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Tateshima, H., Shintani, T., Ohmori, T., and Fujita, H. (2015). Skip search approach for mining frequent itemsets from uncerdain dataset. In *DBSJ Japanese Journal*.

Wang, L., Cheung, D., and Cheung, R. (2012). Efficient mining of frequent item sets on large uncertain databases. In *IEEE Transactions on Data Engineering*.

Wang, L., Feng, L., and Wu, M. (2013). At-mine: an efficient algorithm of frequent itemset mininf on uncertain dataset. In *Journal of Computers*.

Zhang, Q., Li, F., and Yi, K. (2008). Finding frequent items in probabilistic data. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.