一般論文

DBSJ Japanese Journal
Vol. 14-J, Article No. 6, March 2016

スキップ探索を用いた不確実データからの頻出パターンの抽出

Skip search approach for mining frequent patterns from uncertain database

建島 広翔♡　新谷 隆彦◇
大森 匡♠　藤田 秀之♣

Hiroto TATESHIMA　Takahiko SHINTANI
Tadashi OHMORI　Hideyuki FUJITA

不確実データに対する頻出パターン抽出の研究が進められている．不確実データはトランザクションが存在確率を持っているため，パターンの頻度は確率変数となる．それぞれのトランザクションが起こる場合を考慮して頻度の確率を計算しなければならないため，頻度計算の処理コストが高い．従来手法では，探索候補を存在確率が考慮しない場合に最小頻度以上となるパターンに限定していたが，存在確率を考慮した場合に頻出の条件を満たさないパターンの頻度計算も行っていた．本研究では，存在確率を考慮した場合に頻出とならないパターンの頻度計算を回避することによって頻度計算の処理コストを低減する手法を検討し，実験により有効性を評価した．

Mining frequent itemsets from uncertain database has attracted much attention. In uncertain database, each transaction have existential probability values, and support of an itemset is modeled by possibility mass function(pmf). Because an uncertain database has an exponential number of possible worlds, calculating support pmf is costly. In this paper, we propose our approach to avoid calculating support pmf for probabilistic infrequent itemsets and show its effectiveness by experiment.

♡ 非会員　電気通信大学大学院情報システム学研究科
tateshima@hol.is.uec.ac.jp
◇ 正会員　電気通信大学大学院情報システム学研究科
shintani@is.uec.ac.jp
♠ 正会員　電気通信大学大学院情報システム学研究科
omori@is.uec.ac.jp
♣ 正会員　電気通信大学大学院情報システム学研究科
fujita@is.uec.ac.jp

1. はじめに

通信技術を含むハードウェアの急速な進歩により，データ量が爆発的に増大した．これに伴い，大量のデータの中に潜む有用な情報を抽出するデータマイニング技術が研究されている．その一つとして，不確実データに対する頻出パターン抽出がある．不確実データでは各トランザクションが存在確率を持っている．不確実データにおける頻出パターン抽出は，それぞれのトランザクションが起こる確率を考慮しなければならないため，頻度計算の処理コストが膨大になる．これまでに不確実データから頻出パターンを抽出する手法の研究が進められてきた．Apriori[1] のような候補作成型の U-Apriori[2]，P-Apriori，TODIS[3]，MBP[4]，FP-Growth[5] のようなパターン成長型の UF-Growth[6, 7]，UH-Miner[8]，AT-Mine[9] などが研究されてきた．候補作成型の手法は Apriori の処理対象データを不確実データに拡張し，パス毎に探索候補を作成し，頻度の数え上げと同時に存在確率を考慮して頻度計算を行う手法である．Ariori と同様に大量の探索候補を作成し，データベースのスキャンを繰り返す必要がある．MBP はポアソン分布を用いた概算値を計算することで高速化とメモリ効率向上を実現しているが，近似解しか求めることができない．また，パターン成長型は FP-Growth と同様にデータを木構造に変換することでデータベーススキャンの繰り返しと探索候補作成を回避する手法である．しかし，存在確率を考慮することで同じアイテムでもノードを共有させることができないなど木構造が大きくなってしまう問題点がある．

本研究では，存在確率を考慮した場合に頻出とならないパターンの頻度計算を回避することによって頻度計算の処理コストを低減する候補作成型の手法を検討する．TODIS は存在確率を考慮しない場合に頻度の条件を満たすパターンに対して，長いパターンから順に差分計算によって頻度計算をすることで処理コストを抑えることができる．しかし，TODIS は長い候補から順に探索するため，存在確率を考慮した場合には頻出とならないパターンの頻度計算も行っていたが，これらパターンの頻度計算は不要である．提案手法では，TODIS と同様に探索候補とするパターンを限定し，中程度の長さのパターンから順に頻度計算を行うスキップ探索のアプローチを取る．頻度計算を行ったパターンが頻出でない場合，そのパターンを含むパターンは頻出でないため，これらパターンの頻度計算を省略できる．提案手法を実装し，実験によって有効性を評価する．

2. 定義

2.1 不確実データ

アイテム全体の集合を I とする．アイテム $i \in I$ の集合をトランザクション T という．トランザクションに存在確率を付与したデータを不確実データと呼び，不確実データの集合を不確実データベースとする．存在確率はトランザクションが起きる確率 p であり，その取り得る範囲は $0 \leq p \leq 1$ である．表 1 に不確実データベースの例を示す．この不確実データベースはアイテム a,b,c,d,e,f,g,h,i,j から成る 4 つのトランザクションを持つ．例えば，T1 は {a,e,f,g} が起こる確率が 0.7 であることを示す．不確実データではトランザクションに存在確率があるため，どのトランザクションが起きるかによって場合分けできる．表 1 の不確実データベースのそれぞれのトランザクションが起こりうる場合を表 2 に示す．W1 は T2 のみが起きる場合を示しており，その場合が起きる確率は T2 の起きる確率 1.0 と T1，T3，T4 の起きない確率 0.3，0.5，0.2 の積から 0.03 と算出される．W2 は T1 と T2 のみが起きる場合を示しており，その場合が起きる確率は T1，T2 が起きる確率と T3 と T4 が起きない確率の積から 0.07 と算出される．ここで，W1 から W8 までの確率の和は 1 となる．

表 1: 不確実データの例

ID	トランザクション	存在確率
T1	a,e,f,g	0.7
T2	a,b,e,g	1.0
T3	b,c,h,i,j	0.5
T4	b,d,f,h,j	0.8

表 2: トランザクションの場合分け

場合の ID（W）	起きたトランザクションの場合	確率
W1	T2	0.03
W2	T1,T2	0.07
W3	T2,T3	0.03
W4	T2,T4	0.12
W5	T1,T2,T3	0.07
W6	T1,T2,T4	0.28
W7	T2,T3,T4	0.12
W8	T1,T2,T3,T4	0.28

2.2 頻出パターン

1 個以上のアイテムの組合せをパターンと呼ぶ．パターンは頻度の評価値を持つ．頻度はパターンを含むトランザクションの数であるが，トランザクションが存在確率を持つため，パターンの頻度は確率変数となる．パターン X の頻度の確率質量関数を頻度確率質量変数 (Support Probability Mass Funciton,SPMF) と呼ぶ．例えば パターン $\{b,h\}$ はトランザクション T3 と T4 に含まれるが，T3 と T4 が共に起きる場合に頻度が 2，T3 または T4 のどちらか一方が起きた場合は頻度が 1，T3 と T4 がどちらも起きない場合は頻度が 0 となる．パターン X の頻度が k となる確率 $f_X(k)$ は，頻度が k となる全ての場合の確率の和となる．例えば，T3 と T4 が共に起きる場合は W7 と W8 であるため，パターン $\{b,h\}$ の頻度が 2 となる確率は $f_{\{b,h\}}(2) = p(W7) + p(W8) = 0.4$ となる．ここで，$p(W)$ は W の起きる確率である．頻度が 1 になる場合は W3 から W6 の 4 パターンのため，その確率は $f_{\{b,h\}}(1) = p(W3) + p(W4) + p(W5) + p(W6) = 0.5$，頻度が 0 になる場合は W1 と W2 であるため，その確率は $f_{\{b,h\}}(0) = p(W1) + p(W2) = 0.1$ となる．

不確実データにおいてパターン X が，

$$p(sup(X) \geq minsup)) \geq minprob \quad (1)$$

を満たすとき，パターン X は頻出であると表現し，X を頻出パターンと呼ぶ．ここで，$minsup$ と $minprob$ はユーザが指定した頻度の最小値（最小頻度）と確率の最小値（最小確率）である．式 1 は，X の頻度が $minsup$ 以上となる確率が $minprob$ 以上になる場合に頻出パターンであることを示す．例えば，$minsup = 1$，$minprob = 0.7$ としたとき，パターン $\{b,h\}$ の頻度が 1 以上となる確率は $f_{\{b,h\}}(1) + f_{\{b,h\}}(2) = 0.9$ であるため，パターン $\{b,h\}$ は頻出パターンである．

2.3 問題定義

不確実データからの頻出パターン抽出問題は，ユーザが指定した $minsup$ と $minprob$ において式 1 を満たすすべてのパターンを抽出することである．

3. 従来手法

3.1 TODIS

不確実データからの頻出パターン抽出ではパターンの SPMF を計算する必要がある．パターンの SPMF 計算では，表 2 のようにトランザクションが起きる場合を考慮し，それぞれの頻度となる確率を計算しなければならないため，処理コストが高い．そこで TODIS では，SPMF 計算を行うパターンである探索候補を限定すること，SPMF が計算済みのパターンを用いた差分計算により SPMF を求めることによって，処理コストの低減を実現している．

3.1.1 探索候補の限定

存在確率を無視した場合の頻度が $minsup$ 未満となる探索候補は，存在確率を考慮した場合にも頻出とはなり得ないため，この探索候補の SPMF 計算を行う必要がない．TODIS ではこの性質を利用して，探索候補を限定する．最初にトランザクションの存在確率を 1 としてパターンの頻度を数え上げ，頻度が $minsup$ 以上となるパターンを取り出し，探索候補とする．ここで頻度が $minsup$ 未満となる探索候補は，存在確率を考慮した場合に頻出の条件を満たさないため，SPMF 計算を省略するする．

TODIS では探索候補と共に，それぞれの探索候補を含むトランザクションのリストを保持する．探索候補 X を含むトランザクションのリストを X の id-list と呼ぶ．この id-list は SPMF の計算の際に利用する．

3.1.2 SPMF の差分計算

探索候補の SPMF 計算は，その探索候補を含むトランザクションが起きる場合を表 2 のように列挙した上で頻度を数え上げ，その確率を計算する必要があるため，SPMF 計算の処理コストが高いという問題がある．TODIS では SPMF が計算済みのパターンの情報を用いた差分計算によって SPMF の計算コストを低減する．探索候補 X の SPMF を計算する場合を考える．単純な SPMF 計算では，X の id-list のトランザクションについて，それらトランザクションが起きる場合をすべて列挙して X の SPMF を計算する．ここで，X を含むパターン Y があり，Y の SPMF が計算済みである場合，Y の id-list に存在しないトランザクションと Y の SPMF を用いて X の SPMF を計算することが可能である．Y の id-list に対する SPMF は計算済みであるため，Y の id-list に存在するトランザクションが起きる場合は既に列挙してある．これに X の id-list に存在するが Y の id-list に存在しないトランザクションを加えてトランザクションが起きる場合を列挙することは，X の id-list のすべてのトランザクションを用いて SPMF を計算したことと等しい．このとき，Y の id-list に存在するトランザクションに対する SPMF 計算処理を省略できるため，X の id-list に存在するすべてのトランザクションから SPMF 計算を行うよりも処理コストを低く抑えることが可能となる．このように X と包含関係にあるパターンの SPMF を利用して X の SPMF を計算する事を差分計算と呼ぶ．

Y の SPMF を利用した X の SPMF 計算を以下に示す．X は Y に含まれるため，X の id-list には Y の id-list には存在せず X の id-list にのみ存在するトランザクションが含まれている．ここで，X の id-list にのみ存在するトランザクションを T_0 から T_{n-1} の n 個とする．Y の SPMF を用いて X の SPMF 計算する処理は，Y の SPMF に T_0 から T_{n-1} のトランザクションを 1 つずつ追加する処理を n 回繰り返すことである．ここで，T_i が起きる確率を p_i，起きない確率を q_i $(= 1 - p_i)$，Y と X の頻度 k の確率を $f_Y(k)$，$f_X(k)$ とする．$f_Y^i(k)$ は $f_Y(k)$ に T_0 から T_i のトランザクションを追加したときに頻度が k となる確率とする．

頻度が 0$(k = 0)$ となるのは全てのトランザクションが起きない場合であるため，$f_Y(0)$ と q_0 から q_{n-1} までの積 $f_Y^{n-1}(0)$ は X の頻度が 0 の確率 $f_X(0)$ となる．$f_X(0)$ は次式で計算できる．

$$f_X(0) = f_Y^{n-1}(0) = f_Y(0) * \Pi_{m=0}^{n-1} q_m \quad (2)$$

頻度が $k(k \geq 1)$ となるのは，トランザクションが k 個起きた場合である．Y の id-list に T_0 から T_i までを追加した SPMF に T_{i+1} を追加したときに頻度が k になる確率 $f_Y^{i+1}(k)$ は，$f_Y^i(k-1)$ と p_{i+1} の積と $f_Y^i(k)$ と q_{i+1} の積の和となる．$f_Y^{i+1}(k)$ は次式で計算できる．

$$f_Y^{i+1}(k) = f_Y^i(k-1) * p_{i+1} + f_Y^i(k) * q_{i+1} \quad (3)$$

T_0 から T_{n-1} までを Y の id-list に追加した $f_Y^{n-1}(k)$ が $f_X(k)$ であるため，$f_X(k)$ は式 3 を $i=0$ から $n-1$ まで n 回繰り返すことで求めることができる．

X を含むパターン Y の SPMF が既に計算してある場合，X のとり得るすべての頻度 k を $k=0$ から順に式 2 と式 3 を用いて計算することで X の SPMF を得ることができる．

3.2 TODIS の問題点

TODIS では探索候補を限定し，SPMF も差分計算によって頻出パターンを抽出する．しかし，長い探索候補から順にすべての探索候補の SPMF を計算する．存在確率を考慮した場合に頻出となる探索候補のみでなく，非頻出となる探索候補も SPMF を計算する．非頻出となる探索候補は SPMF の計算は不要であるが，TODIS はすべての探索候補の SPMF 計算を行うため，非頻出な探索候補の SPMF 計算を回避できないことが問題点である．

4. 提案手法

存在確率を考慮したときに非頻出となる探索候補の SPMF 計算を回避することで性能向上が期待できる．不確実データにおいてもパターン X が頻出パターンとなるには，X に含まれるパターンはすべて頻出パターンでなければならない．そのため，パターン X が非頻出である場合，X を含むパターンは非頻出であることが分かる．この性質を利用することによって，非頻出である探索候補の SPMF 計算を省略し，処理コストの低減を図る．

4.1 探索候補の選出方法

TODIS のように長い探索候補から順に SPMF 計算を行った場合，非頻出パターンを見つけることができても，そのパターンを含む探索候補は既に SPMF を計算済みであるため，SPMF の計算を省略することができない．

本研究では，探索候補のうち中程度の長さの探索候補から順に SPMF 計算を行う．中程度の長さの探索候補から SPMF 計算を行う場合，非頻出のパターンを見つけたときにそのパターンを含み SPMF 計算を行っていない探索候補が残っているため，SPMF 計算を省略することができる．探索候補 X の SPMF 計算を行ったときに X が非頻出となった場合には，X を含むパターンは非頻出であることが分かるため，探索候補から除外する．その場合，X よりも短い探索候補が頻出であると考えられるため，次に X より短い探索候補を選出し，その探索候補の SPMF を計算する．また，探索候補 X の SPMF 計算を行ったときに X が頻出となった場合には，X に含まれる探索候補はすべて頻出となることが分かるため，これら探索候補の SPMF を計算する．そして，X を含むパターンも頻出であると考えられるため，次に X より長い探索候補を選出し，その探索候補の SPMF を計算する．提案手法における探索候補の選出手順を以下に示す．この探索手順をスキップ探索と呼ぶ．

探索候補の中のパターン X を選出し，SPMF を計算する．

- X が頻出の場合:
  - X を含み，X より長い探索候補から，長さを指定しランダムに選出する．選出する探索候補の長さは，SPMF が計算済みかつ X を含む最長の探索候補と X の長さの中央値とする．
- X が非頻出の場合:
  - X を含み，X より短い頻出候補から，長さを指定しランダムに選出する．選出する探索候補の長さは，SPMF が計算済みかつ X が含む最短のパターンと X の長さの中央値とする．
- X と包含関係にある探索候補の全パターンの SPMF を計算した場合は，再び長さが中程度の探索候補をランダムに選出する．

以下で TODIS とスキップ探索の違いを例で述べる．TODIS における SPMF の計算順序を図 1 に示す．矢印の順に示すように

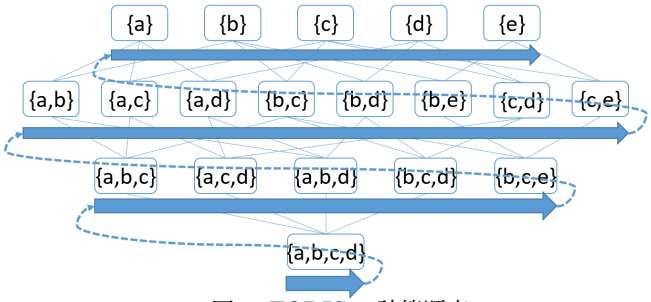

図 1: TODIS の計算順序

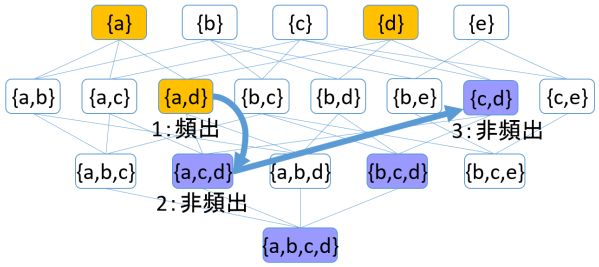

図 2: 提案手法の探索例

長い探索候補から SPMF 計算をする．全ての探索候補の SPMF を計算するため，非頻出の探索候補も SPMF を計算してしまう．

提案手法における SPMF の計算順序の一部を図 2 に示す．最初に探索候補 $\{\mathbf{a,d}\}$ を選出し，SPMF を計算した結果が頻出であったとき，探索候補 $\{\mathbf{a}\}$ と探索候補 $\{\mathbf{d}\}$ も頻出であることが分かるので，これらの SPMF を計算する．次に探索候補 $\{\mathbf{a,c,d}\}$ を選出し，SPMF を計算した結果が非頻出であったとき，$\{\mathbf{a,c,d}\}$ を含む探索候補 $\{\mathbf{a,b,c,d}\}$ は非頻出であることが分かるため SPMF 計算を省略する．さらに次に探索候補 $\{\mathbf{c,d}\}$ を選出して SPMF を計算した結果が非頻出であったとき，探索候補 $\{\mathbf{b,c,d}\}$ も非頻出であることが分かるため SPMF 計算を省略する．3 つの探索候補の SPMF を計算した時点で $\{\mathbf{a,b,c,d}\}$ と $\{\mathbf{b,c,d}\}$ の SPMF 計算が省略できたことが分かる．このようにスキップ探索を行うことによって，非頻出の探索候補を見つけたときに SPMF 計算の省略が可能な探索候補を見つけることができる．

4.2 提案手法の SPMF の差分計算

TODIS では，探索候補 X の SPMF を計算する時に X を含むパターン Y の情報を利用して差分計算を行った．スキップ探索では，SPMF を計算した Y の次に選出される探索候補 X は，Y より短い場合と Y より長い場合がある．X が Y より短く，X が Y に含まれる場合には，TODIS と同様に差分計算をすることができる．X が Y より長く，X が Y を含む場合にも，Y の情報を利用して差分計算によって X の SPMF を求めることが可能である．

パターン Y の SPMF が既知で探索候補 X の SPMF が未知であり，X が Y より長く，X が Y を含む場合の差分計算を示す．Y の id-list には，X の id-list には存在しないトランザクションが含まれている．ここで，Y の id-list にのみ存在するトランザクションを T_0 から T_{n-1} の n 個とする．Y の SPMF を用いて X の SPMF を計算する処理は，Y の SPMF から T_0 から T_{n-1} のトランザクションを 1 つずつ除外する処理を n 回繰り返すことである．ここで，T_i が起きる確率を p_i，起きない確率を q_i，Y と X の頻度が k の確率を $f_Y(k)$，$f_X(k)$ とする．$f_Y^i(k)$ は Y の SPMF から T_0 から T_i までを除外したときに頻度が k となる確率とする．

頻度が 0$(k=0)$ の場合を考える．Y から T_0 を除外したときの頻度が 0 の確率 $f_Y^0(0)$ と Y の頻度が 0 となる確率 $f_Y(0)$ は

$$f_Y(0) = f_Y^0(0) * q_0 \tag{4}$$

となるため，$f_Y^0(0)$ は

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

となる．また，Y から T_0 から T_i までを除外したときの頻度が 0 の確率 $f_Y^i(0)$ とさらに T_{i+1} を除外したときの頻度が 0 の確率 $f_Y^{i+1}(0)$ は

$$f_Y^{i+1}(0) = \frac{f_Y^i(0)}{q_{i+1}} \tag{6}$$

となる．$f_X(0)$ は Y から T_0 から T_{n-1} までの n 個のトランザクションを除外したときの頻度が 0 の確率 $f_Y^{n-1}(0)$ でるため，$f_X(0)$ は次式で計算できる．

$$f_X(0) = f_Y^{n-1}(0) = f_Y(0) * \Pi_{m=0}^{n-1} \frac{1}{q_m} \tag{7}$$

頻度が $k(k \geq 1)$ となるのは，トランザクションが k 個起きたときである．Y の SPMF からトランザクション T_0 を除外したときに頻度が k となる確率 $f_Y^0(k)$ と Y の頻度が k となる確率 $f_Y(k)$ は，

$$f_Y(k) = f_Y^0(k) * q_0 + f_Y^0(k-1) * p_0 \tag{8}$$

となるため，$f_Y^0(k)$ は

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

となる．また，Y から T_0 から T_i までを除外したときの頻度が k の確率 $f_Y^i(k)$ とさらに T_{i+1} を除外したときの頻度が k の確率 $f_Y^{i+1}(k)$ は

$$f_Y^i(k) = f_Y^{i+1}(k) * q_{i+1} + f_Y^{i+1}(k-1) * p_{i+1} \tag{10}$$

となるため，$f_Y^{i+1}(k)$ は次式となる．

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

X の頻度が k となる確率 $f_X(k)$ は，$f_Y(k)$ の頻度 $k = 0$ から X の頻度が取り得る範囲まで順に $i = 0$ から $n-1$ までの T_i を除外する処理を繰り返し計算した値 $f_Y^{n-1}(k)$ となる．X に含まれる探索候補 Y の SPMF が既に計算してあり差分計算が使用可能なときには，式 7 と式 11 を用いて，$k = 0$ から順に計算することで X の SPMF を得ることができる．

4.3 処理手順

提案手法のスキップ探索の処理手順を以下に示す．

1. TODIS と同様に存在確率を 1 としてパターンの頻度を数え上げ，頻度が *minsup* 以上となるパターンを探索候補とする．このとき，それぞれのパターンの id-list も取得する．
2. 初めに，最長の探索候補の $\frac{1}{2}$ の長さの探索候補をランダムに 1 つ選択する．
3. 選択した探索候補 X の SPMF を計算する．SPMF が計算済みであり，X を含むまたは X が含むパターンがある場合，差分計算を行う．
   - X が頻出の場合，X に含まれる探索候補の SPMF を計算する．
   - X が非頻出の場合，X を含むパターンを探索候補から除外する．
4. X の SPMF 計算結果に従って次の探索候補を 1 つ選出する．
5. すべての探索候補の SPMF 計算を行うまたは非頻出と判定されるまで 3 と 4 の処理を繰り返す．

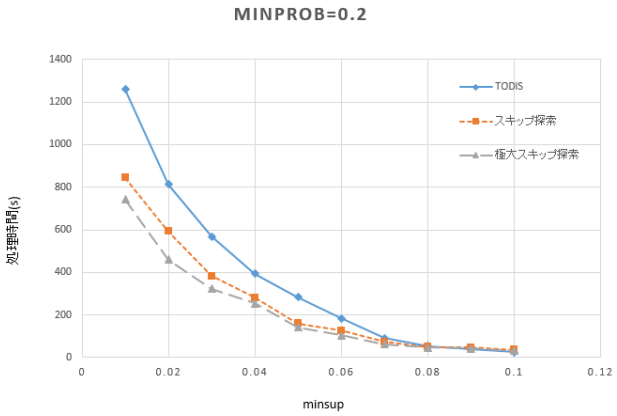

図 3: *minsup* を変化させた場合の処理時間

スキップ探索では，最初に中程度のパターンをランダムに選出した．包含関係にある探索候補がない場合にも同様に中程度のパターンをランダムに選出していた．これらの SPMF 計算では差分計算を行えないため，この探索候補を含むすべてのトランザクションを用いて SPMF を計算する必要があった．差分計算を行えない場合の SPMF 計算の処理コストは高くなる．この問題を回避する手法を極大スキップ探索と呼ぶ．極大スキップ探索では，最初に極大パターンである探索候補の SPMF 計算を行う．極大パターンは他のパターンに含まれないパターンである．極大パターンの SPMF を計算しておくことで，その他の探索候補の SPMF 計算は常に差分計算が可能となる．また，極大パターンは探索候補の中で長いパターンであるため，そのパターンを含むトランザクションの数が短いパターンと比較して少ないため，SPMF 計算の処理コストも高くない．

5. 性能評価

スキップ探索と極大スキップ探索の二つの提案手法と TODIS を実装し，評価実験を行う．実験には Frequent Itemset Mining Dataset Repository(FIMI)[10] により公開されている，1990 年から 2000 年までのフランダース，ベルリン間における交通事故のデータを利用した．本データは，アイテム数が 572，トランザクションの平均長が 45，トランザクション数は約 34 万件である．

5.1 処理時間の比較

提案するスキップ探索，極大スキップ探索と TODIS の処理時間を測定した．ここで測定した処理時間には探索候補を求める処理は含めていない．探索候補を求める処理は従来の頻出パターンマイニング技術が利用可能であり，最も代表的な Apriori アルゴリズムを利用した場合でも処理時間は 230 秒から 410 秒であり，探索候補に対して SPMF 計算を行いながら頻出パターンを抽出する処理と比べて短い．そのため，本実験では探索候補を求める処理時間は除外した．

図 3 に *minsup* を変化させた場合の処理時間を示す．ここで，*minprob* を 0.2 に固定し，縦軸を処理時間，横軸を *minsup* とした．*minsup* が高いときは処理時間の差が見られない．しかし，*minsup* の値が低くなるに従って処理時間が長くなるが，二つの提案手法と TODIS の処理時間の差も増大している．*minsup* が 0.01 の時，TODIS と比較してスキップ探索では約 17%，極大スキップ探索では約 25%の処理時間を短縮できた．*minsup* の値が低いときに TODIS と二つの提案手法の処理時間に大きな差が見られたが，これは交通事故のデータは頻出パターンの平均長が長く，非頻出として SPMF 計算を省略した探索候補の数が多くなっていたためである．

次に，図 4 に *minprob* を変化させた場合の処理時間を示す．ここで，*minsup* を 0.02 に固定し，縦軸を処理時間，横軸を *minprob*

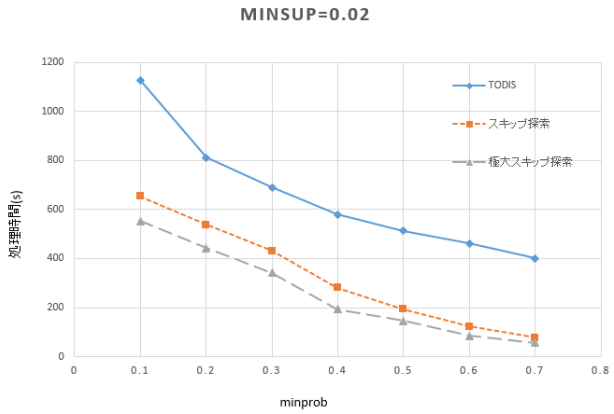

図 4: *minprob* を変化させた場合の処理時間

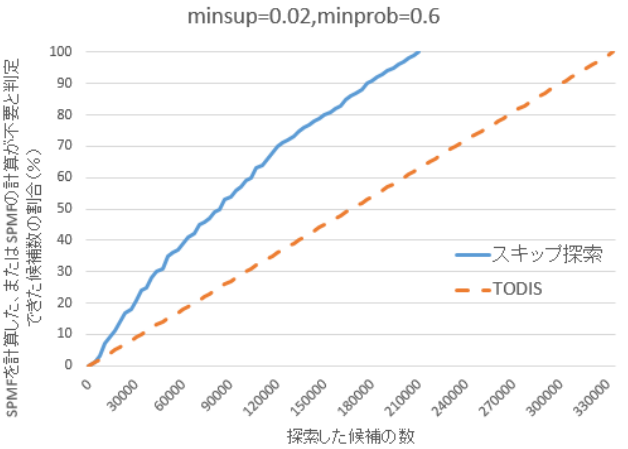

図 5: 探索効率の比較

を横軸とした．*minprob* が増大するにしたがって TODIS と二つの提案手法の処理時間の差が増大している．TODIS と比較して，スキップ探索では最低 25%，最大 75%，平均約 51%の処理時間を削減した．極大スキップ探索では最低 39%，最大 76%，平均 56%の処理時間を削減した．本実験では *minsup* が固定であるため，*minprob* を変化させても存在確率を無視して抽出した探索候補の数が変わらない．*minprob* が小さい場合は非頻出として SPMF 計算を省略できる探索候補数が少なくなるため，TODIS と提案手法との処理時間の差が減少する．また，*minprob* の増大するに従って提案手法であるスキップ探索と極大スキップ探索の処理時間の差が減少していることが分かる．*minprob* が増大するに従って頻出となるパターンが少なくなり，非頻出となるパターンが多くなる．探索候補の中で非頻出となるパターンが多いことからスキップ探索と極大スキップ探索の差が少なくなっていると考えられる．

5.2 探索の効率の比較

次に，提案手法の SPMF 計算の省略の効果を確認する．図 5 に何個目の探索候補の SPMF を計算した時点で，どのくらいの探索候補の判定が完了したかをプロットした図を示す．*minsup* を 0.02，*minprob* を 0.6 とし，頻出パターン抽出を行った．横軸は SPMF 計算を行った探索候補数であり，縦軸は全探索候補に対して SPMF 計算を行ったまたは非頻出と判定された探索候補数の割合である．例えば，探索候補の SPMF 計算を 90000 個行った時点で，TODIS では約 28%，スキップ探索では約 56%の探索候補の処理が完了していることが分かる．TODIS では長い探索候補から順に 1 つずつ SPMF 計算を行うため，線形になる．スキップ探索では SPMF 計算を行った探索候補が頻出の場合には TODIS と同様に線形になる．しかし非頻出の場合には，その時点で複数の探索候補が非頻出と判定されるため，傾きが急になる．スキップ探索では，約 120000 番目以前の傾きが急であり，120000 番目以降から緩やかになっている．これは，SPMF 計算を開始した直後は SPMF 計算が完了したまたは非頻出と判定された探索候補が少なく，非頻出の探索候補を見つけやすいためと考えられる．100%まで完了した時点でのスキップ探索の SPMF を計算した探索候補数は TODIS の約 66%となっており，スキップ探索における SPMF 計算の省略の効果が示されている．SPMF 計算を行った探索候補数は，スキップ探索では理想値の 140%から 150%であった．

6. おわりに

本研究では，TODIS における非頻出となる探索候補に対する SPMF 計算の問題点を解決する手法として，スキップ探索による非頻出な探索候補の SPMF 計算の省略と差分計算を提案した．実験により，SPMF 計算を行う探索候補数の削減と処理効率の向上を実現できたことを示し，提案手法の有効性を示した．

[謝辞]

本研究は JSPS 科研費 25280022 の助成を受けたものです．

[文献]

[1] R. Agrawal, R.Srikant, ”Fast algorithms for mining association rules”, VLDB, pp.487–499, 1994.

[2] C. Chui, B. Kao, E. Hung, ”Mining frequent itemsets from uncertain data”, PAKDD, pp.47–58, 2007.

[3] L. Sun, R. Cheng, D. W. Cheung, J. Cheng, ”Mining uncertain data with probalilistic guarantees”, KDD, pp.273–282, 2010.

[4] L. Wang, D. W. Cheung, R. Cheng, s. Lee, X. Yang, ”Efficient mining of frequent itemsets on large uncertain databases”, IEEE TKDE, 24(12), pp.2170–2183, 2012.

[5] J. Han, J. Pei, Y. Yin, ”Mining frequent patterns without candidate generation”, ACM SIGMOD, pp.1–12, 2000.

[6] C. K. Leung, C. I. Carmichael, B. Hao, ”Efficient mining of frequent patterns from uncertain data”, IEEE ICDM Workshops, pp.489–494, 2007.

[7] C. K. Leung, M. A. F. Mateo, D. A. Brajczuk, ”A tree-based approarch for frequent pattern mining from uncertain data”, PAKDD, pp.653–661, 2008.

[8] C. C. Aggarwal, Y. Li, J. Y. Wang, J.Wang, ”Frequent pattern mining with uncertain data”, KDD, pp.29–38, 2009.

[9] L. Wang, L. Feng, M. Wu, ”AT-Mine: an efficient algorithm of frequent itemset mining on uncertain dataset”, J. of Computers, 8(6), pp.1417–1426, 2013.

[10] Frequent Itemset Mining Dataset Repository(FIMI), http://fimi.ua.ac.be/data

建島 大翔　Hiroto TATESHIMA
2015 年 電気通信大学大学院情報システム学研究科修了，修士（工学）．同年（株）エス・ジー入社．

新谷 隆彦　Takahiko SHINTANI
電気通信大学大学院情報システム学研究科准教授．1999 年東京大学大学院情報工学専攻博士課程修了，博士（工学）．2000 年（株）日立製作所中央研究所研究員，同主任研究員を経て，2011

年より現職．データマイニング，ライフログ活用に関する研究に従事．情報処理学会，電子情報通信学会各会員．

大森 匡　Tadashi OMORI

電気通信大学大学院情報システム学研究科教授．1990 年東京大学大学院情報工学専攻博士課程修了，工学博士．1994 年より電気通信大学大学院情報システム学研究科助教授．データベースシステムとトランザクション処理の研究に従事．ACM，IEEE，情報処理学会各会員．

藤田 秀之　Hideyuki FUJITA

電気通信大学大学院情報システム学研究科助教．2006 年東京大学大学院新領域創成科学研究科博士課程修了，博士（環境学）．東京大学空間情報科学センター研究機関研究員，同助教を経て，2013 年より現職．空間情報科学，インタラクティブソフトウェアに関する研究に従事．情報処理学会，地理情報システム学会，日本地図学会各会員．