1F3-03

相関の違いに基づくローカルトランザクションデータベースにおける隠れた相関の発見

Discovery of Implicit Itemset Pair based on Differences of Correlations in a Local Transaction Database

谷口剛　　原口誠
Tsuyoshi TANIGUCHI　Makoto HARAGUCHI

*1北海道大学大学院情報科学研究科コンピュータサイエンス専攻
Graduate School of Information Science and Technology Hokkaido University

Given a transaction database and its sub-database, we consider a pair of itemsets whose degrees of correlations are higher in the local database than in the global one. Since some of them show high correlation, they are considered to be characteristic in the local database and detectable by some previous study methods. On the other hand, there exist another kind of paired itemsets such that they are not regarded as characteristic ones but that their degrees of correlations become drastically higher by the conditioning to the local database. We pay much attention to the latter kind of paired itemsets, as such pairs of itemsets can be an implicit and hidden evidence showing that something particular to the local database occurs even though they are not yet realized as characteristic ones. From this viewpoint, we define a notion of DC pairs whose differences of correlations are high. In this paper we consider a problem of efficient search of DC pair and a solution of the problem.

1. はじめに

大規模なトランザクションデータベースを対象としたデータマイニングの研究において，相関ルール [1] や強く相関しているアイテム集合 (の組) [3, 4] のようなある特徴を持ったアイテム集合を見つける研究が高い注目を集めている．これらの研究は，あるデータベースにおいて特徴的なアイテム集合を識別するために有用である．また Emerging Patterns [2] のような，比較しているデータベース間におけるアイテム集合の支持度の変化に基づく研究が，与えられた 2 つのデータベースのうちのいずれかのデータベースを特徴付けるようなアイテム集合を見つけることに対して成果を上げている．このように，アイテム集合に関する研究の多くは，与えられた単一のデータベース，あるいは与えられたいくつかのデータベースのうちのいずれかのデータベースにおいて，特徴的であるアイテム集合あるいはアイテム集合の組を抽出することに主眼を置いている．

一方，Chance Discovery の研究 [5] のように，上述の意味で特徴的でないアイテム集合もまた，ある条件において潜在的に重要な情報になりえるので有用である．例えば，ある特定の地区におけるスーパーマーケットのデータベースがあると仮定しよう．そのデータベースの中には，実際に同時に買われた商品の数と統計的独立を仮定したときのその商品が同時に買われると期待される数のデータベースにおける割合がほぼ同じ値である商品 (アイテム集合) の組が存在するかもしれない．この商品の組が統計的独立の状態にあるならば，それぞれの割合は同じ値になるので，この関係は高い相関を持っているわけではなく，特徴的なアイテム集合の組であるとは考えられない．しかしここで分析しようとしている特定の地区を含む全ての地区のデータベースを考えたとき，実際に同時に買われた割合が統計的独立を仮定したときに期待される割合に対して極端に低かった場合を考えてほしい．この場合，その地区におけるこの商品の組の相関は高いというわけではないが，その地区においてこの商品の組み合わせに何かが起こっていると考えることができるだろう．なぜならば全体のデータベースからある特定の条件付けをしてローカルデータベースを考えることにより，相関の度合いが非常に低い値から大きな変化を見せているからである．したがって，この現象に注目し，この変化がなぜ生じたかを分析し，その分析を考慮して新たな販売戦略を実現することには価値があると考えられる．

ここまで述べてきたような観点に基づき，与えられた全体のデータベースとある条件付けによって得られるそのローカルデータベースに対して，本研究の目的は，次のような性質を持つアイテム集合の組を見つけるアルゴリズムを提案することである．(1) アイテム集合間の相関が全体のデータベースにおいてよりもローカルデータベースにおいての方が非常に高くなるが，(2) 必ずしも特徴ではないアイテム集合の組．ここであるデータベースにおいてアイテム集合の組が特徴的であるとは，それらの間の相関が高いことを言う．つまり，本研究ではローカルデータベースに条件付ける前後の相関の違いを観測する．全体のデータベースとローカルデータベースで相関の違いが大きいアイテム集合の組のことを DC pair と呼ぶ．

DC pair を求める問題は一般に難しい問題である．なぜならば，相関の違いを評価する関数は非単調に変化するからである．そこで，本研究ではこの問題を解決するためにいくつかの枝刈り規則を提案してきた．しかし，現在のところ，十分な成果を挙げているとは言えない．そこで，本稿では本研究の問題点を明らかにし，現在取り組んでいる課題についてまとめる．

2. 準備

$I = \{i_1, i_2, \ldots, i_m\}$ をアイテムの集合とする．I の部分集合 $X \subseteq I$ をアイテム集合という．トランザクションデータベース $\mathcal{D}$ はトランザクションの集合とする．ここで，トランザクションはアイテム集合である．もし $X \subseteq t$ ならば，トランザクション t はアイテム集合 X を含むという．トランザクションデータベース $\mathcal{D}$ とアイテム集合 X に対して，$\mathcal{D}$ における X を含むトランザクションの集合を，$O(X, \mathcal{D})$ と記述し，$O(X, \mathcal{D}) = \{t | t \in \mathcal{D} \wedge X \subseteq t\}$ と定義する．そして，$\mathcal{D}$ にお

連絡先: 谷口剛，北海道大学大学院情報科学研究科，〒 060-0814 札幌市北区北 14 条西 9 丁目，TEL(FAX):011-706-7161, E-mail:tsuyoshi@kb.ist.hokudai.ac.jp

ける X の確率を $P(X)$ と記述し，$P(X) = |O(X,\mathcal{D})|/|\mathcal{D}|$ と定義する．

アイテム集合 C に対し，C に関する $\mathcal{D}$ のサブデータベースは，$\mathcal{D}_C$ と記述し，$\mathcal{D}$ において C を含むトランザクションの集合，つまり $\mathcal{D}_C = O(C,\mathcal{D})$ と定義する．$\mathcal{D}$ に関する $\mathcal{D}_C$ の補データベースは $\overline{\mathcal{D}_C}$ と記述し，$\overline{\mathcal{D}_C} = \mathcal{D} - \mathcal{D}_C$ と定義する．

アイテム集合 X と Y に対し，トランザクションデータベース $\mathcal{D}$ における X と Y の相関 $correl(X,Y)$ を，$correl(X,Y) = P(X|Y)/P(X)$ と定義する．ここで，$P(X|Y) = P(X \cup Y)/P(Y)$ である．サブデータベース $\mathcal{D}_C$ に対して，$\mathcal{D}_C$ における X と Y の相関を $correl_C(X,Y)$ と記述し，$correl_C(X,Y) = P(X|Y \cup C)/P(X|C)$ と定義する．ここで相関は $\mathcal{D}$ と $\mathcal{D}_C$ において支持度が 0 でないアイテム集合に対してのみ定義されることに注意する．本研究において，$correl(X,Y) > 1$ を満たす X と Y の組を特徴的であると考える．なぜならば $P(X|Y) > P(X)$ であるからである．ここで，$P(Y|X) > P(Y)$ も同様に成り立つことに注意する．同様の理由で，$correl(X,Y) \leq 1$ が成立するような X と Y の組を特徴的ではないと考える．

3. DC pair 探索問題

この節では，DC pair と DC pair を探索する問題について定義する．

アイテム集合 X と Y の組に対して，本研究では，"サブデータベースに条件付けることによって観測される相関の違い"に注目する．ここで，相関の違いは以下の比率によって評価される．

$$change(X,Y;C) = \frac{correl_C(X,Y)}{correl(X,Y)} = \frac{P(C)P(C|X \cup Y)}{P(C|X)P(C|Y)}. \quad (1)$$

$\rho(>1)$ を相関の違いのパラメータとすると，本研究では，アイテム集合 X と Y の組に対して，$change(X,Y;C) \geq \rho$ を満たすならば重要な関係であると考える．ここで C はユーザによって与えられると仮定し，$P(C)$ を定数とみなす．したがって，実際には相関の違いを以下の関数 g によって評価する．

$$g(X,Y;C) = \frac{P(C|X \cup Y)}{P(C|X)P(C|Y)}. \quad (2)$$

ここで，$g(X,Y;C) \geq \rho/P(C)$ を満たすようなアイテム集合 X と Y の組を DC pair と呼ぶ．本研究では，全ての DC pair を効率的に見つけたい．ここで g がアイテム集合 X と Y のいずれかのアイテム数の増加に関して非単調に変化することに注目しなければならない．このために，DC pair を求める際に apriori [1] のような単純な枝刈りを行うことができない．それゆえ，以下のような素朴な考え方によって上記の問題を近似する．

> $P(C|X), P(C|Y)$ を低く保ちながら，$P(C|X \cup Y)$ が高くなるような X, Y の組を見つける．

もう 1 つの新たなパラメータ $\zeta(0 \leq \zeta \leq 1)$ を用いて，本研究で扱う問題を以下のように定義する．

定義 1. DC pair 探索問題
C を条件付けのためのアイテム集合とする．ρ と ζ が与えられたとき, DC pair 探索問題は $P(C|X \cup Y) > \zeta$, $P(C|X) < \epsilon$ and $P(C|Y) < \epsilon$ を満たすような全ての X と Y の組を見つけることである．ここで $\epsilon = \sqrt{\zeta \cdot P(C)/\rho}$ とする．

4. アルゴリズム

この節では，DC pair 探索問題を解くためのアルゴリズムについて説明する．3. において，ζ と ϵ というパラメータを用いて，求める DC pair を限定した．しかし，$P(C|Z)$ も Z のアイテム数の増加に関して g と同様に非単調に変化する．つまり，このままでは単調性に基づく枝刈りは望み得ない．そこで，今調べているアイテム集合からトップダウンに問題を考えることにより，DC pair 探索問題を効率的に計算できるかもしれない枝刈り規則を導くことができた．したがって，ここではトップダウンに $P(C|Z) > \zeta$ なる Z から見つけていくアルゴリズムについて説明する．ここで，トップダウンにアイテム集合を調べていくために，最初に $\mathcal{D}_C$ における極大アイテム集合を計算する．なぜならば $P(Z|C) > 0$ を満たさなければならないからである．結局 DC pair 探索問題は以下の 2 つの段階に分けられる．

Phase1: 組み合わせの候補を識別
$P(C|Z) > \zeta$ であるようなアイテム集合 Z は，DC pairX, Y から得られる組み合わせ $Z = X \cup Y$ の候補として識別される．

Phase2: 組み合わせを分割
それぞれの組み合わせの候補 Z は，$Z = X \cup Y$，$X \cap Y \neq \emptyset$，$P(C|X) < \epsilon, P(C|Y) < \epsilon$ であるようなアイテム集合 X と Y に分割される．

上記のアルゴリズムにおいて，ある候補の Z が DC pair に分解できる可能性がないことがあらかじめわかる場合もありえる．したがって，Phase1 においても Z が DC pair に分解できる可能性があるかを調べながら探索することが考えられる．そこで，まずはアルゴリズムの基本的な概要について述べ，その後に DC pair への分解性を考慮して，より洗練されたアルゴリズムを導入する．

4.1 ドロップ規則による枝刈り

ここでは組み合わせの候補を識別する探索における基本的な枝刈りについて議論する．g と同様 $P(C|Z)$ も Z のアイテム数の増加に関して非単調に変化するため単調性に基づく枝刈りは望み得ない．そこで，$\mathcal{D}_C$ において見つかったそれぞれの極大トランザクション Z_{max} に対して，まず Z_{max} を調べ，その後にその部分集合を調べていくようなトップダウン探索を考える．この探索を考えたときに，以下に示すような枝刈り規則を導くことができ，有用でない枝 (アイテム集合) を枝刈りできる．

アイテム i を含むアイテム集合を Z とし，$i \in Z' \subset Z$ かつ $P(C|Z') > \zeta$ を満たす Z の部分集合 Z' が存在すると仮定する．$P(C|Z') = P(C)P(Z'|C)/P(Z') > \zeta$ より，$P(Z'|C) > \zeta \cdot P(Z')/P(C)$ である．よって，$P(i|C) \geq P(Z'|C) > \zeta \cdot P(Z')/P(C) \geq \zeta \cdot P(Z)/P(C)$ となる．したがって，$P(C \cup i) > \zeta \cdot P(Z)$ を得る．このことは $P(C \cup i) \leq \zeta \cdot P(Z)$ が成り立つならば，i を含む Z の部分集合 Z' 中には $P(C|Z') > \zeta$ を満たすような Z' が存在し得ないことを意味する．つまり，Phase1 において探索ノードとして Z を仮定したとき，$P(C \cup i) \leq \zeta \cdot P(Z)$ が成り立つならば，i を含む Z の部分集合は調べる必要はないことがわかる．それゆえ，Z から i を安全に除去 (ドロップ) することができる．

ドロップ規則:
探索ノード (アイテム集合)Z とアイテム $i \in Z$ に対して，トップダウン探索の過程において $P(C \cup i) \leq \zeta \cdot P(Z)$ ならば，i を含む部分集合 Z' は Z の子ノードにはなりえない．つまり，

全ての子ノードは Z においてドロップされていないアイテムから成る．

ドロップ規則の特別な場合として，全てのアイテム $i \in Z$ がドロップされ，Z の部分集合は調べる必要がないことがわかる場合がある．

停止条件:
探索ノード (アイテム集合)Z に対して，$\max\{P(C \cup i)|i \in Z\}/\zeta \leq P(Z)$ が成り立つならば，Z の子ノードは生成しない．

上記の停止条件は Phase1 における探索の理論的な下限を与える．$i \in Z$ および $P(Z|C) > 0$ より，$P(i|C) > 0$ が成り立つ．したがって，以下を得る．

Phase1 における探索の理論的下限 :
Phase1 において探索ノード Z が生成されるならば，$P(Z) \leq maxsup_\zeta$.
ここで，$maxsup_\zeta = \max\{P(C \cup i)|P(i|C) > 0\}$．つまり，Phase1 において $maxsup_\zeta$ を超える支持度を持つ Z は生成されることは決してあり得ない．

4.2 DC pair への分解を考慮した枝刈り

4.1 で説明した枝刈りは，Phase2 における制約を考慮することによって，より強力にすることができる．つまり，組み合わせの候補を識別する探索において DC pair への分解性を調べながら探索を行うことにより，さらに有用でない枝 (アイテム集合) を刈ることができる．

Phase2 において，Phase1 で見つけた組み合わせの候補 Z は $P(C|X) < \epsilon$，$P(C|Y) < \epsilon$ を満たすような X と Y の組に分解される．上述の議論と同様に，$i \in X \cup Y (= Z)$ に対し，$P(C \cup Z) < \epsilon \cdot P(i)$ が成り立つ．それゆえ，$P(C \cup Z) \geq \epsilon \cdot P(i)$ を満たすようなアイテム $i \in Z$ が存在するならば，i を含む Z の部分集合は ϵ の制約を満たす 2 つの部分に分解することができない．したがって，Z から i をドロップすることができる．このように，探索ノードをより強力に枝刈りするドロップ規則 (Revised) を得ることができる．

ドロップ規則 (Revised):
探索ノード Z とアイテム $i \in Z$ に対して，$P(C \cup i) \leq \zeta \cdot P(Z)$ あるいは $P(i) \leq P(C \cup Z)/\epsilon$ ならば，Z から i をドロップすることができる．

ドロップ規則 (Revised) によって，新たな停止条件と探索の理論的下限が与えられる:

停止条件 (Revised):
探索ノード Z に対して，$\max\{P(C \cup i)|i \in Z\}/\zeta \leq P(Z)$ あるいは $\max\{P(i)|i \in Z\} \leq P(C \cup Z)/\epsilon$ ならば，Z の子ノードは生成しない．

Phase1 における探索の理論的下限 (Revised):
Phase1 において探索ノード Z が生成されるならば，$P(Z) \leq maxsup_\zeta$ かつ $P(C \cup Z) \leq \epsilon \cdot maxsup_\epsilon$．ここで $maxsup_\epsilon = \max\{P(i)|P(i|C) > 0\}$.

4.3 組み合わせの候補の分解

Phase2 において，Phase1 で得られた組み合わせの候補であるアイテム集合 Z は $Z = X \cup Y$，$X \cap Y = \emptyset$，$P(C|X) < \epsilon, P(C|Y) < \epsilon$ を満たすようなアイテム集合 X, Y に分解される．DC pair への分解において，Z を最大限とするアイテム集合束を考え，そのアイテム集合束を単独のアイテムから Z まで以下のような枝刈り規則を用いながら，ボトムアップ探索で $X \subset Z$ を列挙していく．

Phase 2 におけるドロップ規則:
探索ノード (アイテム集合)X とアイテム $i \in X$ に対して，$P(C \cup Z) \leq \epsilon \cdot P(i \cup X)$ ならば，i を含む X の上位集合は調べる必要がない．

上記の規則は Phase1 におけるドロップ規則と完全に双対であるため同様に証明され，ドロップされうるアイテムを含む次のノードに対して枝刈りを行うことができる．

5. 実験

5.1 実験データと実装

まずは，用いたデータベースについて説明する．本実験では，UCI KDD Archive (http://kdd.ics.uci.edu) におけるデータベースのうち，Entree Chicago Recommendation Data を用いた．このデータはレストランの特徴を表しており，アメリカの Los Angeles や New Orleans などの 8 つの地区のデータベースから成る．全体の地区と比較してある特定の地区における DC pair を見つけるため，それぞれの地区を表すアイテムを考え，全てのトランザクションに対して追加した．この操作によって，4160 トランザクション，265 アイテムから成る統合されたデータベースを得た．データベースに元々与えられていたアイテムは”Italian”，”romantic”，”parking”などであり，様々なレストランの特徴を表している．この統合されたデータベースを用いて，C 言語でシステムを実装し，DC pair を見つけた．全ての実験は，PentiumIV， 1.5GHz，主記憶 896MB のスペックを持つ PC によって行った．

本実験の前に行った予備実験において，本研究の枝刈り規則はサイズの大きいアイテム集合には効きづらいということがわかった．その理由については後の節で述べる．本実験ではサイズが小さいアイテム集合における枝刈り規則の性能を調べるために，サイズが 6 以下のアイテム集合の中で極大なアイテム集合から探索をはじめることにする．

5.2 実験結果

実験結果を図 1 に示す．本実験において，ρ, ζ はそれぞれ 3.0，0.4 に設定した．ここで，$|N_{full}|$ は $\mathcal{D}_\mathcal{C}$ におけるサイズ 6 以下のアイテム集合の数であり，$|N_{drop}|$ は Phase1 において実際に調べられたアイテム集合の数である．そして $|N_{P(C|Z)>\zeta}|$ は $\mathcal{D}_\mathcal{C}$ において $P(C|Z) > \zeta$ を満たすようなサイズが 6 以下のアイテム集合 Z の数である．$|DC|$ は見つかった DC pair の数である．最後に $|DC_{NotCor}|$ は DC pair となるアイテム集合の間の相関の値が 1 以下の数である．

実験データには様々な種類の DC pair が存在した．例えば，ニューオリンズにおいて $X = \{$*Entertainment*, *Quirky*, *Up and Coming*$\}$, $Y = \{$*$ 15-$ 30*, *Private Parties*, *Spanish*$\}$ という DC pair が見つかった．この pair はニューオリンズに条件付けることによって，高い相関性変化を示し，もちろん重要な関係である．しかし，この pair は相関性の変化の結果として，ニューオリンズにおいて非常に高い相関を示すので，従来手法で見つけることができる．さらに，実験データから見つかったこのような DC pair は全体のデータベースにおいても相関していることが多かった．つまり，わざわざニューオリンズにおける DC pair として見つける必要のない情報であると言えるかもしれない．一方，DC pair の中には，$X = \{$*Quirky*$\}$, $Y = \{$*Good Decor*, *Italian*, *$15-$30*, *Good Service*$\}$ というものも存在した．この pair は全体のデータベースにおいても部分のデータベースにおいても相関していなかった．したがって，従来手法で見つけることはできない．しかし，相関性変化量は非常に大きい値を示す．このことは，X と Y を同時に特徴として持つレストランの期待される値に対する割合は非常に低いが，ニューオリンズでは中間くらい (高くはない) とい

$\rho = 3.0, \zeta = 0.4$							
region	$sup(C)$	ϵ	$\|N_{full}\|$	$\|N_{drop}\|$	$\|N_{P(C\|Z)>\zeta}\|$	$\|DC\|$	$\|DC_{NotCor}\|$
Atlanta	6.4	0.0922	1922264	1826678	1575575	112877	269
Los Angeles	10.7	0.118	1857501	1760522	1769705	30404	97
New Orleans	7.9	0.102	1120224	1071306	1027241	39158	55
San Francisco	10.0	0.114	2154595	2113443	1735822	134520	312

図 1: 実験結果

うことを意味している．この関係が，本研究が注目している関係である．本研究ではこのような情報もある場面では有効に働きうると考えている．例えば，ニューオリンズでレストランを探している人は，ニューオリンズにおいて特徴的な quirky spanish restaurant よりニューオリンズにおいて隠れた特徴である quirky Italian restraunt の方に行ってみたいと思うかもしれない．以上より，実験データには実際に潜在的に重要な DC pair が存在し，本研究のアルゴリズムはそのような DC pair を見つけることができることが示された．実験データからは上記の pair 以外にも様々な潜在的に重要な DC pair が発見された．

5.3 本研究の課題

この節では，現在本研究が抱えている課題についてまとめる．

DC pair の探索において本研究で提案していた枝刈り規則の効果は十分とは言えなかった．そこで，現実の問題に応用するには更なる工夫が必要である．DC pair を効率的な探索を妨げる要因としては，大きく以下の 2 つが挙げられる．

1. 枝刈り規則の適用機会の少なさ．
   本研究の枝刈り規則は，簡単な評価で多くのアイテム集合を枝刈りできる可能性がある．しかし，実際には枝刈りの機会が少なく，結果として有効に探索対象を減らすことができなかった．そこで，枝刈り規則の性質を分析したところ，全体，部分いずれのデータベースであっても今調べているアイテム集合とそれに含まれるアイテムの支持度の差が大きい場合，本研究の枝刈り規則は有効に働くことが難しいことがわかった．このことが，サイズが大きいアイテム集合に規則が効かなかった原因でもある．したがって，この問題を解決し枝刈りの機会をもっと増やすために，枝刈り規則の適用条件を緩和し，手続きを修正する必要がある．

2. 今回の実験データにおける組み合わせの候補の多さ．
   図 1 において，Altanta では全探索対象である 192 万のアイテム集合のうち，10 万のアイテム集合しか枝刈りできなかったように見える．しかし，データには Step1 で求めている組み合わせの候補の数が 157 万個あるので，実は 35 万のアイテム集合しか探す必要がないものがないということもできる．ここで，Los Angeles において，実際に調べた探索対象の数が $P(C|Z) > \zeta$ を満たす Z の数を下回っていることに注目してほしい．この現象には，DC pair の分解性の制約が影響を与えている．したがって，DC pair の分解性をもっと考慮することよって，探すべき探索対象を少なくできる可能性がある．

6. まとめと今後の課題

与えられたトランザクションデータベース $\mathcal{D}$ とそのサブデータベース $\mathcal{D}_\mathcal{C}$ に対し，本研究では DC pair というアイテム集合の組の考え方を提案した．アイテム集合の組 X と Y に対し，$\mathcal{D}_\mathcal{C}$ における X と Y の相関が $\mathcal{D}$ の相関と比べてある割合よりも高いとき，そのアイテム集合の組 X と Y のことを DC pair と呼ぶ．注目すべきことは，たとえ $\mathcal{D}$ と $\mathcal{D}_\mathcal{C}$ における相関の違いがある程度観測できたとしても $\mathcal{D}_\mathcal{C}$ における相関は高いとは限らない点である．この意味で，そのような組み合わせは $\mathcal{D}_\mathcal{C}$ において特徴的ではないと考えられる．このように，DC pair はサブデータベースにおいて潜在的な特徴であると考えられる．

本研究で行った実験において，潜在的に重要な DC pair は確かに見つけることができたが，提案した枝刈り規則によって十分に効率的な計算ができているとはいえなかった．そこで，(1) 枝刈り規則の緩和と手続きの修正，(2)DC pair の分解性のさらなる考慮，の 2 つの対策を行っているところである．予備実験の結果より，特に (2) よって探索すべきアイテム集合をいままでの 2 分の 1 以下にできる可能性がある．実験結果については別の機会に報告したい．

最後に，今後の展望について議論する．本研究では今後の展望として時系列データへの応用を考えている．本稿においては，全体のデータベースとローカルデータベースにおける相関の違いに注目した．この考え方に基づいて，ある時間 t_1 とある時間 t_2 における相関の違いに注目すれば，もちろん時系列データ特有の知識も考慮しなければならないが，本研究の考え方は簡単に時系列データに応用することができる．この問題において，それぞれの時間 t_1，t_2 における特徴的な相関は従来の手法を用いて見つけることができるだろう．しかし，t_2 の後に来る t_3 において特徴的になる可能性のあるアイテム集合の組 X，Y を見つけたい場合には，もちろん時間の区切り方も問題になるが，t_1 から t_2 における時間の変化における大きな相関性変化をとらえることによって実現できる可能性がある．本研究では時系列データへの応用を検討している段階である．

参考文献

[1] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules, in *Proc. of the Int'l Conf. on Very Large Data Bases*, pp. 487–499 (1994).

[2] Dong, G. and Li, J.: Efficient mining of emerging patterns: discovering trends and differences, in *Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 43–52 (1999).

[3] Brin, S., Motwani, R. and Silverstein, C.: Beyond market baskets: generalizing association rules to correlations, in *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, Vol. 26, No. 2, pp. 265–276 (1997).

[4] Morishita, S. and Sese, J.: Traversing itemset lattices with statistical metric pruning, in *Proc. of the ACM SIGACT-SIGMOD-SIGART Symposium on Database Systems (PODS)*. pp. 226–236 (2000).

[5] Ohsawa, Y. and Nara, Y.: Understanding internet users on double helical model of chance-discovery process. in *Proc. of the IEEE Int'l Symposium on Intelligent Control*, pp. 844–849 (2002).