3A3-4

ローカルデータベースにおけるアイテム集合の相関の違いに基づく隠れた相関の発見

Discovery of Hidden Correlations 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
Division of Computer Science, Hokkaido University

Given a transaction database as a global set of transactions and its local database obtained by some conditioning of the global database, we consider pairs of itemsets whose degrees of correlation are higher in the local database than in the global one. A problem of finding paired itemsets with high correlation in one database is already known as Discovery of Correlation, and has been studied as the highly correlated itemsets are characteristic in the database. However, even noncharacteristic paired itemsets are also meaningful provided the degree of correlation increases significantly in the local database compared with the global one. They can be implicit and hidden evidences showing that something particular to the local database occurs, even though they were not previously realized to be characteristic. From this viewpoint, we have proposed measurement of the significance of paired itemsets by the difference of two correlations before and after the conditioning of the global database, and have defined a notion of DC pairs, whose degrees of difference of correlation are high. Since the measurement of DC pairs is nonmonotonic, DC pair mining problem is difficult. In this paper, for our difficult problem, we show DC pairs can be found efficiently by using properties of closed itemsets.

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は，相関ルール [1] や相関しているアイテム集合 (の組) [3, 6]，あるいは比較しているデータベースにおけるあるデータベースにおいて高い出現確率を持つアイテム集合 [4] のように，与えられた (いくつかの) データベースにおいて，特徴的であるアイテム集合 (の組) が注目を集めてきた．

上記の意味で特徴的なアイテム集合 (の組) は，例えば一般的な関係をとらえる際には有用である．しかし一方で，ユーザはそのような関係を当たり前であると考えるかもしれない．ここで，チャンス発見の研究 [7] のように，上記の意味で特徴的でないアイテム集合もある条件の下では潜在的に重要である．

著者らは上記のような潜在的に重要なアイテム集合を発見するための 1 つのアプローチとして，与えられたデータベースとある条件付けによるローカルデータベースにおけるアイテム集合の相関の変化に注目した．もし，その相関の変化が顕著であれば，変化後の相関が特徴的でなくても，その顕著な相関変化を何かが起こっているかもしれないエビデンスとし，潜在的に重要な関係として注目する価値があると考える．著者らは相関の度合いがローカルデータベースにおいて非常に高くなるアイテム集合の組を DC pair と呼び，上記のような DC pair の考え方と DC pair を求めるためのアルゴリズムを提案してきた [9, 10, 11].

ここで，DC pair を発見する問題は，非常に難しい問題である．なぜならば，相関の変化を評価するための関数は非単調に変化するため，単調性に基づく枝刈りの実現が難しいからである．出現確率が非常に低いアイテム集合の組でさえも顕著な相関変化を示しうると言い換えてもよい．この難しい問題に対し，著者らは既に DC pair になりえないアイテム集合の条件（枝刈り規則）とそのアイテム集合郡を同定するための方法を用い，ある程度効率的に DC pair を発見できる可能性を示した．しかし，より大規模で，より複雑な問題，例えば顕著な相関変化を引き起こす条件さえも見つける問題等に対処するためにはアルゴリズムの更なる改良が必要である．本稿では，出現確率が同じであるという性質を持つ閉包元と閉包の関係を利用し，頻出度計算の省略，計算すべきアイテム集合のまとめあげ等を行うことにより，DC pair をさらに効率的に探索できる可能性を示す．

以下において，本稿の構成を述べる．2. 章において本稿における議論のための準備を行う．3. 章において DC pair の考え方を導入し，DC pair 探索問題の定義を行う．4. 章において DC pair を探すためのアルゴリズムについて述べる．特に，飽和アイテム集合の性質の利用について議論する．5. 章において実験結果を示す．最終章において，本稿についてまとめ，今後の課題について議論する．

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

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それぞれがユニークな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．

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，$correl(X, Y) = P(X \cup Y)/P(X)P(Y)$ と定義する．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，$correl_C(X,Y) = P(X \cup Y|C)/P(X|C)P(Y|C)$ と定義する．ここで相関は $\mathcal{D}$ と $\mathcal{D}_C$ において確率が 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．

$G \subseteq G' \subseteq X, O(G) = O(G') = O(X)$ を満たす全ての G' に対し，X を**飽和アイテム集合**，G を X の**閉包元**，G' を G の**閉包**と呼ぶ．

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，”ローカル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．相関の違いを以下の比率によって評価し，本研究における目的関数とする．ここで，C はユーザによって与えられる条件を表すアイテム集合とする．

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

$\rho(>1)$ を相関の違いを表すパラメータとし，$g(X,Y;C) \geq \rho$ を満たすような X と Y の組を顕著な相関変化を示す重要な関係であると考える．ここで，(1) はベイズの定理を用いて以下のように書きなおすことができる．

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

(2) を C とそれぞれのアイテム集合の関係を意識し書きかえると，以下のように表現することができる．

$$g(X,Y;C) = \frac{correl(C, X \cup Y)}{correl(C,X)correl(C,Y)}. \quad (3)$$

(3) の分母の値が低くなればなるほど，$g(X,Y;C)$ の値は高くなる．このことは，ローカルデータベースと相関していないアイテム集合ほど，ほかのアイテム集合と組み合わせたとき，顕著な相関性変化を示す可能性が高いということを意味する．逆に言うと，ローカルデータベースと高く相関しているアイテム集合が本研究の求めるアイテム集合となる確率は非常に低い．

したがって，新たなパラメータ ϵ を導入し，$correl(C,X) < \epsilon$ を満たすような X を本研究の目的関数を最大化する可能性の高いアイテム集合として生成する．$g(X,Y;C)$ は非単調に変化するため，一般には本研究が求めるアイテム集合の組を発見することは難しいが，この制約を入れることにより，本研究の求めるアイテム集合である可能性の低い多くの探索対象を枝刈りし，探索の効率化をはかることができる．

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 ϵ が与えられたとき，**DC pair 探索問題**は $g(X,Y;C) > \rho$, $correl(C,X) < \epsilon$, $correl(C,Y) < \epsilon$ 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．
以下の議論において，$X \cup Y$ を DC pair の**組み合わせアイテム集合**，X，Y を DC pair の**要素アイテム集合**と呼ぶ．

4. アルゴリズム

本研究では，DC pair 探索問題を以下の 2 つのフェーズに分け，DC pair を導出する．

要素アイテム集合識別フェーズ
$correl(C,X) < \epsilon$ を満たすアイテム集合 X は要素アイテム集合の候補として識別される．

DC pair 導出フェーズ
要素アイテム集合識別フェーズで得られた候補は $g(X,Y;C) > \rho$ を満たすかどうか調べられる．

この節では，主にそれぞれのフェーズにおける工夫について議論する．本稿では，特に飽和アイテム集合の性質を用いた工夫について議論する．

4.1 基本的なアルゴリズム

要素アイテム集合識別フェーズにおいて，要素アイテム集合の候補を生成するため，バックトラックアルゴリズム [2, 12] を用いて探索を行う．バックトラックアルゴリズムは，再帰呼び出しを用いた反復的なアルゴリズムである．各反復は現在解を受け取り，それに各アイテムを追加したアイテム集合が満たすべき条件を満たすアイテム集合であれば，再帰呼び出しを行う．解を重複して発見することを防ぐため，各反復で現在解の末尾よりも添え字が大きいアイテムのみを追加していく．

4.2 枝刈り規則と終了条件

要素アイテム集合識別フェーズにおいては，ボトムアップに要素アイテム集合を探索する問題を考える．この探索においては，以下に示す枝刈り規則を用いることにより，効率的な探索の実現が可能である．

枝刈り規則
探索ノード (アイテム集合)X とその上位集合 $Z(X \subseteq Z' \subseteq Z)$ に対して，もし $P(C \cup Z) \geq \epsilon' \cdot P(X)$ が成り立つならば，探索において Z' は X の候補ノードにはならない．ここで，$\epsilon' = P(C) \cdot \epsilon$ である．

上記の枝刈り規則を用いるためには，サブデータベース $\mathcal{D}_C$ において X を調べるのと同時に，X の上位集合 Z も調べなければならない．そのような Z を識別するために，本研究では，max-miner アルゴリズム [2] において極大頻出アイテム集合を識別するために用いられている先読みの考え方を利用する．この考え方を用いるために，アイテムに対しある辞書順番 (本研究では経験的に頻度の昇順) を仮定し，X を現在調べているアイテム集合とする．また，$tail(X)$ をその辞書順番における X の最大アイテムとし，$T(tail(X))$ を $tail(X)$ よりも大きいアイテムの集合とする．以下の条件を満たす時，それ以降の探索を打ち切ってバックトラックする．

探索の終了条件
アイテム集合 X と $Z = X \cup T(tail(X))$ に対し，$P(C \cup Z) \geq \epsilon \cdot P(X)$ ならば，X はそれ以上調べなくてよい．

4.3 DC pair 導出フェーズにおける性質

この節では，DC pair 導出フェーズにおける性質について議論する．DC pair 導出フェーズにおいて，もし要素アイテム集合の数が多ければ，その組み合わせの数は非常に膨大である．しかし，以下の性質を利用することにより，調べるべき組み合わせを限定することができる．

要素アイテム集合の候補 X, Y を考え，$O(X, \mathcal{D}_\mathcal{C}) = \{t | t \in \mathcal{D}_\mathcal{C} \wedge X \subseteq t\}$ とする．$t \in O(X, \mathcal{D}_\mathcal{C})$ に対し，$Y \subseteq t$ なる t が存在しないならば，Y はそれ以上調べる必要はない．また，$O(X, \mathcal{D}_\mathcal{C})$ に出現するアイテムの集合 J に対し，Y が J の部分集合でないならば，Y は全く調べる必要はない．

上記の性質について，簡単に説明する．もし，X と Y が DC pair であるならば，DC pair の定義より必ず $X \cup Y$ はローカルデータベースのトランザクションに出現する．したがって，X に対し，ローカルデータベースにおける X を含むトランザクションに Y が出現しないなら，Y は X と DC pair になる要素アイテム集合ではない．ローカルデータベースにおいて要素アイテム集合の候補を含むトランザクションはそれほど多くないことが期待でき，そのようなトランザクションを調べるだけで探索を終了できる場合がある．さらに，ローカルデータベースすらスキャンせずに上記のことを判定できる場合もある．

また，要素アイテム集合識別フェーズにおいて，要素アイテム集合の候補を識別する際に，それがどのトランザクションに含まれるかの情報を得ることができる．その情報を利用することにより，DC pair 導出フェーズにおける無駄な探索を避けることができる．

4.4 飽和アイテム集合の性質の利用

この節では，本研究におけるアルゴリズムの効率化のために飽和アイテム集合の性質を利用することについて議論する．

飽和アイテム集合は，同じ支持度を持つアイテム集合郡の中で他のアイテム集合の含まれないアイテム集合である．飽和アイテム集合は，主に頻出度計算の省略 [12] 等，計算の効率化のために利用されることが多い．つまり，閉包元と閉包は同じトランザクションに含まれるので，それぞれ同じ出現確率であり，閉包の頻出度計算はする必要がない．本研究においても，要素アイテム集合生成フェーズにおける頻出度計算の省略のために，この性質を用いることができる．

ここで，本研究が特に期待したいのは，DC pair 導出フェーズにおける計算効率の向上である．DC pair 導出フェーズにおいては，求められた要素アイテム集合の候補を利用して DC pair を求める．その際に，要素アイテム集合の候補の組み合わせは，単純に考えると求められた N 個の要素アイテム集合の候補に対し，2 乗のオーダで計算負担がかかる．この負担に対し，閉包の関係にある要素アイテム集合は同じトランザクションに含まれるので，そのような要素アイテム集合の候補をまとめて扱い DC pair を求めることにより，計算負担は軽減する．DC pair となった閉包元の組のそれぞれの閉包同士が重複することもありえるが，その場合は，出力時に除外することにする．本研究では，DC pair を求める際に，閉包元を基に DC pair を求め，DC pair が見つかった際にそれぞれの要素アイテム集合の閉包の重複しない組み合わせを出力することにより，DC pair を求める．

5. 実験

この節では，本研究で行った実験の結果を示す．実験の目的は，飽和アイテム集合の性質を利用することにより，DC pair の効率的な探索が実現できるか確かめることである．

5.1 データセットと実装

本研究では，*Entree Chicago Recommendation Data* とアメリカの国勢調査のデータ (IPUMS)[8] を用いて実験を行った．

Entree Chicago Recommendation Data は UCI KDD Archive [5] におけるデータセットの 1 つであり，それぞれが Atlanta や Boston などにおけるレストランの特徴を含む 8 つのデータセットから成る．本研究では，8 つのデータセットをグローバルデータベース $\mathcal{D}$ として 1 つのデータベースに合成した．そして，それぞれの地域 C による条件付けによって，$\mathcal{D}$ におけるローカルデータベース $\mathcal{D}_\mathcal{C}$ を定義する．$\mathcal{D}$ はそれぞれのアイテムがレストランの特徴を表す 265 アイテムの部分集合である 4160 トランザクションから成る．

IPUMS データはアメリカの国勢調査と 2000 年から 2003 年における”American Community Surveys”という調査からの 37 の信頼性の高いサンプルから成る．IPUMS data extraction system を使用することにより，必要なサンプルと変数を選択しデータを抜き出すことができる．本研究では，Washington における 1980 年，1990 年，2000 年のサンプルを抽出した．変数は，年齢，性別，人種等である．グローバルデータベースは 1980 年と 1990 年，1990 年と 2000 年のサンプルを合成したそれぞれ 90031 トランザクションと 107697 トランザクションからなり，条件はそれぞれ 1990 年と 2000 年として DC pair を導出した．それぞれのトランザクションは選択した変数のそれぞれの値をアイテムとした 131 アイテムの部分集合である．

本研究のシステムは C 言語で実装され，全ての実験は 1.00 GB RAM，Xeon 3.60 GHz プロセッサのスペックを持つ PC 上で行った．

5.2 飽和アイテム集合の性質の効果

図 1,2 に IPUMS データを用いた際の実験結果を示す．*Entree Chicago Recommendation Data* を用いた実験結果については，紙面の都合により割愛し，議論のみ行う．

要素アイテム集合識別フェーズ

要素アイテム集合識別フェーズにおいては，飽和アイテム集合の性質を利用し，頻出度計算を省略することにより，計算効率が上昇するかを確かめた．頻出度計算を省略しただけで，本研究の枝刈りを利用しなくても，十分効率的な探索を実現できる可能性もあるので，本研究の枝刈りのみ行う今までの探索，本研究の枝刈りに加えて頻出度計算を省略する本実験の探索に付け加えて，頻出度計算の省略のみを行う探索に対する探索対象数と実行時間を調べた．

図 1 より，今までの探索（枝刈りのみ）と本実験の探索（枝刈り＆飽和）を比較することにより，約 2 分の 1 の探索対象数と実行時間で要素アイテム集合の生成が実現できていることがわかる．さらに，頻出度計算の省略のみを利用した探索との比較により，頻出度計算の省略のみが探索の効率化に寄与しているわけではなく，本研究で提案している枝刈りが探索の効率化に十分に寄与していることもわかる．*Entree Chicago Recommendation Data* においても全ての条件において，同様の結果が得られた．この結果により，要素アイテム集合識別フェーズにおいて飽和アイテム集合の性質を利用することにより，本研究の枝刈り規則によって探索を省略できない部分の頻出度計算を省略するという形で，探索の効率化が実現できることが示された．

DC pair 導出フェーズ

図 2 より，飽和アイテム集合の性質を利用することにより，DC pair 導出フェーズにおいて 4 分の 1 以下の探索対象数と実行時間で DC pair を導出できていることがわかる．しかし，*Entree Chicago Recommendation Data* においては確実に探索の効率は上昇したが，IPUMS データほどの探索の効率化は実現できなかった．その原因は，IPUMS データでは，1980–1990 census，1990–2000 census において見つかったそれぞれ 16854 個，6796 個の要素の候補に対する 6319 個，2281 個の

$\rho = 2.0, \epsilon = 0.35$	1980–1990 census		1990–2000 census	
	探索対象数	時間 (sec)	探索対象数	時間 (sec)
枝刈りのみ	1229553	350.532	1347128	402.922
飽和のみ	1093019	267.469	1099774	299.063
枝刈り＆飽和	658653	197.797	734137	234.875
要素の候補	29106		8244	

図 1: 要素アイテム集合識別フェーズにおける飽和アイテム集合の性質の利用の効果

	1980–1990 census		1990–2000 census	
	探索対象数	時間 (sec)	探索対象数	時間 (sec)
飽和利用なし	985158	49.547	323825	13.657
飽和利用あり	383237	10.781	77063	3.328
DC pair	16854		6796	

図 2: DC pair 探索フェーズにおける飽和アイテム集合の性質の利用の効果

閉包元の組み合わせを調べればよかったのに対して，*Entree Chicago Recommendation Data* では，見つかった要素アイテム集合とその閉包元に大きな差が見られなかったため，調べるべき組み合わせもそれほど減らせなかったことが挙げられる．以上の結果より，DC pair 導出フェーズにおいては，見つかった要素の候補がある程度少ない数の閉包元によって表現できている際には，DC pair の効率的な導出が可能であることが示された．

6. まとめと今後の課題

本稿では，潜在的に重要なアイテム集合の組を発見するための１つの考え方として，我々が既に提案している DC pair を発見するためのアルゴリズムの改良について議論した．我々は現在，さらに複雑な問題を扱うための準備を進めており，本稿における実験の結果は，その有効なエビデンスとなりえる．今後は，本稿における実験によって明らかになってきた DC pair の性質を利用し，さらに効率的に DC pair を導出するシステムを目指すことが課題となる．

参考文献

[1] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, In: J. B. Bocca, M. Jarke and C. Zaniolo (Eds.), the 20th Int'l Conf. on Very Large Data Bases, Morgan Kaufmann, VLDB'94, pp. 487–499, 1994.

[2] R. J. Bayardo Jr., Efficiently Mining Long Patterns from Databases. In: L. M. Haas and A. Tiwary (Eds.), the ACM-SIGMOD Int'l Conf. on Management of Data, ACM Press, pp. 85–93, 1998.

[3] S. Brin, R. Motwani and C. Silverstein, Beyond Market Baskets: Generalizing Association Rules to Correlations. In: J. Peckham (Ed.), the ACM SIGMOD Int'l Conf. on Management of Data, ACM Press, vol. 26, no. 2, pp. 265–276, 1997.

[4] G. Dong and J. Li, Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In: the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, ACM, pp. 43–52, 1999.

[5] S. Hettich, and S. D. Bay, The UCI KDD Archive, Department of Information and Computer Science, University of California, Irvine, CA, http://kdd.ics.uci.edu, 1999.

[6] S. Morishita and J. Sese, Traversing Itemset Lattices with Statistical Metric Pruning. In: the ACM SIGACT-SIGMOD-SIGART Symposium on Database Systems, ACM, PODS 2000, pp. 226–236, 2000.

[7] Y. Ohsawa and Y. Nara, Understanding Internet Users on Double Helical Model of Chance-Discovery Process. In: the IEEE Int'l Symposium on Intelligent Control, IEEE, pp. 844–849, 2002.

[8] S. Ruggles, M. Sobek, T. Alexander, C. A. Fitch, R. Goeken, P. K. Hall, M. King and C. Ronnander, Integrated Public Use Microdata Series: Version 3.0 [Machine-readable database]. Minneapolis, MN: Minnesota Population Center [producer and distributor], 2004.

[9] T. Taniguchi, M. Haraguchi and Y. Okubo, Discovery of Hidden Correlations in a Local Transaction Database based on Differences of Correlations, In: P. Perner and A. Imiya (Eds.), Machine Learning and Data Mining in Pattern Recognition, Springer Verlag, lnai 3587, MLDM 2005, pp. 537–548, 2005.

[10] T. Taniguchi and M. Haraguchi, An Algorithm for Mining Implicit Itemset Pairs based on Differences of Correlations. In: A. Hoffmann, H. Motoda and T. Scheffer (Eds.), the 8th Int'l Conf. on Discovery Science, Springer Verlag, lnai 3735, DS 2005, pp. 227–240, 2005.

[11] T. Taniguchi and M. Haraguchi, Discovery of Hidden Correlations in a Local Transaction Database based on Differences of Correlations, P. Perner (Issue Ed.) " Advances in Data Mining ", In: Engineering Applications of Artificial Intelligence, ISSN 0952-1976, Elsevier, (to appear).

[12] T. Uno, M. Kiyomi and H. Arimura, LCM ver.2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets. In: R. J. Bayardo Jr., B. Goethals and M. J. Zaki (Eds.), the IEEE International Conference on data mining, 2nd Workshop on Frequent Itemset Mining Implementations (FIMI'04), CEUR-WS.org, CEUR Workshop Proceedings, vol. 126, 2004.