様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１２　日現在

機関番号：　１３５０１
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2016
課題番号：　２５３３０２５６

研究課題名（和文）潜在的相関ルールマイニングと高次イベント系列コーパスの自動構築

研究課題名（英文）Efficient Mining Methods for Latent Association Rules and their Application for Generating Latent Event Sequence Corpora

研究代表者

岩沼　宏治（IWANUMA, Koji）

山梨大学・総合研究部・教授

研究者番号：３０１７６５５７

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：本研究では，潜在的相関ルールとして負の相関ルールに着目し，マイニングアルゴリズムの高速化と高度化に関する研究を行った．またイベント時系列コーパスの半自動合成へ応用すための基礎的研究を行った．具体的には，負の相関ルールを高速マイニングするために，上昇型と下降型計算を接尾辞木上で融合した新しい計算法を開発した．また膨大な数となる負ルールを圧縮するために，極小生成子を用いた無損失な圧縮手法を新しく開発した．更に大規模テキスト時系列データから潜在因子を考慮したイベント時系列を生成するために必要な基盤として，多重データストームから頻出アイテム集合や飽和集合をオンライン型近似計算で抽出する手法を開発した．

研究成果の概要（英文）：In this research, we developed several efficient algorithms for negative association rule mining, which can be regraded as a concrete form of a latent association mining. We also studied some online approximation algorithms for a huge transaction stream, which is an essential tool for generating latent event sequence corpora from a very large sequential text data such as newspaper data. The details are as follows: First, we proposed a new efficient top-down search algorithm for valid negative association rules, which uses a suffix tree over frequent itemsets. Second, we studied lossless compression for a set of negative rules, and gave a novel lossless compression method based on minimal generators. Third, we developed two online approximation algorithms for mining a huge transaction stream: one achieves a resource-oriented computation, and another uses an incremental intersection computation for frequent closed itemsets, both of which can avoid the combinatorial explosion phenomena.

研究分野：　人工知能基礎

キーワード：　データマイニング　負の相関ルール　極小生成子　飽和アイテム集合　オンライン型アルゴリズム　データストリーム　潜在因子　無損失圧縮

１．研究開始当初の背景

データマイニングは Agrawal らによる相関ルール抽出法 Apriori の成功以降，多くの研究が行われてきた．マイニングの対象となるデータも，データ系列，木とグラフ，連続数値データなど多種多様な形式のものが研究されている．我々は長大な単一のデータ系列に着目し，新しく系列全体頻度を提案し，頻出な部分系列を高速に抽出法について研究を行ってきた，新聞記事コーパスに適用し，意味のあるイベント時系列の自動抽出に試験的に成功している．

しかしながら，これらの全てのデータマイニング技術は，陽に出現する事象の関係，即ち，頻出な事象の間の正の共起関係を，相関ルールの形で抽出する技術である．出現しないデータ，即ち隠れ因子に関する関係，即ち潜在的共起（反共起）関係を表現する相関ルールのマイニングについては殆ど研究が行われていない．一般に，潜在変数とそのパラメータの推定問題は，機械学習において極めて重要な課題であり，非常に多くの研究がなされている．大規模データ中の隠れ因子・事象間の共起規則，即ち潜在的相関ルールのマイニングも極めて重要な問題であり，応用範囲も広いと考えられるが，これまで殆ど研究されていなかった．我々の知る限りにおいて，負の相関ルールマイニングに関して幾つかの研究が存在する程度である．負の相関ルールとは

$$\neg X \Rightarrow Y \quad \text{（左否定形）または}$$
$$Y \Rightarrow \neg X \quad \text{（右否定形）}$$

なる形の表現である．$\neg X$ は負のアイテム集合と呼ばれ，X が出現しないことを表現するものであり，隠れ因子と見なせる．例えば，左否定形のルール $\neg X \Rightarrow Y$ は，「X が出現しない場合に Y が出現する場合が多い」という反共起関係を表す表現となる．

負の相関ルールマイニングでは，「頻出でない（非頻出）」アイテム集合を取り扱う必要があるために，マイニングに必要となる計算量は膨大であり，効率化が難しいことが知られている．既存の研究は，正の共起関係を高速マイニングするための Apriori 法などを，反共起関係を表す負ルールのマイニングに直接的に適用したものが殆どである．負ルールの特性を積極的に利用したものはなく，負ルールの効果的なマイニング手法は開発されていなかった．

２．研究の目的

本研究の目的は，新しい相関ルールマイニングとしての潜在的相関ルールマイニングの提案，特に負の相関ルールマイニングの高速化と高度化に関する研究を行う．またイベント時系列コーパスの半自動合成に応用するための基礎的研究を行う．より具体的には以下通りである．

(1) 負の相関ルールを抽出するときに，正ルールの抽出に倣ってルールの台集合を生成することから始めると，非頻出なアイテム集合を網羅的に生成する必要が生じるために，莫大な計算が必要となる．本研究では，この問題を避けるために，頻出集合 X と Y を組み合わせて妥当な負ルール $\neg X \Rightarrow Y$ と $Y \Rightarrow \neg X$ を生成することにより，負ルールを効果的に抽出枚挙する手法を新しく開発する．

(2) 潜在的相関ルールを抽出するためには，負のアイテム集合に代表される潜在因子の抽出が本質的な課題となる．しかし，頻度尺度で明確に決定できる非頻出な集合（負のアイテム集合）以外のものは，その抽出は本質的に極めて難しい．その解決には，仮説推論の導入と融合が効果的と考えられ，それに必要な技術を開発する．

(3) 本研究で構築をめざすイベント時系列コーパスは，種々の知的情報処理において極めて重要な役割を果たすと期待できる．例えば，「地震」の後に「火事」が起こり，その後に「災害対策本部設立」などのイベントが連続的に生じることを，従来の言語資源などを用いて予測同定することは不可能である．大規模なテキスト時系列データから抽出することが必要になるが，潜在因子まで考慮したイベント系列コーパスを効果的に構築することは簡単ではない．長大なデータストリーム上でオンライン型計算により負ルールのマイニングを行うことが望ましく，それに必要な技術を開発する．

３．研究の方法

本研究では，主に以下の 3 つの研究を行っていく．

(1) まず，負の相関ルール抽出のためのアルゴリズムの高速化について研究を行う．これまでの負ルールの抽出技術は Apriori 法の直接的な拡張となっており，まず膨大な非頻出なアイテム集合を生成し，それらを適宜分割して妥当な負ルールを抽出しており，極めて効率が悪い．本研究では，頻出集合 X と Y の組み合わせを，接尾辞木等に基づいた深さ優先探索で効率的に検査し，妥当な負ルールを高速に生成する手法を開発する．

(2) 負ルールは正ルールと比較しても，その数が極めて多い．そのため，個々の負ルールの生成を高速にしても，負ルール全体の生成には多くの時間がかかるのが通常である．これを改善するためには，負ルール全体の集合を圧縮する手法が非常に重要な研究課題となる．正ルール集合の圧縮には飽和集合の技

術が有効であることが知られているが，負ルール集合への適用する妥当性は不明である．その適用可能性や代替え手法等について研究を行う．

また潜在的相関ルールを抽出するためには，負のアイテム集合に代表される潜在因子の抽出が本質的な課題となるが，その効果的な計算は一般には非常に難しい．解決に向けては，仮説推論の導入が有効と考えられるが，仮説推論には領域知識の事前抽出が必要となる．対象データの領域知識は論理型ルールで表現する必要があるが，これはデータから確信度 100%の正負の相関ルールを抽出することに相当する．このため本研究では，確信度 100%の相関ルールを事前に高速に列挙抽出する手法を開発することも目標としている．

(4) 実用に供することができるイベント系列コーパスを構築するためには，巨大なテキスト時系列データから意味のある単語集合の系列を抜き出す必要がある．よって，アイテム集合のストリームデータから頻出な部分集合の部分系列を抜き出す技術を開発する必要がある．アイテム集合の系列を取り扱う場合には，極めて厳しい組み合わせ爆発現象を克服しなければならない．この問題を解決するために，本研究では，多重データのストリームでのオンライン型近似計算に基づくマイニング技術を開発する．

４．研究成果

(1) まず，負の相関ルールの抽出を行う高速アルゴリズムを開発した．正の相関ルールを抽出する第 1 世代の技術は Apriori 法に代表される上昇型計算法であるが，これは現在の技術水準でみればかなり効率が悪い．第 2 世代の技術としては，データ射影を利用した分割統治計算を行う下降型計算法で著名である．残念ながら，負ルールの抽出ではデータ射影が原理的に行えないことから，分割統治法の適用も困難である．そこで本研究では，両者を融合した新しい計算法を開発した．即ち，まず上昇型計算により頻出アイテム集合を全て抽出する．次に抽出した頻出集合の組み合わせを検査し，妥当な負ルールを選出する．検査と選出は，接尾辞木を用いた下降型探索により高速計算を実現した．接尾辞木の上の下降型探索は，右否定型ルール X ⇒ ¬ Y の右極小性の判定が極めて高速に実行でき，探索木の効果的な枝刈が実現できる．左否定型 ¬ X ⇒ Y に対しては，負ルールの確信度関数が逆単調性を満たさないために，その近似関数として，逆単調性をみたす上界関数を考え，効果的な探索木の枝刈を実現している．実証システムを試作して性能評価実験を行い，良好な結果を確認している[論文③] ．

(2) 負ルール集合の圧縮に関する研究成果は以下の通りである．まず，正のアイテム集合の無損失圧縮に用いられた飽和集合の技術が，負ルール集合の圧縮には本質的に不十分である，即ち，無損失圧縮を保証することが原理的にできないことを明らかにした．そこで本研究では，飽和集合に代わるものとして，極小生成子を用いた負ルール集合の圧縮法を，新しく開発を行った．提案した極小生成子に基づく圧縮法の完全性（無損失性）を理論的に証明し，更に実証実験を通して，圧縮率等に関する有用性を検証した．実験の結果，疎なデータに内在する負ルール集合の圧縮には効果が確認できなかったが，密なデータ中の妥当な負ルールの集合の圧縮においては，100 分の 1 以下に圧縮できるなど，大きな圧縮効果を持つことが確認できた[論文①]．この圧縮技術の高速計算を実現するためには，極小生成子を効果的に生成する必要があるが，これについては，飽和集合から極小生成子を生成する手法について研究を行っている．膨大な数に及ぶ頻出アイテム集合の生成を一切行わないことから，極小生成子の生成計算の効率化，ひいては負ルールの生成計算の効率化につながる手法である．確信度 100%の論理型ルールも，極小生成子を用いることによって生成が容易になることが明らかにしている．

(3) イベント系列コーパスを抽出する大規模テキスト時系列データは，個々のテキストを単語が多重出現する集合と考えれば，長大な多重データストリームとみなすことができる．よって，多重データストリームから負ルールの集合を抽出する必要がある．極めて厳しい組み合わせ爆発現象を克服しなければならないために，本研究では，多重データのストリームでのオンライン型近似計算に基づくマイニング技術を開発した．

個々の負ルールは頻出アイテム集合を組み合わせて構築するので，まず初めに，頻出アイテム集合を多重データストリームからオンラインで抽出する必要があるが，これまでに効果的な計算手法は知られていなかった．そこで本研究では，オンライン型近似計算の枠組みを用いて，計算時間やメモリなどのリソースの状況に基づいて，マイニング計算を適応的に制御する新しい計算機構を開発した．より具体的には，リソースの使用状況に応じて，アイテム出現頻度の許容誤差を動的に制御し，あわせてデータ処理を適応的にスキップする機構を開発し，頻出アイテム集合を抽出する新しいオンライン型近似アルゴリズムを提案した．その有効性について理論的な性能保証を行い，実証実験を通してその有用性を確認している．この研究成果はデータ工学の分野で世界最高レベルの国際会議 ACM-SIGMOD 14 において regular paper として採録され，発表を行っている[学会発表⑤].

上記の手法はストリーム中の頻出アイテム集合に着目して処理を行っているが，これを飽和アイテム集合に変更すれば，頻出集合の候補を削減でき，より効率的な抽出計算が行える．そこで本研究では，前述の研究を発展させ，頻出飽和アイテム集合を多重データストリームからオンライン抽出を可能とする，漸近的集合積計算とε近似を用いた近似計算法を開発した．これまでに提案された飽和アイテム集合のオンライン型近似計算法の全ては，理論的保証の無いヒューリスティック算法であった．本研究では，提案手法が，頻出アイテム集合の抽出に関する完全性を持つことと，抽出したアイテム集合の出現頻度の相対誤差がε以下であることを，世界で初めて理論的に証明し，保証を与えることに成功している[論文②]．この研究成果をデータ工学の欧州でのトップ会議であるEDBT2016に発表[学会発表①]している．またその予備的な研究発表は，2014 年度人工知能学会研究会優秀賞を受賞している[学会発表④]

⑷ 以上の結果を踏まえて，本研究では，ストリームデータからのオンラインでの負ルールの生成を試みている．オンライン近似計算で抽出した飽和アイテム集合から極小生成子を順次生成し，その組み合わせを検査して，妥当な負ルールの集合を準オンライン計算で抽出する手法を開発した．本研究事項でまず問題となるのは，基礎となる飽和アイテム集合の頻度に誤差が存在することである．このため出現頻度に基づくルールの評価値にも本質的に誤差が混入してしまうことである．本研究では，提案手法で算出した頻度の誤差には一定の保証を与えていたことに着目し，負ルールの評価尺度の誤差にも一定の保証を与えることができることを新たに示した，負ルールの集合の計算は，極小生成子の差に基づく漸近的差分を計算して，順次更新していくことで，計算時間を軽減できることを示した［学会発表②］．この漸近的計算においては，差分を計算する時間区間の設定が非常に重要である．設定した時間区間があまりに短いと，オーバーヘッドが大きくなり，効果がでない．逆に，長いと差分計算が無意味になる．適切な区間長の設定は難しい課題であり，現在も実験的評価によって研究を継続しておこなっている．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

① 岩沼宏治，佐生隼一，黒岩健歩，山本泰生：負の相関ルール集合の極小生成子に基づく圧縮表現，情報処理学会論文誌．査読有，Vol.57, No.8, 2016, pp.1845－1849

② 岩沼宏治，山本泰生，福田翔士： ストリーム中の頻出飽和集合を抽出するオンライン型ε-近似アルゴリズムの完全性．人工知能学会論文誌，査読有，Vol.31, No.5, 2016, p.B-G52_1-10 http://doi.org/10.1527/tjsai.B-G52

③ 井出典子，岩沼宏治，山本泰生：負の相関ルールを抽出する高速トップダウン型アルゴリズム．人工知能学会論文誌，査読有, Vol.29, 2015, pp.406－415 http://doi.org/10.1527/tjsai.29.406

〔学会発表〕（計 21 件）

① Koji Iwanuma, Yoshitaka Yamamoto and Shoshi Fukuda: An On-Line Approximation Algorithm for Mining Frequent Closed Itemsets Based on Incremental Intersection. *Proc. of Intl. Conf. on 19th Extended Database Technology (EDBT2016)*, Bordeaux (France), 2016, pp.704-705.

② 黒岩健歩，岩沼宏治，山本泰生：負相関ルールを抽出する準オンラインアルゴリズム．人工知能学会第 100 回人工知能基本問題研究会資料 SIG-FPAI-B503-01, pp.1-6, 2016．熊本市民会館（熊本県・熊本市）

③ Yoshitaka Yamamoto and Koji Iwanuma: Online Pattern Mining for High-Dimensional Data Streams. *Proc. of IEEE BigData 2015*, Santa Clara (USA), 2015, pp.2615-2617.

④ 福田翔士，岩沼宏治，山本泰生: トランザクションストリーム上のオンライン型頻出飽和集合マイニング．人工知能学会第 97 回人工知能基本問題研究会, pp.1-6, 2015. 別府ビーコンプラザ（大分県・別府市）

⑤ Yoshitaka Yamamoto, Koji Iwanuma and Shoshi Fukuda: Resource-oriented Approximation for Frequent Itemset Mining from Bursty Data Streams. *Proc. of the 2014 ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD'14)*, Utah (USA), 2014, pp. 205-216, doi>10.1145/2588555.2612171

⑥ Yoshitaka Yamamoto, Adrien Rougny, Hidetomo Nabeshima, Katsumi Inoue, Hisao Moriya, Christine Froidevaux and Koji Iwanuma. Completing SBGN-AF Networks by Logic-Based

Hypothesis Finding. *Proc. of the 1st Intl. Conf. on Formal Methods in Macro-Biology (FMMB2014)*, *Lect. Notes in Bioinformatics*, Vol.8738, 2014 , pp.165-179. Noume'a (New Caledonia).

〔その他〕
ホームページ等：潜在的相関ルール
http://www.kki.yamanashi.ac.jp/~iwanuma/Kaken2013

６．研究組織

(1)研究代表者
岩沼　宏冶（IWANUMA, Koji）
山梨大学・総合研究部・教授
研究者番号：30176557

(2)研究分担者
山本泰生（YAMAMOTO, Yoshitaka）
山梨大学・総合研究部・助教
研究者番号： 30550793