様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　２４　年　５　月　２５　日現在

機関番号：10101
研究種目：基盤研究（Ｂ）
研究期間：　1//8　～　1/00
課題番号：21300047
研究課題名（和文）　構造変化マイニング

研究課題名（英文）　Mining Structural Changes

研究代表者 原口 誠　(HARAGUCHI MAKOTO)
北海道大学・大学院情報科学研究科・教授
研究者番号：40128450

研究成果の概要（和文）：
変数間の関係変化をマイニングの検出目標とする．具体的には，正および負の相関，さらには，偏相関も扱いうるカルバック・ライブラー情報量が，文脈変化の前後で増加する変数パターンを検出する．情報量の算出に要する集合分割の計算コストに対処するために，変数を頂点にもつ離散グラフにおけるクリーク制約の活用，非相関から相関への変化をより高速に検出できる最適化手法の実現（jumping emerging correlation change）などを活用し，高速な相関変化検出を達成した．

研究成果の概要（英文）：
We have targeted patterns of variables whose correlations get increased after an event, while the correlations are uncorrelated before the event.　To take into account positive, negative and even partial correlations among variables, we adopt Kullback-Leibler divergence for two contexts, before and after the event, and take their difference as the measure of changes.　To reduce the computational cost for set partitioning, we implemented two strategies, Double Clique Constraints and Jumping Emerging Correlation Change, and showed that the correlation change problem can be solved efficiently even for large data sets.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
21 年度	1,400,000	420,000	1,820,000
22 年度	2,200,000	660,000	2,860,000
23 年度	1,400,000	420,000	1,820,000
年度			
年度			
総　計	5,000,000	1,500,000	6,500,000

研究分野：　総合領域
科研費の分科・細目：　情報学・知能情報学 1005
キーワード：変化検出問題，情報量差分，制約マイニング

１．研究開始当初の背景

トランザクションデータや多次元クロス表において，興味深い変数（アイテム）集合を求める研究はデータマイニングにおいて主要なテーマの一つである．一方，変数間の関係は時間やトピック等の文脈的要因によって変化し，そうした変化は変数間の構造的関係変化として具象化される．顕著な構造的変化であっても，単一の文脈において顕著なパターンであるとは限らず，このことが変化

検出問題を解決する際の一つの大きな難点となっていた．単純な頻度変化においてこの問題を試みた研究もあるが，構造変化を考える際に重要な相関変化をも扱いうる形で変化検出手法を新たに開発する必要性があった．

２．研究の目的

重要な変数間の関係は，変数依存性を現す相関ルールや構造方程式の形で明示的に示されている場合も多い．文脈変化の前後において，変数間の関係に無視できない変化が生じる場合でも，変化の前後において変数依存性・変数間相関が変わらないものも多数存在する．変化検出の立場からは，そうした不変なものはむしろ除外し，変化の兆しを持つものだけを効率的に検出することがポイントとなる．ここで，変化前において依存性・相関性がないものを事前に枚挙し，その中から変化後に依存度・相関度が向上するものを拾いあげる方式もないわけではないが，組合せ数において依存性のない変数の組合せは依存性のある変数のそれよりも爆発的に多いという経験則を指摘しなければならない．すなわち，そうした素朴な探索戦略ではなく，変化しないものを枝刈しながら変化するものだけを抽出する手法の提案が技術的な目標の核となる．

３．研究の方法

(1) 対象となるデータセット

テキストマイニング等において良く使われているトランザクションデータを考える．具体的には，社会的に一定の注目を浴びたイベントの前後における新聞記事を，イベント前およびイベント後のデータとして収集する．各データは，記事 - 単語行列の形式のデータに前処理しておく．

(2) 変数間の関係変化を計測する尺度を定める．統計的相関も一つの考え方だが，ここでは，データベースのサイズの影響を受けにくい情報量（ダイバージェンス）を用いる．情報量の比もしくは差として変化量を計測するアプローチは，非線形・非凸関数に対する多数の局所最適解が存在する最適化問題となる．この難点を避けるために，制約マイニングの考えを採用し，制約を満たす解（パターン）を列挙する手法を開発する．

(3) 手法の評価は，計算時間と解の品質の両面から評価を行う．前者に関しては，制約と最適化による枝刈効果を，後者に関しては，伝統的な変化尺度との対比も行い，今回提案手法によって初めてマイニングされるものの存在を示す．

４．研究成果

(1) まず，変数パターンのマイニングにおいて，情報量の大きなものを求める手法は，サブスペース法の一種として２０００年前後から既に研究されている．しかるに，情報量の計算のために，集合分割（細分）を繰り返すことから，多大なスペースを一般に必要とする．その問題もあり，主流の方法論としては進化していない．本研究は，文脈前後の２つのデータベースの各々に対し，情報量の計算を結果的に要求しており，素朴な戦略を使う限りにおいて，効率の良いアルゴリズムの開発は困難であることをまず指摘しておく．また，頻度等の計算負荷が軽い評価尺度を用いる場合においても，「ボーダーアルゴリズム」という，極小もしくは極大パターン集合を事前に求める方式が最も標準的な手法として用いられているが，これも極小パターン数が膨大であることから，少なくとも本研究が採用すべき手法としては現実的ではない．

(2) これに対し本研究では，情報量がクラス変化の前後で増大する条件を，変数を頂点にもつ離散グラフにおけるクリーク制約として翻訳し，クラス変化の前後で相関が変わらない変数パターンを枝刈する方式を実現した．クリーク制約を満たす変化パターンは，情報量に関する制約を必ず満たすわけではないが，後者に対する必要条件になっており，可能な解を全て枚挙できるという完全性を保持し，かつ，不要な組合せを大幅にカットする探索を実現している．

(3) さらに，単に上記の要請を満たす変数群をマイニングするのではなく，最適化条件を満たす変数群のみを検出する新たな方式を策定・実装・実験した．具体に述べると，文脈変化前においては，ほとんど無相関な変数群で，文脈変化後にある一定の情報量を持つ変数群のみを選択的に抽出する研究である．クラス変化前の無相関性制約を使うのではなく，その最小化により結果的に無相関制約を達成することを意図している．計算論的には，最小化基準に従い，最適化に寄与しない変数パターンを探索時に枝刈できるので，制約の組み合わせのみのマイニングと比較し，さらに高速化できることを検証済みである．

(4) (2) および (3) のいずれの方式の場合も，頻度変化においては顕現的でない変化を，３-(1) のデータに対して検出可能なことを確認した．これは，変化後において顕現的でないが相関変化においては重要な変数関係が実際に生じることを示しており，本研究によって初めて実証された知見である．

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 12 件）

1. A. Li, M. Haraguchi and Y. Okubo, Contrasting Correlations by an Efficient Double-Clique Condition, Transactions on Machine Learning and Data Mining, 5(1), pp. 3 - 22, ibai Publishing, 2012, 査読有.

2. A. Li, M. Haraguchi and Y. Okubo, Top-N Minimization Approach for Indicative Correlation Change Mining, Proc. of the 8th Int'l Conf. on Machine Learning and Data Mining in Pattern Recognition - MLDM'12, Springer-LNAI, 2012 (Accepted), 査読有.

3. A. Li, M. Haraguchi and Y. Okubo, Contrasting Correlations by an Efficient Double-Clique Condition, Proc. of the 7th Int'l Conf. on Machine Learning and Data Mining in Pattern Recognition - MLDM'11, Springer-LNAI 6871, pp. 469 - 483, 2011, 査読有.

4. Y. Okubo and M. Haraguchi, An Algorithm for Finding Indicative Concepts Connecting Larger Concepts Based on Structural Constraints, Contributions to ICFCA 2011, The 9th Int'l Conf. on Formal Concept Analysis - ICFCA'11, pp. 53 - 68, 2011, 査読有.

5. Y. Okubo and M. Haraguchi, An Algorithm for Finding Conceptual Document Clusters Based on Top-N Formal Concept Search: Pruning Mechanism and Empirical Effectiveness, Search Algorithms and Applications, Nashat Mansour (ed.), pp. 385 - 408, InTech, 2011, 査読有.

6. E. Tomita, T. Akutsu and T. Matsunaga, Biomedical Engineering, Trends in Electronics, Communications and Software, Anthony N. Laskovski (ed.), pp. 625 - 640, InTech, 2011, 査読有.

7. M. Haraguchi and Y. Okubo, Pinpoint Clustering of Web Pages and Mining Implicit Crossover Concepts, Web Intelligence and Intelligent Agents, Zeeshan-ul-hassan Usmani (ed.), pp. 391 - 410, InTech, 2010, 査読有.

8. Y. Okubo, M. Haraguchi and T. Nakajima, Finding Rare Patterns with Weak Correlation Constraint, Proc. of the 2010 IEEE Int'l Conf. on Data Mining Workshops - ICDMW'10, pp. 822 - 829, 2010, 査読有.

9. 中西 裕陽, 富田 悦次, 最大クリーク問題の多項式時間的可解性の一結果, 電子情報通信学会論文誌 D, J93-D, 417 - 425, 2010, 査読有.

10. Y. Okubo and M. Haraguchi, An Algorithm for Extracting Rare Concepts with Concise Intents, Proc. of the 8th Int'l Conf. on Formal Concept Analysis - ICFCA'10, Springer-LNAI 5986, pp. 145 - 160, 2010, 査読有.

11. Y. Okubo and M. Haraguchi, Finding Top-N Pseudo Formal Concepts with Core Intents, Proc. of the 6th Int'l Conf. on Machine Learning and Data Mining in Pattern Recognition - MLDM'09, Springer-LNAI 5632, pp. 479 - 493, 2009, 査読有.

12. E. Tomita, Y. Sutani, T. Higashi, S. Takahashi and M. Wakatsuki, A Simple and Faster Branch-and-Bound Algorithm for Finding a Maximum Clique, Proc. of International Workshop on Algorithms and Computation, Springer-LNCS 5942, pp. 191 - 203, 2010, 査読有.

〔学会発表〕（計 13 件）

1. エラウィンディ サラ・原口 誠・大久保 好章・富田 悦次, クリーク全列挙に基づく構造変化検出アルゴリズム, 情報処理学会 数理モデル化と問題解決研究会, 2012 年 3 月 2 日, 指宿市民会館(鹿児島県指宿市).

2. 原口 誠・大久保 好章・富田 悦次・吉岡真治, 変化検出のための極大整合連結集合, 人工知能学会 基本問題研究会, 2011 年 12 月 15 日, 慶応義塾大学(神奈川県横浜市).

3. 鶴田 哲章 ・原口 誠, モデュラリティの差異に基づくコントラスト法, 人工知

能学会全国大会(第 25 回), 2011 年 6 月 2 日, アイーナ いわて(岩手県盛岡市).

4. 大久保 好章・原口 誠, 接続概念間の構造制約に基づくレア概念抽出, 情報処理学会 数理モデル化と問題解決研究会, 2011 年 3 月 7 日, 青島パームビーチホテル(宮崎県宮崎市).

5. A. Li and M. Haraguchi, Contrasting Correlations by an Efficient Double-Clique Search Method, Poster Abstract of the 2nd Asian Conf. on Machine Learning - ACML'10, 2010 年 11 月 10 日, 東京工業大学(東京都目黒区).

6. A. Li and M. Haraguchi, Contrasting Correlations Based on Double-Clique Search, 第9回情報科学技術フォーラム (FIT2010), 2010 年 9 月 7 日, 九州大学(福岡県福岡市).

7. E. Tomita, PLENARY LECTURE, The Maximum Clique Problem (招待講演), The 14th WSEAS International Conference on Computers, 2010 年 7 月 23 日, Corfu Island, Greece.

8. 中島 健志・原口 誠・大久保 好章, 萌芽的概念抽出のための局所分枝限定探索を用いた概念プール掘削法, 人工知能学会全国大会(第 24 回), 2010 年 6 月 9 日, 長崎ブリックホール(長崎県長崎市).

9. S. Muwazi Simona and M. Haraguchi, Extracting Approximate Biclusters/Patterns from Time Series Medical Data Using Suffix Trees, 人工知能学会全国大会(第 24 回), 2010 年 6 月 9 日, 長崎ブリックホール(長崎県長崎市).

10. 中島 健志・原口 誠・大久保 好章, 局所分枝限定探索による概念プール更新操作に基づく萌芽的概念のボトムアップ抽出, 情報処理学会 数理モデル化と問題解決研究会, 2010 年 5 月 21 日, 群馬大学(群馬県前橋市).

11. E. Tomita, Y. Sutani, T. Higashi, S. Takahashi, M. Wakatsuki, A Simple and Faster Algorithm for Finding a Maximum Clique, 情報処理学会研究報告アルゴリズム研究会, 2010 年 1 月 26 日, 九州大学(福岡県福岡市).

12. 中島 健志・原口 誠・大久保 好章, 萌芽的閉包を枚挙する分枝限定法について, 情報処理学会 数理モデル化と問題解決研究会, 2009 年 12 月 17 日, 電気通信大学(東京都調布市).

13. 李 愛香・原口 誠・大久保 好章, 相関コントラストの最適化, 人工知能学会全国大会(第 23 回), 2009 年 6 月 19 日, サンポートホール高松(香川県高松市).

〔図書〕(計 1 件)

1. 富田 悦次 (広中平祐 編, 分担執筆), 現代数理科学辞典 第2版, 1450 頁, 丸善, 2009.

〔産業財産権〕

○出願状況 (計 0 件)

○取得状況 (計 0 件)

〔その他〕

６. 研究組織

(1) 研究代表者
原口 誠 (HARAGUCHI MAKOTO)
北海道大学・大学院情報科学研究科・教授
研究者番号：40128450

(2) 研究分担者
富田 悦次 (TOMITA ETSUJI)
電気通信大学・名誉教授
研究者番号：40016598

(3) 研究分担者
大久保 好章 (OKUBO YOSHIAKI)
北海道大学・大学院情報科学研究科・助教
研究者番号：40271639