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　25 年　5 月　30 日現在

機関番号：１３５０１
研究種目：　基盤研究（C）
研究期間：　2010～2012
課題番号：　２２５００１２７
研究課題名（和文）
系列データマイニングと高次推論の統合による大規模テキスト時系列からの知識発見
研究課題名（英文）　Knowledge Discovery from Large-scale Text Sequences by Integrating Sequential Data Mining and Advanced Reasoning.
研究代表者
岩沼　宏治（IWANUMA KOJI）
山梨大学・大学院医学工学総合研究部・教授
研究者番号：　30176557

研究成果の概要（和文）：
本研究では，大規模テキスト系列データからの知識発見・抽出手法の確立を最終目的として，幾つかの高速なテキスト系列データマイニング技術を開発し，また抽出したデータの構造化と圧縮を行う手法も新しく提案した.更に潜在的な因子やルールの発見･抽出手法，および欠落情報の補完などを行うための高次推論法の開発も行った．これらに関して種々の理論考察，及び実証評価の両方の側面から研究を推進した.

研究成果の概要（英文）：

In this research, we developed some new fast and effective sequential data-mining technology for the knowledge discovery from series of large-scale text data. We also proposed a new method for structuring and compressing a huge amount of data which are extracted in data-mining process. Furthermore, we gave some new algorithms for extracting latent rules in the form of negative association rule mining, and also for discovering missing factors in the form of inductive reasoning. We study these issues not only from a theoretical viewpoint but also from an experimental evaluation.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,200,000	360,000	1,560,000
2011 年度	900,000	270,000	1,170,000
2012 年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,200,000	960,000	4,160,000

研究分野：　総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：　データマイニング，オンライン型アルゴリズム，近似アルゴリズム，相関ルール，潜在的因子，コーパス，テキスト系列

１．研究開始当初の背景

日々変化していく WEB コンテンツを「老若男女誰でも簡単・便利に利用できる」ためには，時系列テキスト中の重要語を認識・予測する技術が極めて重要である．例えば「地

震」の後に「火事」が起こり，更にその後に「災害対策本部設立」などのイベントが生じることを予想・自動追跡することは，従来型の言語資源（ 国語辞典やシソーラス，用例辞典など）では極めて困難である．統計などに基づく従来手法にも大きな限界がある．この壁を打ち破るためには，イベント系列コーパスのような新しい言語・知識資源の開発が必要不可欠である．イベント系列コーパスの自動構築は大変困難な課題と考えられるが，近年発展してきた自然言語処理技術，データマイニング技術，機械学習や仮説推論，および「情報爆発」とも形容されるほどの巨大なテキスト時系列データを利用すれば，その自動構築はかなりの程度可能と推測される．

データマイニングは Agrawal らによる相関ルール高速抽出法Apriori が提案されて以降，多くの研究がなされている．FP-growth 法による分割統治法，頻出飽和アイテム集合などの抽出データの圧縮など第 2 世代の技術が開発され，対象データもグラフや数値データなど多種多様な構造が考察されている．系列データ集合に対しては，飽和頻出系列を pattern-growth 法に基づき抽出する CoSpan 法や BIDE 法などが開発されている．

本研究では，次世代系列データマイニング技術と，潜在因子抽出・仮説推論などの高次推論技術を新たに開発統合し，長大な単一の時系列テキストデータからイベント系列コーパスを自動生成することを目的としている．これまで，複数の系列データの頻出部分列の抽出問題は，顧客の購買履歴分析などの応用を目的として，多くの研究が行われてきた．一方で，単一データ系列中に重複出現する部分系列の同定・抽出問題は殆ど研究されていなかった．そもそも合理的な出現頻度尺度そのものが不明であったため，我々は系列全体頻度という新しい評価尺度を開発している（FIT2004 優秀論文賞を受賞）．また頻出な部分系列を，ある限定的な条件下において高速に抽出するアルゴリズム等を開発し，国際的に高い評価を得ている(国際会議 ICDM2005 発表)．更に時系列上のイベント系列の抽出に向けて，重要単語句の抽出その他について研究を継続してきた．これらを 2004 年度上半期の新聞記事コーパス（社会面 13,499 記事）という小規模データに試験的に適用したところ，幾つもの有用なイベント系列の自動抽出に実際に成功している．

しかしながら，データマイニングにより抽出・生成されるルールや部分構造は，通常極めて大量になる．より有用な知識を抽出するためには，データの意味理解に基づくフィルタリングや構造化，抽象化と圧縮などの処理が重要である．フィルタリングに関しては，これまで統計的尺度や制約などを導入する試みが多数あるが，時系列テキスト中の系列データに対する手法は，研究代表者の知る限り，まだ提案されていない．抽出したルールや系列の構造化は，Mannila らのエピソード構造や Garriga の半順序構造の抽出などがあるが， それ以外には殆ど研究されていない．抽象化に関しても， 1995 年に Han らによって提案された分類階層に基づく抽象化法があるが，それ以降はあまり進展が無い．また更に，データ中の潜在因子の発見や欠落データの補完なども極めて重要な研究課題であるが，非常に難しい課題であり，これまでのところ殆ど研究されていない．

２．研究の目的

本研究の目的は，次世代の大規模テキスト時系列データマイニング技術を開発を目的として，高速かつ効率的な系列データマイニング手法，及び抽出したデータの構造化と圧縮を行う技術を開発することにある．また 潜在因子の発見および欠落情報の補完などを行うデータマイニングのための高次推論法の開発も行う．提案技術の実証評価の枠組みとしては，イベント時系列コーパスの半自動生成を考え，大規模テキスト時系列からの知識発見問題に取り組む．理論及び実証応用の両方の側面から研究を推進する．

３．研究の方法

本研究では，高速かつ高機能な系列データマイニングと高次推論の統合による知識発見・抽出技術を開発し，新聞記事 10 年分を超える大規模テキスト系列データから，高品質で実用規模のイベント系列コーパスを自動生成する技術の開発に取り組む．具体的には以下の項目について，実証的な性能確認と考察改良を加えながら，研究を順次遂行していく．

(1) 長大な単一データ系列に対して，高速かつ効果的なフィルタリングを行う次世代の系列データマイニング技術の開発

(2) 抽出データ系列の構造化および圧縮を行う技術，また潜在因子の抽出や欠落情報の補完などを目的とした高次推論技術の開発と，その系列データマイニング技術との融合．

４．研究成果

本研究の成果は以下のとおりである．

(1) 情報理論および統計学的尺度に基づく興味深い非同期・非周期パターンの高速

フィルタリング・抽出法の開発を行った．非同期パターンの抽出は，実世界のイベントの系列を抽出するためには重要である．また先行研究の InfoMiner の情報量利得では，有用なイベント系列を大規模テキスト時系列から抽出することは困難であることが分かっている．これに代わる種々の情報量利得基準を新たに導入し，その高速実行アルゴリズムを開発した．新聞記事コーパスを用いて実証的評価を行い，新しいフィルタリング基準および高速化手法双方の良好な性能を確認した．また併せて統計学的な評価尺度も新しく提案し，有効なイベントおよびその系列の抽出に有効であることを確認している．

以下は，2006 年度毎日新聞記事コーパス社会面から自動的に抽出できた重要イベントの集合の例である．2006 年度に特徴的なイベントが抽出されている．

① 教育，センバツ，理事長，テレクラ放火殺人事件，教委贈収賄，ライブドア強制捜査，姉歯事件 ,伊福部昭，台湾補償，リスニングトラブル

② 北海道，拉致，滋賀，東京地裁，偽装，サッカー，発注工事談合，全国高校野球，飛鳥会事件，高校履修不足，小 1 男児殺害

③ トリノ冬季五輪，ドイツ W 杯，全国高校野球，宇和島 パロマ湯沸かし器事故，高校履修不足，未来通信，錬金術，救い, GW, 春センバツ

(2) より大規模なテキスト時系列データを取り扱いを目標として，既存のオンライン型マイニング手法である Lossy Counting 法を拡張し，高速な系列マイニングのアルゴリズムを開発した．オンライン型アルゴリズムの弱点として，系列おけるデータ爆発，即ちバーストへの脆弱性がある．本研究では，このバースト的なデータの出現を効果的に取り扱うために，更にオンライン型の近似アルゴリズムを開発した．具体的には，バーストデータの格納に必要とされる莫大なメモリ空間を大幅に縮減するために，オンライン実行中にメモリに蓄積するデータを適宜交換する手法を新しく開発した．これにより，実行中の最大メモリ容量を一定の値に制限することが可能になり，バースト現象によるメモリ不足によるプログラムの実行停止などの危機的現象を回避することができるようになる．本研究では，提案アルゴリズムの完全性や導入される誤差の最大値の保証などに関する理論的考察を行った．また提案アルゴリズムを実装し，性能評価実験を行った結果，良好な性能を確認している．

(3) 負の相関パターンの概念に基づく潜在因子の抽出に関するデータマイニング技術の開発を行った．大規模データ中の隠れ因子・事象の間の共起規則，即ち潜在的相関ルールのマイニングは極めて重要な問題であるが，これまで殆ど研究されておらず，僅かに負の相関ルールマイニングの研究が少数ある程度であった．基本的に，負の相関ルールのマイニングでは，非頻出アイテム集合を取り扱う必要があるため，計算量が膨大で効率化が難しい。これまで完全かつ効果的な負の相関ルールの生成アルゴリズムは知られていなかったが，我々は新たに負の相関ルールの生成に関して完全なアルゴリズムを提案した。提案手法は既存手法とは異なり，負の相関ルールの台集合（非頻出アイテム集合の一種）は生成せずに，頻出集合だけを使って負の相関ルールを生成する．そのため基本的に非常に効率的なアルゴリズムとなっている。更に接尾木上の極小性チェックに基づくデーマイニングの高速化技術を開発している．実験的評価により 100 倍から 1000 倍の高速化性能を確認している。

(4) 帰納推論に基づく欠落情報の補完と抽出する手法を開発した．逆包摂法に基づく帰納推論を考察し，その応用としての仮説推論による欠落情報の補間手法を新たに提案した．試作システムを実装し性能評価実験を行い，良好な結果を得た．

(5) 上記の(3)で提案した負の相関ルールマイニングアルゴリズムは「頻出集合だけを使って負の相関ルールを抽出する」ことに特徴があるが，これは本質的にオンライン型の高速実行に適した性質である。このため，データストリーム上の頻出アイテム集合を準オンライン型で抽出する高速アルゴリズムも開発した．我々が提案した準オンライン型アルゴリズムでは，抽出した頻出アイテム集合を代表パターンとよばれるものへ非可逆的に圧縮を行いながら，マイニングを行うことができる．誤差保証も与えることができるため，潜在的相関ルールマイニングに一定の誤差保証を与えることが可能となる。本手法により，巨大なテキスト時系列からも，潜在因子まで考慮

した高次のイベント系列コーパスを実用時間で生成することが可能となったと考えられる。

５．主な発表論文等

〔雑誌論文〕（計 6 件）

(1) 岩沼宏治: テキスト系列マイニングにおける有用性尺度について. 人工知能学会誌,Vol.27, No.2 pp.136-145, 2012, 査読有
(2) Yoshitaka Yamamoto, Katsumi Inoue and Koji Iwanuma: Inverse Subsumption for Complete Explanatory Induction. Machine Learning, Vol.86, pp.115-139 (2012). DOI 10.1007/s10994-011-5250-y, 査読有
(3) 岩沼宏治, 鍋島英知, 井上克巳: 一階論理上の等号推論：理論と実際, コンピュータソフトウエア, Vol.28, No.4, pp.282--305 (2011), 査読有.
(4) Hidetomo Nabeshima, Koji Iwanuma, Katsumi Inoue and Oliver Ray: SOLAR: An Automated Deduction System for Consequence Finding, AI Communications, Vol.23, No.2-3, pp.183-203, 2010., 査読有
(5) 村田順平, 岩沼宏治, 大塚尚貴: 情報量と頻度に基づく非同期かつ有用な系列パターンの高速抽出. 人工知能学会論文誌, Vol.25, No.3, pp.464-474, 2010, 査読有.

〔学会発表〕（計 26 件）

(1) 井出典子（発表者）, 岩沼宏治, 山本泰生：極小性を用いた負の相関ルールの効率的な抽出法. 人工知能学会全国大会（第 27 回）2C1-4, 2013 年 6 月 05 日, 富山国際会議場（富山市）
(2) 福田翔士（発表者）, 岩沼 宏治, 山本 泰生: 頻出アイテム集合の即時圧縮を行う準オンライン型ストリームマイニング. 人工知能学会全国大会(第 27 回), 2N5-OS-21b-3, 2013 年 6 月 05 日, 富山国際会議場（富山市）
(3) 小野裕美（発表者）, 岩沼宏治, 山本泰生: アンサンブル法に基づく検索隠し味の精度向上.第４０回知能システムシンポジウムプログラム論文, pp.245-250, 2013 年 3 月 15 日, 京都工芸大（京都市）
(4) 井出典子（発表者）, 岩沼宏治, 山本泰生: 負の相関ルールの完全かつ効率的な抽出法. 第 26 回人工知能学会全国大会, 2B1-R-3-2, 2012 年 6 月 12 日, 山口県教育会館（山口市）
(5) 大柴亮(発表者), 岩沼宏治, 山本泰生: 系列パターン抽出における各種の評価尺度の関係性人工知能学会人工知能基本問題研究会（第 85 回）SIG-FPAI-B104-07, pp.35-40, (2012 年 2 月 2 日, 下呂交流会館（岐阜）
(6) Yoshitaka Yamamoto（発表者）, Koji Iwanuma and Katsumi Inoue, “Non-monotone dualization via monotone dualization” Proc. of the 22th Int'l Conf. on Inductive Logic Programming (ILP2012) , LNCS Vol.7842, 2012/9/17 , Dubrovnik (Croatia) 査読有
(7) Yoshitaka Yamamoto（発表者）, Katsumi Inoue and Koji Iwanuma. Heuristic Inverse Subsumption in Full-clausal Theories. Proceedings of the 22nd International Conference on Inductive Logic Programming (ILP2012), 2012/9/18 , Dubrovnik (Croatia) 査読有
(8) 大柴亮（発表者）, 岩沼宏治, 山本泰生：単一系列データマイニングにおける情報量基準とその補完尺度. 第 25 回人工知能学会全国大会, 2G3-2, 2011 年 6 月 2 日, アイーナいわて県民情報交流センター（盛岡）
(9) 伊藤秀志（発表者）, 岩沼宏治, 山本泰生: 多重データストリーム中のバースト出現に対応したオンライン型頻出系列マイニング. 第 25 回人工知能学会全国大会, 2G3-3, 2011 年 6 月 2 日, アイーナいわて県民情報交流センター（盛岡）
(10) 小野裕美（発表者）, 岩沼宏治, 山本泰生: 専門検索エンジンの半自動構築を目的とした少数データ上のアンサンブル学習, 第 4 回楽天研究開発シンポジウム , 2011 年 11 月 19 日, 楽天タワー2 号館（東京都）優秀研究開発賞受賞 http://rit.rakuten.co.jp/conf/rrds4/papers/RRDS4-009.pdf.
(11) 伊藤秀志（発表者）, 岩沼宏治, 山本泰生: バースト出現へ対応を目的としたオンライン型系列マイニングへのメモリ制限の導入. 人工知能学会データ指向構成マイニングとシミュレーション研究会, 人工知能学会創立 25 周年記念合同研究会予稿集, pp.(2-42)--(2-49), 2011 年 12 月 15 日, 慶応大日吉（神奈川）
(12) Yoshitaka Yamamoto（発表者）, Katsumi Inoue and Koji Iwanuma: Comparison of Upward and Downward Generalization in CF-induction, Proc. of the 21th Int'l.

Conf. on Inductive Logic Programming (ILP2011). LNCS Vol.7207, pp.373-388, 2011, 査読有

他 14 件

〔その他〕
ホームページ：URL: http://www.kki.yamanashi.ac.jp/~iwanuma/Kaken2010/

６．研究組織
(1) 研究代表者
岩沼　宏治（IWANUMA KOJI）
山梨大学・大学院医学工学総合研究部・教授
研究者番号：30176557

(2) 研究分担者
山本　泰生（YAMAMOTO YOSHITAKA）
山梨大学・大学院医学工学総合研究部・助教
研究者番号：30550793

(3) 連携研究者
なし