様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　１９　日現在

機関番号：　３２４０８
研究種目：　基盤研究(C)
研究期間：　2012～2014
課題番号：　２４５００１７５

研究課題名（和文）系列パターンマイニングに基づく属性構築手法の開発

研究課題名（英文）Development of a feature construction method based on sequential pattern evaluation index and temporal pattern extraction

研究代表者

阿部　秀尚（Abe, Hidenao）

文教大学・情報学部・講師

研究者番号：００３９７８５３

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究では，従来，別々の研究分野において開発が行われてきた，自然言語処理における特徴的な語句抽出のための評価指標と系列データにおける系列パターン評価指標を統一した視点から評価指標群として定義し，これらを利用した複合形式データからの有用情報抽出手法の開発を行った。
この結果，定義した指標群がテキストデータおよび系列データ双方で適用可能であることを示した。また，時間経過とともに生成される系列データ集合およびテキスト集合において，分析者が着目する行動の時間的な要因となりうる特徴的なパターンや語句とそれらの出現時系列パターンを同時に抽出可能であることを示した。

研究成果の概要（英文）：This study aims to develop a feature construction method from complex dataset, which consist of not only numerical and nominal values but also their temporal values. However, conventional feature extraction methods have been only developed for each kind of values for constructing features for performing some statistical methods and machine learning methods.
In this study, existing importance evaluating indices from the natural language processing field and sequential pattern evaluation indices were formulated to an integrated viewpoint for evaluating sequential patterns. Then, by combining these indices with temporal pattern extraction method, the method has enable to extract temporal causalities for understanding user behaviors on Web clickstream and temporal text corpus on SNS from more data-centric viewpoint.

研究分野：　知能情報学

キーワード：　系列パターン評価指標　時系列クラスタリング　属性構築　系列パターンマイニング　分類学習

１．研究開始当初の背景

近年，各分野における情報システムの導入が進んでおり，種々の形式でのデータの蓄積が行われ，これらの有効な利活用が求められている。多くのデータ蓄積システムにおいて，表形式によるデータベースが構築され，運用されている。しかしながら，名義値による値が多い，あるいは値の間に時間的・位置的な順序関係がある場合，表形式の項目としてデータを表現することは困難を伴う。このため，名義値の羅列や名義値を文字に置き換えた文字列とした系列データが用いられる。系列データは，名義値として表現する対象の単位を任意に設定することにより，DNA/RNA の塩基配列や文字列といった元始的な対象からイベント系列といった情報粒度の大きな対象まで，多くの対象を観測し，記録することが可能である。

例えば，病院情報システムにおいて，患者選択から次の患者選択までを一区切りとした場合，選択された患者に対する種々の指示やカルテ記載，予約までの一連の流れが一系列データとなる。このような系列データを約 600 床で 1 日の外来患者が約 1000 人の大学病院において取得すると，1 ヶ月で 100 万件以上の系列データが得られる。このようにシステム上で得られた系列データは記録を保存することが現状の主用途であるが，医療安全や診療プロセス理解に繋がる不具合事象に関連する異常操作系列の抽出などの活用が求められている。

以上のような複数の形式の値から成るデータに対しては，数値や名義値を対象に，特徴量を抽出するための属性構築手法が開発されてきた。従来の属性構築手法では，質的な値をもつ名義属性への論理演算や量的な値をもつ数値属性への算術演算を用いて，所与の属性から新たな属性の構築を行ってきた。しかしながら，非構造かつ名義値や数値，テキストなど多様なデータからなる複合形式のデータからの表形式データセットの構築は不可能であった。

これに対し，本研究では，名義値の列として与えられる系列データから，表形式データセットでの属性を構築する手法を開発する。本手法は，系列データから得られる部分系列である系列パターンについて，複数の計量化指標を用いて出現情報を数値化することにより，これらの出現情報と従来の表形式データセットでの属性を併合可能とする手法として開発を行う。これにより，目的概念を説明する有用な系列パターンが属性として構築されたそれらの出現情報を基に同定可能であることを示す。

２．研究の目的

本研究では，系列データから得た有用な系列パターンを用いて，データマイニングプロセスにおける属性構築を行う手法の開発を目的とする。このため，各系列データに対する系列パターンの出現・非出現を表形式データとして変換することにより，目的とする事象（クラス）を説明する有用な系列パターンをデータに基づく規則性から同定することが必要と考えた。このため，系列パターンの有用性を評価する指標を複数用意し，より適応的にクラスを説明する系列パターンの出現情報の規則性が得られる手法の開発を目指す。

３．研究の方法

平成 24 年度においては，系列データベース中で系列パターンが出現する頻度などから種々の性質を数値として計量化する系列パターン評価指標の開発と情報システムの利用者による操作に関する系列データの収集を行った。

本研究では，これまで別々の対象領域の系列データについて開発が行われてきた指標について，系列パターンの評価指標として統一した視点を与える。このため，従来のテキストマイニングにおける重要度指標として用いられた頻度・tf-idf などの他，相関ルールにおける指標を基とした系列マイニングにおける共起指標を系列パターンの評価指標とする。本年度は，これらを定式化し，計算モジュールとして実装する。また，これらに閾値処理などを加えてネットワーク上の中心度などを測るメタ指標について，基本評価指標とそれに対する演算処理とに分けて整理することで，新たなメタ指標の開発を行う。新たなメタ指標は，任意の系列パターンの基本評価指標と併せて，系列パターンの評価が行えるよう，実装した。

情報システムにおける操作系列データの収集については，Web をインタフェースとする情報システムを利用し，利用者の操作状況をマウスカーソル位置，クリック操作，発話などをデータとして蓄積する。蓄積された系列データから，操作系列パターンを生成し，系列パターン評価指標により評価する。この結果を整列結果などとして開発者に示し，利用者の実際の操作手順が開発者の想定と合致したことを評価する指標や不一致を検出する指標を明らかにする。

平成 25 年度においては，利用者のタスク意図やそれに対する不具合の発生を目的とする事象（クラス）として与え，系列パターンの出現との関連性について，教師有り学習アルゴリズムを用いて両者の間の規則性を抽出する。さらに，系列パターン評価指標の値の変化を時系列パターンとして，時間経過による利用者の情報システム操作の変化がタスク意図や不具合事象の発生に与える影響を明らかにする。

まず，前年度収集した発話などの自然言語によるテキストを含む情報システム上での利用者タスク意図の検出に関して，操作と語句の双方に対する系列パターンによる属性を含む表形式データセットへのルールマイ

ニングを実行する。出力された if-then 形式のルールが指摘する語句や操作系列について，タスク意図や不具合事象につながるものとしての有用性については，専門家に評価を依頼する。

さらに，時点毎に計量化を行った時系列データからパターンを得るため，複数の時系列クラスタリング手法を適用する。ここで得られた時系列パターンの組み合わせについて，複数の手法による時系列パターンを属性とした分類ルール学習の正解率をもって，客観的な評価を行う。

また，語句や系列パターンの出現の多寡や構成要素の組み合わせの特異性を表す指標を基本指標とするメタ指標の開発を行う。従来のメタ指標では，特定の基本指標に閾値処理を加えることによって中心性や名望度を計量化してきた。しかし，それ以外の基本指標との組み合わせによってメタ指標がどのような性質を計量化するかについては，未知である。以上の数値時系列パターン，系列パターン評価指標（基本指標およびメタ指標）により，複合形式データから従来の表形式データセットを作成し，評価用データセットとして公開する。表形式データセットに対しては，ルールマイニング以外に種々の分類モデルマイニングなど従来のデータマイニング手法の適用が可能となることを示す。

平成 26 年度においては，時系列パターンの視覚化と目的とする事象との規則性に基づく記述を提示するシステムを作成し，専門家による分類ルールの評価作業支援について効率性向上の観点から評価を行う。また，時間粒度の異なる時系列データセットを作成し，得られる操作系列の差異についても専門家の視点から評価を行う。

利用者タスク意図やその問題点を検出する分類ルールについては，ルールの条件部で指摘される系列パターンの出現傾向や語句の用法変化の傾向が目的とする事象の説明として適切かを専門家により評価する。さらに，専門家による知識提供を受け，目的とする事象のより適切な設定と，知識をより適切に記述すると思われる形式の分類モデルの生成を行う。以上の評価プロセスを繰り返すことにより，専門家による専門知識や専門分野での経験に基づいた判断結果を得る。また，ここで得られた時系列パターンの傾向と代表元である特徴語や特徴的なシステム操作系列パターンの有用性について，専門的見地からの評価を実施する。

以上により，本研究課題で開発する系列パターンと数値時系列パターンを利用した属性構築手法が専門家にとって有用な属性を構築していることを示す。また，本手法の実装が多種の値から成る複合形式のデータから従来の表形式データセットへの変換に有用であることを示す。

以上の結果をとりまとめ，本研究の評価とする。

４．研究成果

本研究では，これまで別々の対象領域の系列データについて開発が行われてきた指標について，自然言語処理における特徴的な語句を抽出するための評価指標と系列データにおける系列パターン評価指標として統一した視点から評価指標群として定義し，これらを利用した複合形式データからの有用情報抽出手法の開発を行ってきた。

平成 24 年度においては，系列データベース中で系列パターンが出現する頻度などから種々の性質を数値として計量化する系列パターン評価指標の開発と情報システムの利用者による操作に関する系列パターン評価指標に基づく操作予測モデルの構築を行った。

本年度は，従来のテキストマイニングにおける重要度指標として用いられる頻度・tf-idf などの他，頻出アイテム集合に対する評価指標として頻繁に用いられる３種類の評価指標を系列パターンの評価指標を定式化し，系列データに対する計算モジュールとして実装した（学会発表⑨）。

これら評価指標の定義および頻度数え上げの基準を組み合わせた計７種の評価指標により，系列パターンを Web サイトのクリックストリームデータに適用し，それぞれの指標での並び替え順序の比較を行った（学会発表⑧）。

また，系列パターン評価指標に基づく Web クリックストリーム予測モデル構築を行った。この結果，共通データセットとして提供される Web クリックストリームデータセットにおいて，高い精度の予測モデルが得られることが示された。さらに，予測モデルに用いられる評価指標とその閾値を用いることで，別々の期間での予測モデルを適用できる可能性について検討を行った（研究発表⑤⑦）。

平成 25 年度においては，系列データベース中で系列パターンが出現する頻度などから種々の性質を数値として計量化する系列パターン評価指標の拡充を行った。また，情報システムの利用者による操作に関する系列パターン評価指標に基づく操作予測モデルの別期間への適用可能性について評価した。さらに，先行研究で示したテキストにおける評価指標群の時系列変化パターンと系列パターン評価指標による時系列変化パターンについて，評価指標間でパターンの検出傾向にどのような関連があるかを比較した。

本年度行った評価指標の拡充では，従来，自然言語処理における用語の自動抽出で用いられてきた評価指標と，語彙の豊富さを表す指標について，系列パターン評価指標としての有用性とそれぞれの指標による並び替え結果の相関について比較評価した（学会発表⑥）。

さらに，系列パターン評価指標に基づく Web クリックストリーム予測モデルの評価で

は，共通データセットとして提供される Web クリックストリームデータセットにおいて，別々の期間での予測モデルを適用できる可能性について評価した結果を示した。

この結果では，系列パターン評価指標による時系列変化パターンの検出では，四半期ごとの Web クリックストリームデータセットにおいて，時系列変化に伴う系列パターンの変遷を検出することを示した。また，複数の系列パターン評価指標間で時系列クラスタに含まれる系列パターンの関連性について，連関係数を用いて評価した。以上より，テキストデータにおける変化パターンの連関よりも指標間での連関が小さなことが示され，各期間における順位相関が高い 2 指標でも時系列変化では異なる傾向をとらえていることが示唆された（学会発表④）。

平成 26 年度においては，系列パターン評価指標として提案された確信度に基づく指標群を時々刻々と算出される時系列テキストデータに適用し，分析者の注目する発信者の行動に繋がる有用な時系列パターンが得られることを明らかにした。

本年度は，語句の重要度評価指標および系列パターン評価指標がテキストデータから抽出する特徴語について，SNS 上での有名アカウントからの情報受信行動を興味の違いとして，特徴語群の違いを比較した（学会発表③）。また，時間経過によって変化する特徴語句の出現傾向から，情報拡散行動の特徴の差異を得るため，各評価指標による時系列パターンについて，抽出と内容の比較を示した（学会発表①②）。

以上より，従来，自然言語処理における特徴的な語句抽出に用いられた評価指標群に加え，テキストデータをアイテムである単語間に順序関係のある系列データとして，系列パターンの抽出と系列パターン評価指標群の適用が可能であることを示した。また，これらの指標群が系列データ集合およびテキスト集合において，分析者が着目する行動や情報の時間的な要因となる時系列パターンが抽出可能であることを示した。

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　1　件）

① Hidenao Abe: Analysis for Finding Innovative Concepts Based on Temporal Patterns of Terms in Documents, Theory and Applications for Advanced Text Mining (S. Sakurai(ed.)), 査読無, pp. 37-50, (2013). DOI: 10.5772/52210

〔学会発表〕（計　9　件）

① Hidenao Abe: Analyzing User Behaviors Based on Temporal Patterns of Sequential Pattern Evaluation Indices on Twitter, The fourth Quality issues, measures of interestingness and evaluation of data mining models workshop (QIMIE'15), 査読有, (2015)

② 阿部秀尚: “Twitter 上での発話履歴の時系列パターンに基づく特定発話行動予測手法の開発“, 情報処理学会　第 178 回知能システム研究会, 情報処理学会研究報告. ICS, [知能と複雑系] 2015-ICS-178(11), 査読無, pp. 1-5 (2015)

③ 阿部秀尚: “フォロワーの興味抽出における系列パターン評価指標利用の検討“, 人工知能学会　第 102 回知識ベースシステム研究会, 査読無, pp. 57-61 (2014)

④ 阿部秀尚: “系列パターン評価指標によるトレンド検出傾向の比較“, 人工知能学会　第 101 回知識ベースシステム研究会, 査読無, pp. 1-4 (2014)

⑤ Hidenao Abe: Developing Transferable Clickstream Analytic Models Using Sequential Pattern Evaluation Indices. In Proc. of AMT 2013, 査読有, LNCS 8210, pp.177-186 (2013)

⑥ 阿部秀尚: “クリック系列パターンマイニングにおける用語性判定指標の算出と比較“, 人工知能学会　査読無, 第 100 回知識ベースシステム研究会, pp. 7-10 (2013)

⑦ 阿部秀尚: “系列パターン評価指標群に基づく転移型クリックストリーム予測モデル構築の検討” 社会システムと情報技術研究ウィーク 2013 (電子情報通信学会 人工知能と知識処理研究会), 査読無, 人工知能と知識処理 112(477), pp.25-30 (2013)

⑧ 阿部秀尚: “プロセスマイニングにおける系列パターン生成と評価指標群に関する考察”, 第 8 回情報システム学会全国大会, 査読無, E2-4 (2012)

⑨ 阿部秀尚: “テキストマイニングにおける語句計量化指標群の利用に関する一考察“, 2012 年度人工知能学会全国大会（第 26 回）, 査読無, 3K2-NFC-3-2 (2012)

〔図書〕（計　0　件）

〔産業財産権〕

○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
http://abe-lab.jp/works/kaken/24500175-2/

６．研究組織
(1)研究代表者
阿部 秀尚（ABE, Hidenao）
文教大学・情報学部・講師
研究者番号：00397853

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：