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　　５　日現在

機関番号：　３２６６５
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０２６２

研究課題名（和文）系列データの動的ネットワーク化による高階データマイニング

研究課題名（英文）Pattern mining in dynamic networks constructed from sequential data

研究代表者
尾崎　知伸（OZAKI, Tomonobu）

日本大学・文理学部・准教授

研究者番号：４０３６５４５８

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：本研究では、ソーシャルメディアに代表されるテキストストリームやセンサーネットワーク等から得られる多次元時系列データを対象とした高度な知識獲得を実現するための基礎技術として、関連性に着目した系列データの異種多次元ネットワーク化技術及び部分類似区間対に着目した系列データの区間イベント系列化とそれに対するパターン検出技術を開発した。
また、実社会における典型的な系列データであるソーシャルメディア、商品購買行動、ニュース記事、株銘柄（株価）などを対象に、相互関連性やバースト現象の観点から分析を行った。

研究成果の概要（英文）：In this research, to develop fundamental technologies on pattern mining from text stream in social media and multi-dimensional time series data in sensor networks, we proposed a method for constructing dynamic networks from text and/or numerical stream as well as a method for converting multi-dimensional time series data into a sequence of interval-based events.
In addition, we conducted analysis on social media, news, purchasing behavior and stock prices, to get insight on these relationship and burst phenomenon.

研究分野：　情報学

キーワード：　パターンマイニング　系列データ

１．研究開始当初の背景

ソーシャルメディアにおけるテキストストリームや、センサーネットワークにおける多次元時系列データなど、近年、系列データはその増加とともに、質の面においても多様化が進んでいる。例えば、代表的なマイクロブログである Twitter では、時間情報はもちろんのこと、GPS 情報やより詳細な情報を得るための短縮 URL など、付加的な属性情報が含まれることも珍しくない。またセンサーデータでは、株価などの従来の時系列データとは異なり、時間的不均一性に加え、常にデータの信頼性（不確実性）が問題となる。これらの「非定型系列データ」を対象とした知識発見・データマイニング技術の確立は、データの急激かつ爆発的な広がりを背景に、緊急性の高い重要な研究課題として認識されており、主にストリームマイニングや社会メディア分析の分野で研究が進められている。

非定型系列データを対象とした知識発見の一つの方法として、データ中のイベントや主体、及びそれらの関係性に着目してデータをネットワーク化するとともに、得られたネットワーク上でネットワーク分析やパターンマイニングを展開することが考えられる。ネットワーク化を通じた非定型系列データからの知識発見では、ネットワークの表現能力の向上やその上での高度な知識発見技術の開発など、適用範囲の拡大や精度向上のために解決すべき課題も少なくない。

ネットワークの表現能力に関しては、非定型系列データの大きな特徴である「主体や関連性の時間的な変化」は動的ネットワークで、また別の特徴である「異なる種類の主体とその間の多様な関係性」は異種多次元ネットワークでそれぞれ対応可能ではあるが、より自然なモデル化を達成するには、これらの特徴を同時に扱うとともに、更なる特徴である付与的な属性情報をも表現できる必要がある。また、得られたネットワークを対象とした分析に関しては、一般的なネットワーク分析に加え、グラフマイニングが代表的ではあるが、両者とも、動的性・異種性・多次元性・不確実性を有するネットワークを直接的に扱えるとは限らない。

２．研究の目的

近年、ソーシャルメディアに代表されるテキストストリームや、センサーネットワーク等から得られる多次元時系列データが爆発的に増加している。本研究課題では、これらの系列データを対象とした高度な知識獲得を実現するための基礎技術として、系列データの動的異種多次元ネットワーク化技術を開発するとともに、より深い知識の獲得を実現するために、大域的な分析を目的としたネットワーク分析と局所的な分析を得意とする構造パターンマイニングを動的ネットワーク上で展開する新たなデータ分析技術を開発することを目的とする。

３．研究の方法

本研究課題の目的を達成するため、研究を（１）系列データの動的ネットワーク化技術の開発、（２）パターン発見技術の拡張、（３）実データを対象とした分析の３つに分け、それぞれ研究を行った。

（１）系列データの動的ネットワーク化技術の開発

各データをノード、強い関連を持つノード間にエッジを張ることで、ネットワーク化を行うことを基本とした。また、対象期間をずらしながらこの操作を繰り返すことで、動的なネットワークを構築する。データ間の多様な関連性を捉えるため、本研究課題では、時系列間の影響を計量する移動エントロピーに加え、部分系列間の類似性を利用した。また、テキストストリームを対象とした場合、頻出単語の出現数系列に加え、トピックモデル（LDA）により得られるトピック分布の系列を考慮することで、より内容を考慮したネットワーク化を行った。これらのアイディアは、Twitter 上の投稿や、商品の売れ行き（消費者行動）、ニュース記事、株価などの実データを対象に検証を行っている。

（２）パターン発見技術の拡張

時系列上の部分類似区間対やモチーフ（繰り返し構造）を抽出することで得られる、単一かつ長大な時間幅を持つ区間イベント系列を対象とした頻出系列パターン発見手法の開発を行った。また、部分類似区間対の系列は、辺に時間を持つグラフとして捉えることもできる。そこで本研究では、時間付きグラフからのパターン発見技術の開発も行った。さらに、パターン発見の基礎研究として、GPGPU を用いた頻出グラフパターン発見の並列実装を行っている。

（３）実データを対象とした分析

主に実社会で生成されるデータの性質を把握することを目的に、Twitter と身体動作データの分析を行った。加えて、論理に基づく手法を用い、人狼ゲーム（人狼 BBS)におけるテキストデータ系列からの特徴抽出・ルール発見を行っている。

４．研究成果

（１）系列データの動的ネットワーク化技術の開発

系列データの動的ネットワーク化技術として、ソーシャルネットワークにおける重要単語（頻出単語）とオンラインストアでの商品をノードとする動的な２部グラフを構築する技術を開発した。具体的には、代表的なソ

ーシャルメディアの一つである Twitter と楽天市場における売り上げランキング情報を対象に、２部グラフを構築する。１日を単位とするツイートにおける各単語の出現数系列とジャンル別の商品売り上げ順位系列とを抽出し、移動エントロピーを用いて情報伝播量を計量することで、有向の２部ネットワークを構築する。また 30 日間を一つの単位とし、10 日間ずつ期間をずらしながら複数のネットワークを得ることで、動的なネットワークを構築している。得られたネットワークに対して、ネットワークに関する基本的な統計量（ノード数やエッジ数など）を分析するとともに、ページランクによる重要ノード（単語や商品）の特定や移動エントロピー値による強い関連性の抽出、さらに可視化による直観的な理解の補助などを行った。

また同様の方法を用いて、株銘柄（株価）とニュース及び SNS（Twitter）との関係の分析を行った。ニュース及びツイートに関しては、トピックモデル（LDA)を適用し、トピックの時間的推移と株価の推移からニューストピック－株銘柄、もしくは SNS トピック－株銘柄の２部動的ネットワークを構築している。さらに、構築されたネットワークに対しコミュニティ発見技術を適用することで、関連の強いトピック株銘柄群の抽出を行った。

一方、移動エントロピー値ではなく、部分系列間の類似性に着目した動的ネットワーク化技術の開発も行った。具体的には、DTW 距離の基づくワンパスアルゴリズムである CrossMatch を利用して多次元時系列の任意の２系列間で類似する部分区間対を抽出する。これにより、各系列を頂点、類似部分区間を（その区間における）辺とする動的ネットワークを構築する。これにより、区間（時間幅）を持つ辺という新たなデータ構造を抽出することに成功している。

（２）パターン発見技術の拡張
多次元時系列データから類似部分区間対に基づき構築される、辺に時間幅（生起時間と消滅時間）を持つ単一の動的ネットワークを対象とした、頻出部分グラフ列挙アルゴリズムを開発した。具体的には、辺に生起時間情報を持つグラフを対象とした手法を拡張するとともに、頻度計算を時間幅を考慮したものに改変している。なお、類似部分区間対に基づくグラフは、グラフにおける辺が持つ情報が一段階抽象化されている。すなわち、元の時系列データに立ち返った際に、各辺に対応する対内では部分時系列の形状が類似しているが、対間では必ずしもその形状は類似しているわけではない。今回開発した手法は、前処理と合わせ、既存手法では捉えることのできない一段階抽象度の高いパターンの抽出を実現しており、新たなパターンの開発という点で、その意義が高いと考えている。

一方、動的ネットワークは、単一辺系列と捉えることが可能である。この点に着目し、動的ネットワークを対象とした、辺系列パターンの抽出アルゴリズムの開発を行った。グラフの場合同様、今回対象とする動的ネットワークは辺に時間幅を持つため、辺系列は区間イベント系列となる。（時間幅を持たない）点系列に対する手法と、複数の区間イベント系列に対する手法を組み合わせ、新たに単一区間イベント系列からのパターン発見を実現している。

一般に、構造データを対象としたパターン発見には高い計算コストが必要となる。今回、この問題を軽減する、すなわち計算時間を短縮することを目的に、グラフパターン発見アルゴリズムの GPGPU による再実装を行っている。

（３）実データを対象とした分析
利用者の行動の特徴を捉えることを目的に、Twitter 上で急激に投稿が増える現象（バースト現象）の分析を行った。具体的には、日本のプロ野球の試合 35 試合に関するツイートを対象とし、バースト時と非バースト時におけるリツイート率やリプライ率、画像や URL を含むツイートの割合や平均文字数の違いを比較した。また、野球における典型的なバースト要因を 4 種選定し、それぞれを要因とするバーストの特徴を比較、分析した。さらに、クラスタリング技術を適用し、バースト現象の類型化を試みた。

SNS とは異なる分野における時系列データとして、身体動作に関する時系列データの分析を行った。具体的には、腕をあげた状態で胸の上の部分でフラフープを回す Chest Hooping という技を対象に、動作的側面と認知的側面の二つの側面から、動作習得過程について考察を行った。動作的側面からの分析では、腰の位置座標時系列に着目し、動的時間伸縮法に基づく非類似度計算を用いて動作の変遷を調査した。また認知的側面からの分析では、メタ認知的言語化の観点から、気づきを対象に頻出語や重要語の変遷を調査した。これらの分析を通じ、Chest Hooping の習得過程の考察を行った。

不完全情報ゲームの一つである人狼ゲームは、近年人工知能研究の新たなターゲットとして注目を集めている。本研究課題では、特にインターネット上のテキストによる人狼ゲーム（人狼 BBS）のログデータを対象に、論理プログラミングの技術を用いて特徴的なルールの抽出を行った。具体的には、ゲーム中に襲撃や処刑などによって追放されるプレイヤの傾向を、帰納論理プログラミングを用いて明らかにした。自然言語で与えられる会話ログを適切な述語に変換するととも

に、それらの前後関係等を背景知識として用いることで、具体的な発言だけでなく、その前後関係やプレイヤ間の関連性を考慮したルールを抽出することに成功している。また人狼ゲームでは、人狼同士は秘密の会話が可能であるが、この秘密の会話グが、実際の行動に与える影響についても調査を行った。これには、論理に基づくアクションルールという新たな形式のパターンを利用している。

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）

〔学会発表〕（計１２件）

① Saki Kawanobe and Tomonobu Ozaki、Extraction of Characteristic Frequent Visual Patterns by Distributed Representation 、 The 2017 31st International Conference on Advanced Information Networking and Applications Workshops、pp.525-530、2017.03.29、Taipei（Taiwan）

② Tomonobu Ozaki, Xia Gao and Mako Mizutani、Extraction of characteristic sets of ingredients and cooking actions on cuisine type、The 2017 31st International Conference on Advanced Information Networking and Applications Workshops、pp.509-513、2017.03.29、Taipei（Taiwan）

③ Tatsuya Toki and Tomonobu Ozaki、Experimental Evaluation of a GPU-based Frequent Subgraph Miner using Synthetic Databases、Proc. of the 2016 Fourth International Symposium on Computing and Networking Symposium/Workshops、 pp.504-507、2016.11.24、Hiroshima（Japan）

④ Ema Nishizaki and Tomonobu Ozaki、Behavior analysis of executed and attacked players in Werewolf game by ILP、Proc. of the 26th International Workshop on Inductive Logic Programming (Short paper)、2016.9.6、London (UK)

⑤ Saki Sakaguchi and Tomonobu Ozaki、An experimental analysis of whispers' effect in Werewolf BBS by relational association rules、Proc. of the 26th International Workshop on Inductive Logic Programming (Short paper)、2016.9.6、London (UK)

⑥ 鈴木 湧人、尾崎 知伸、区間イベント系列からの頻出部分グラフマイニング、2016 年度人工知能学会全国大会（第 30 回）, 3I4-1、2016.6.8、北九州国際会議場（福岡県・北九州市）

⑦ 鈴木 湧人、尾崎 知伸、単一区間イベント系列からの頻出系列パターンマイニング、人工知能学会 第 108 回知識ベースシステム研究会、SIG-KBS-B504-05, pp.24–29、2016.6.5、北九州市立商工貿易会館（福岡県・北九州市）

⑧ 石井 悠加里、尾崎 知伸、調理手順木に基づくレシピクラスタリング、情報処理学会 第 78 回全国大会、5K-06、2016.3.11、慶應義塾大学 矢上キャンパス（神奈川県・横浜市）

⑨ 矢冨 匡祐、尾崎 知伸、プロ野球ツイートのバースト現象の分析、人工知能学会第 10 回データ指向構成マイニングとシミュレーション研究会, SIG-DOCMAS-010-02、2016.3.4、ルスツリゾートホテル（北海道・虻田郡留寿都村）

⑩ Tomonobu Ozaki、Analysis of Hula Hoop Skills by using Dynamic Time Warping and Meta Cognition、Proc. of the Second International Workshop on Skill Science(Poster Presentation) 、2015.11.18、Kanagawa（Japan）

⑪ 尾崎 知伸、金城 敬太、パターンマイニング技術を用いた特徴的食材構造の抽出に関する基礎検討、人工知能学会 第 106 回知識ベースシステム研究会、SIG-KBS-B502-06, pp.30--35, 2015.11.12、慶應義塾大学 日吉キャンパス 来往舎（神奈川県・横浜市）

⑫ 天神 雄貴, 尾崎 知伸、移動エントロピーによる動的ネットワーク化を用いた SNS と商品購買の相互関係の分析、人工知能学会 第 104 回知識ベースシステム研究会、SIG-KBS-B403-03、pp.13–17、2015.3.4、ルスツリゾートホテル（北海道・虻田郡留寿都村）

６．研究組織
(1)研究代表者
尾崎　知伸（OZAKI, Tomonobu）
日本大学・文理学部・准教授
研究者番号：40365458