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

令和　２ 年　 ５ 月 ２９ 日現在

機関番号：３２６６５
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００３１５

研究課題名（和文）深層学習技術を用いた頻出構造パターン発見の高性能化

研究課題名（英文）Improvements of substructure pattern mining using representation learning.

研究代表者
尾崎　知伸（OZAKI, Tomonobu）

日本大学・文理学部・教授

研究者番号：４０３６５４５８

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究課題では、低品質かつ理解困難なパターンの大量生成というパターンマイニング分野における本質的問題を軽減することを目的とし、頻出パターンに対する表現学習技術の開発と応用を行った。その結果として、特徴的なパターンの抽出や、頻出パターン、相関ルールに対する新たな評価関数の開発に成功した。また応用研究として、表現学習技術を用いた料理レシピの分析と不動産間取り図を対象とした解釈容易モデルも抽出を行った。

研究成果の学術的意義や社会的意義
頻出パターン発見分野において、パターンの理解容易性の向上と実応用における有益なパターンの獲得は、それぞれ重要な研究課題として認識されている。本研究課題は、表現学習技術を用いてこれらの研究課題にアプローチするものであり、パターン発見の実応用における障壁を軽減し、より現実的な応用への展開において大きな波及効果が期待される。一方、本研究課題で提案する枠組みの本質の一つは、表現学習技術を用いた分析結果の精緻化であり、深層学習における一段抽象度の高い応用であると考えている。

研究成果の概要（英文）：Generation of large number of low quality and uninterpretable patterns is one of essential drawbacks in pattern mining. To alleviate this drawback, in this research, we developed a framework on learning distributed representation of frequent substructure patterns. Using the framework, we built an algorithm for specifying the characteristic patterns as well as new evaluation functions on frequent patterns and association rules.
In addition, as an application study on graph mining, we conducted the analysis of cooking recipes. We also tried to extract interpretable models for the classification of room layouts in the real estates.

研究分野：データマイニング

キーワード：頻出パターン発見　表現学習　グラフマイニング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

分散表現学習は、埋め込み技術とも呼ばれ、共起情報などを用いて対象データを低次元ベクトル空間へと写像する。得られるベクトルが、加法構成性（ベクトルの計算が意味の計算に相当すること）を満たす可能性があることから応用が広がり、文書や画像、ネットワークを対象とするなど技術的な発展を見せている。これらに加えて、データに対する付帯的な情報を考慮した埋め込み技術も開発され、目的や状況を考慮した種々の低次元ベクトル表現を獲得することが実現されつつある。

一方、データ集合中に頻繁に現れる特徴的な組み合わせをパターンとして抽出する頻出パターン発見は、データマイニングの基本的な問題であり、これまでに動画像や軌跡データなども含め、様々な拡張と幅広い分野での応用が行われている。一般に頻出パターン発見における問題点として、（自明、偶発的という意味で）低品質なパターンが大量に抽出されてしまう点や、得られたパターンの解釈が必ずしも容易ではないという点が指摘されている。これまでに、前者に対する解決策として、頻度に着目した代表元に基づく圧縮表現や、パターンに対する制約の導入、主観的評価尺度の援用、統計的なテストに基づくフィルタリング、直交するパターン集合の抽出など、様々な手法が提案されている。また後者に対する解決策として、各パターンに対して強く関連するパターンを関連付けることで外部（関連付けられたパターン）から解釈を支援する関連パターン集合や、（関連性という意味で）余計な構成要素を排除して本質的な部分だけを取り出すことで内部から解釈を支援する強相関構造パターンなどが提案されている。以上のように、これまでパターン発見の本質的な弱点である「低品質かつ理解困難なパターンの大量生成」の解決に向けて複数の視点から継続的に研究が行われているが、実応用までを考えた場合、必ずしも十分な問題解決が達成されているとは言えない。この原因の一つは、既存アプローチの多くがパターンの出現位置などの表層的な情報しか利用していないことにあると考えられる。これに対し本研究では、表現学習により潜在的な情報として得られる多様なベクトル表現を積極的に利活用する枠組みを提案する。

２．研究の目的

本研究課題では、表現学習技術と頻出構造パターン発見技術を橋渡しする基礎技術を開発することで、両者を深化させることを目的とする。 具体的には、分散表現学習技術を用いて各部分構造パターンを多様なベクトル空間へと展開することで、「低品質かつ理解困難なパターンの大量生成」という頻出パターン発見が本質的に抱える問題を解決する。加えて、ベクトル空間上で各構造パターンとその構成要素の関係性を捉えることで、ロバストかつ多様な特徴的部分構造パターンを新たに開発する。一方で、構造データに対する深層学習技術を開発することで、深層学習の適用分野の拡大を目指す。

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）頻出パターンに対する表現学習技術の構築、（２）分散表現を用いたパターンの評価、（３）部分グラフパターン発見手法の拡張、の３つに分けて研究を行った。

（１）頻出パターンに対する表現学習技術の構築

複数のトランザクションに共通して現れる組み合わせを頻出パターンと呼ぶ。本研究では、不要パターンの排除や、各パターンの評価・解釈基準の基礎を与えるため、各パターンに対する分散表現獲得手法を構築する。また、最もシンプルなアイテム集合パターンに加え、グラフ構造データを対象とした手法を構築する。

（２）分散表現を用いたパターンの評価

各パターンの分散表現を用い、パターン間の大域的な関係性に基づく評価関数を構築する。またパターンの構成要素に着目し、頻出パターン発見問題における構成要素であるアイテム、トランザクション、パターンの各分散表現を利用することで。多様な評価関数を実現する。さらにこれらを発展させ、各ルールにおける前提と帰結の関係性を計量することを念頭に頻出パターンに対する評価関数を拡張し、拡張相関ルールに対する評価関数を構築する。

（３）部分グラフパターン発見手法の拡張

部分グラフパターン発見の発展として、GPU を用いた高速実装、単一区間イベント系列からのパターン列挙技術、グラフマイニングを用いた解釈容易モデルの獲得をそれぞれ行う。

４．研究成果

（１）頻出パターンに対する表現学習技術の構築

頻出アイテム集合に対する分散表現獲得に関し、トランザクションを文書、トランザクションに含まれる頻出パターンを単語にそれぞれ対応付けることで、それぞれの分散表現を獲得する技術を開発した。具体的には、対象パターンを飽和パターン（頻度を基準とした同値類における極大元）に限定し、各トランザクションに包含される飽和パターンの極大元集合を新たなトランザクションとする変換を行い、変換結果に対して従来の表現学習技術を適用する。これにより、

パターンの包含関係を考慮した分散表現の獲得が期待できる。また得られた分散表現を利用し、クラスタリング技術を通じた代表的パターンの特定や、密度に基づく例外発見手法を通じた例外パターンの抽出、更にネットワーク分析を通じた特徴的パターンの同定技術を開発している。

一方、同様のアイディアを部分グラフパターンの表現学習にも援用した。具体的には、ノイズを考慮した飽和パターンと、同じくノイズを考慮した極小パターン（頻度を基準とした同値類の極小元） を対象に、トランザクションを基にした集合を構築し、StarSpace と呼ばれる技術を利用して部分グラフパターン及びトランザクショングラフそのものの分散表現を獲得する。また本手法を料理レシピの有向非循環グラフ表現であるレシピフローグラフへと応用し、料理種分類や代替食材の発見へと応用している。

（２）分散表現を用いたパターンの評価

より多様な側面からの頻出パターンの評価を目的に、局所的な関係に着目し、パターンに加えてアイテムとトランザクションの分散表現を利用する手法を開発した。具体的には、パターンの構成要素であるアイテムの分散表現に基づく評価関数を３種、パターン導出に利用されたトランザクションの分散表現に基づく評価関数を２種、パターンの分散表現に基づく評価関数を３種、提出した。アイテムの分散表現に基づく評価関数に関しては、各パターンの構成要素であるアイテム同士の距離に着目し、多様なアイテムのみで構成されるパターンを高く評価する最短距離の基づく関数と類似アイテムのみで構成されるパターンを高く評価する最遠距離に基づく関数に加え、平均距離に基づく関数を提案している。また、トランザクションの分散表現に基づく評価関数としては、パターンの支持集合（パターンを含むトランザクションの集合）がパターンの特徴を表していると仮定し、上位集合パターンの支持集合または下位集合パターンの支持集合との差が大きいパターンを高く評価する関数をそれぞれ設計している。加えて、パターンの分散表現が各パターンの性質や役割を表していると仮定し、分散表現空間上で、上位集合パターンまたは下位集合パターンと大きく離れているパターンを高く評価する関数をそれぞれ考案した。またパターンの分散表現と、その構成要素であるサイズ１のパターンの分散表現の重み付き平均との差を基準とする関数も準備した。これは、構成要素の単純な合成では予想・説明することのできないパターンに対して高い評価を与えることに相当する。

一方、相関ルールに対する評価関数の開発に関しては、各ルールにおける前提と帰結の関係性を計量することを念頭に頻出パターンに対する評価関数を拡張し、新たに 6 種の関数を提案した。まず、アイテムの分散表現が、各アイテムの性質・役割を表していると仮定し、ルールの前件と後件の差が大きい方ルールを高く評価する関数を設計した。前件と後件はそれぞれアイテムの集合であるため、集合間距離として最短距離、最長距離、平均距離が考えられ、それぞれを基準とする関数を準備している。またトランザクションの分散表現を利用した関数として、ルール前件の支持集合を、後件も満たすグループと満たさないグループの２つに分け、両グループでのクラス間分散を評価値とする基準を開発した。加えて、前件と後件それぞれのパターンとして分散表現の差を直接評価値として用いる基準や、前件の構成要素の単純な重み付き和と後件との差、すなわち前件を用いた後件の予測困難性を基準とすることも提案している。

相関ルールの拡張として、ある条件下での対比関係を捉える間接相関ルールと属性値間の変化に関する関連性を表す相関行動ルールがあげられる。両者はいずれも、相関ルールの対として形式化される。このことに着目し、両拡張相関ルールに対する評価関数の開発を行った。間接相関ルールに関しては、パターンの分散表現が、各パターンの性質・役割を表していると仮定し、メディエータ（前提条件）を考慮した場合に帰結対の距離が小さくなるようなパターンに高い評価値を与える関数を準備した。また加えて、一方の相関ルールから他方の相関ルールの帰結を推定することが困難な場合に評価が高くなる関数を開発した。更に、相関行動ルールに対し、前提の小さな変化が帰結の大きな変化に繋がる場合に高評価となる関数を設計している。

（３）部分グラフパターン発見手法の拡張

(i) GPGPU を用いたノイズ許容頻出飽和部分グラフ発見の高速化

頻出パターン発見技術をより汎用的かつ実用的にするためには、その高速化実装が不可欠である。本研究では、ノイズを許容した飽和部分グラフパターンの列挙を対象に、GPGPU を用いた実装を行った。具体的には、探索空間の枝刈りに必要とされる出現マッチング及び飽和性のチェックに必要とされるノイズ許容トランザクションマッチングをそれぞれ GPU 上で実装し、これらを既存の GPU 版頻出部分グラフ発見アルゴリズムと連動させることで全体の並列化を実現している。

(ii) 単一区間イベント系列からのパターン列挙

時間幅、すなわち開始時間と終了時間を持つイベントを区間イベントと呼ぶ。本研究では、単一で長大な区間イベント系列を入力としたパターン発見技術の開発を行った。既存の区間イベントの列挙手法と単一系列におけるパターンの頻度尺度を利用し、逆探索に基づくアルゴリズムを提案すると共に、イベントを起こした主体を変数化することで、より表現力の高いパターンの列挙を実現した。また開発した手法を、株価データとスポーツデータに適用し、初歩的ではあるがその効果を検証した。

(iii) グラフマイニングを用いた解釈容易モデルの獲得

部分グラフパターンを属性とする分類・回帰問題に対し、解釈可能モデルの抽出を試みた。まず、各トランザクションにおける部分グラフパターンの出現数は必ずしも自明ではないことに着目し、複数の視点から新たな支持度を提案した。またこれらの支持度を用いることで、グラフデータベースを，部分グラフパターンを属性、その出現数（支持度）を値とする属性＝値表に変換し、解釈性研究における既存手法を援用することを可能とした。

一方、代表的な解釈性研究の一つである解釈可能決定集合のアイディアをグラフ構造データへと拡張した。具体的には、部分グラフパターンの大きさや構成要素に関する評価関数を導入し、ルール群抽出における最適化関数へと組み込むことで、解釈容易な少数ルール群の抽出を試みている。なおこれらの手法を不動産間取り図の分析に応用し、一定の知見を得ることに成功した。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計18件（うち招待講演　0件／うち国際学会　8件）

１．発表者名 Tomonobu Ozaki
２．発表標題 Extraction of Characteristic Subgraph Patterns with Support Threshold from Databases of Floor Plans
３．学会等名 The 2019 Seventh International Symposium on Computing and Networking（国際学会）
４．発表年 2019年

１．発表者名 Akari Ninomiya and Tomonobu Ozaki
２．発表標題 Learning Distributed Representation of Recipe Flow Graphs via Frequent Subgraphs
３．学会等名 The 11th Workshop on Multimedia for Cooking and Eating Activities（国際学会）
４．発表年 2019年

１．発表者名 二宮 あかり，尾崎 知伸
２．発表標題 レシピテキストと調理画像系列の埋め込みを用いたケーキレシピの分析
３．学会等名 人工知能学会 第119回知識ベースシステム研究会
４．発表年 2020年

１．発表者名 島田 研太，尾崎 知伸
２．発表標題 ベクトル単位での重み共有化と符号化によるカプセルネットワーク軽量化の試み
３．学会等名 人工知能学会 第24回インタラクティブ情報アクセスと可視化マイニング研究会
４．発表年 2020年

１．発表者名
松山 航太，尾崎 知伸

２．発表標題
グラフ構造データを対象とした解釈可能決定集合の拡張

３．学会等名
人工知能学会 第24回インタラクティブ情報アクセスと可視化マイニング研究会

４．発表年
2020年

１．発表者名
尾崎 知伸

２．発表標題
賃料予測残差に着目した特徴的部屋配置の抽出

３．学会等名
人工知能学会 第117回知識ベースシステム研究会

４．発表年
2019年

１．発表者名
二宮 あかり，尾崎 知伸

２．発表標題
部分グラフに基づくレシピフローグラフ分散表現の比較評価

３．学会等名
人工知能学会 第117回知識ベースシステム研究会

４．発表年
2019年

１．発表者名
Tomonobu Ozaki

２．発表標題
Evaluation measures for frequent itemsets based on distributed representations

３．学会等名
The 2018 Sixth International Symposium on Computing and Networking（国際学会）

４．発表年
2018年

１．発表者名
Tomonobu Ozaki

２．発表標題
Evaluation measures for extended association rules based on distributed representations

３．学会等名
The workshops of the 33rd International Conference on Advanced Information Networking and Applications（国際学会）

４．発表年
2019年

１．発表者名
Kazuki Sato and Tomonobu Ozaki

２．発表標題
Estimation of emotion type and intensity in Japanese Tweets using multi-task deep learning

３．学会等名
The workshops of the 33rd International Conference on Advanced Information Networking and Applications（国際学会）

４．発表年
2019年

１．発表者名
尾崎知伸

２．発表標題
分散表現を利用した特徴的相関ルールの抽出

３．学会等名
人工知能学会 第114回知識ベースシステム研究会

４．発表年
2018年

１．発表者名
尾崎知伸

２．発表標題
分散表現に基づく拡張相関ルールに対する評価関数の提案

３．学会等名
人工知能学会 第115回知識ベースシステム研究会

４．発表年
2018年

１．発表者名
佐藤一輝、尾崎知伸

２．発表標題
複数の表現学習手法を用いた日本語ツイートの感情強度推定

３．学会等名
人工知能学会 第115回知識ベースシステム研究会

４．発表年
2018年

１．発表者名
Tatsuya Toki and Tomonobu Ozaki

２．発表標題
Discovery of \delta-tolerance closed subgraphs on GPGPU

３．学会等名
The 2017 Fifth International Symposium on Computing and Networking（国際学会）

４．発表年
2017年

１．発表者名
Yuto Suzuki and Tomonobu Ozaki

２．発表標題
Frequent Pattern Mining in Multiple Trajectories of Football Players

３．学会等名
The 2018 IAENG International Conference on Data Mining and Applications（国際学会）

４．発表年
2018年

１．発表者名
Saki Kawanobe and Tomonobu Ozaki

２．発表標題
Experimental Study of Characterizing Frequent Itemsets using Representation Learning

３．学会等名
The 2018 32nd International Conference on Advanced Information Networking and Applications Workshops（国際学会）

４．発表年
2018年

１．発表者名 土岐 達哉、尾崎 知伸
２．発表標題 GPGPUによる\delta-許容飽和頻出部分グラフマイニングの実装と評価
３．学会等名 人工知能学会 第111回知識ベースシステム研究会
４．発表年 2017年

１．発表者名 鈴木 湧人、尾崎 知伸
２．発表標題 アノテーション付きサッカー軌跡データからのチーム戦術パターン抽出
３．学会等名 人工知能学会 第112回知識ベースシステム研究会
４．発表年 2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考