研 究 報 告 書

「大規模グラフ系列からの知識体系化と理解支援手法の開発」

研究期間: 平成 20 年 10 月～平成 24 年 3 月
研 究 者: 猪口 明博

1, 研究のねらい

計算機技術やネットワーク技術の発展により人間の処理能力を超えるほどの, 多様で, 大規模なデータの生成, 蓄積が可能となった. 多様で, 大規模なデータを分析するための手法を確立することで, 従来手法では獲得できなかった有用な情報を獲得でき, 社会活動を向上させることが可能である. 本研究では, 構造が変化するグラフを分析対象とする. 例えば, 人間関係ネットワークの人間をグラフの頂点, 人間関係をグラフの辺とすると, ある時点での人間関係はグラフで表現することができる. 人間関係は常に一定ではなく, 時間とともに変化する. すなわち変化するグラフで表現することができる. 人間関係に限らず, 遺伝子が頂点, 相互関係が辺である遺伝子ネットワークは進化の過程で遺伝子を新規獲得, 欠落, 突然変異する変化するグラフにより表現可能である. このようなデータを対象として, 共通する変化を見出すことができれば, 将来の構造変化の予測に役に立つと考えられる. 本研究の目的は, 構造が変化するグラフを分析対象として, 変化するグラフに特徴的に現れる共通の変化をマイニングするための手法を確立することである.

2, 研究成果

(1) 変化するグラフから共通の変化をマイニングする手法 GTRACE

本研究の目的は, 構造が変化するグラフを分析対象として, 変化するグラフに特徴的に表れる共通の変化をマイニングする手法を確立することである. 対象とするグラフ系列は以下の通りである.

- グラフ系列において, 頂点数や辺数は増減する.
- グラフ系列において, 頂点ラベルや辺ラベルは変化する.
- グラフ系列において, 連続する 2 つのグラフの構造は大きくは変化しない.

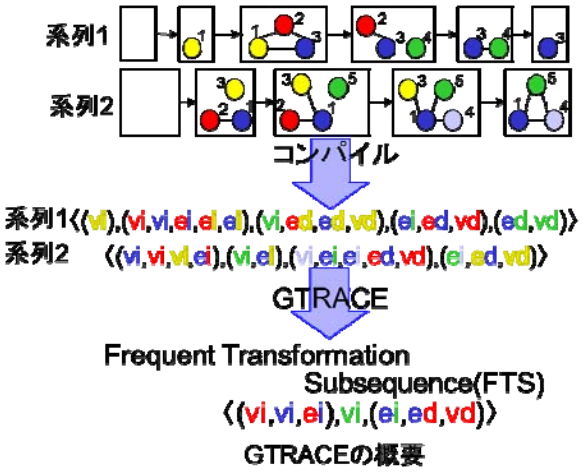
グラフ系列中において, 連続する 2 つのグラフの構造変化は小さいという仮定に基づいて, グラフの変化を表現するために, グラフの変換規則を提案した. この場合グラフ系列の各グラフ全体の情報を保持するのは冗長である. 従って, 連続する 2 つのグラフの差分のみを保持すれば十分であり, 解析手法の領域量を大きく削減することができ, それにより計算量を大きく下げることが可能となる. この変換規則の系列により任意のグラフ系列を表現することが可能である. また, 任意のグラフ系列を頂点数と系列長に比例する計算量で変換規則の系列にコンパイルできる.

さらに変化するグラフに現れる共通の変化をマイニングするために, 変換規則の系列から頻出する部分系列をマイニングする手法 GTRACE (Graph TRAnsformation sequenCE mining) を開発した. GTRACE は, 入力として与えられた n 本の変換規則の系列から k 本以上の変換系列に頻繁に表れる頻出変換部分系列(FTS: Frequent Transformation Subsequence)を効率良くマイニングする手法であり, 頻出変換部分系列(FTS)は変化するグラフに特徴的に表れる

共通の変化である.

提案手法の概要は図に示す通りである. 入力として, グラフ系列の集合と閾値 k が与えられる. 各グラフ系列を変換規則の系列にコンパイルする. コンパイルにより得られた変換規則の系列から GTRACE により頻出変換部分系列(FTS)をマイニングする. さらに, GTRACE に逆探索の概念を取り込むことで, GTRACE を効率化した手法 GTRACE-RS を提案した.

GTRACE を提案して以降, 変化するグラフから特徴的なパターンをマイニングする手法が提案されているが, 辺のみが変化し, 頂点数は一定である手法や頂点の増加のみを許容する手法などグラフ系列に制約を課したものが多い. このような制約を課すことで, グラフ系列のより単純な構造であるグラフに変換でき, 2000 年より知られる既存のグラフマイニング手法で解くことが可能となる. GTRACE は, 他の手法が適用できるグラフ系列に比べ汎用なクラスのグラフ系列の適用可能な手法である.



GTRACEの概要

(2) 変化するグラフから共通の構造をマイニングする手法 FRISSMiner

成果 1 において, 構造が変化するグラフを分析対象として, 変化するグラフに特徴的に表れる共通の変化をマイニングする手法 GTRACE を確立したことを述べた. GTRACE では, グラフ系列中の連続する 2 つのグラフで, その構造は大きく変化しないことを仮定している. しかし, グラフ系列を観測する際(グラフデータを収集する際)に, 時間分解能が低い場合, 観測されたグラフ系列の連続する 2 つのグラフの間で, グラフの大部分が変化する可能性がある. 従って, このようなデータに対する解析に, GTRACE は適さない. そこで, このような課題を克服するために, 我々は, FRISSMiner(Frequent, Relevant, and Induced Subgraph Subsequence Miner)を開発した. FRSSMiner は, グラフ系列の集合と閾値 k を入力として受け取り, グラフ系列中に頻繁に, かつ共通して表れる構造をパターンとしてマイニングする手法である. FRISSMiner では人間に理解容易で, 可読な共通パターンをマイニングするために, 得られるパターンにおける頂点の関連性(relevancy)を連結グラフにより定義し, パターンと入力データとの頂点の関係の再現性を頂点誘導部分グラフ(vertex-induced subgraph)により定義している. FRISSMiner が出力する頻出パターンを関連があるグラフ系列であり, かつ誘導部分グラフ系列として含まれるものに限定したため, FRISSMiner がマイニングするパターンは理解容易であり, それらを効率良く探索することができる.

以上, 成果 1 と 2 により, 構造が変化するグラフを分析対象として, 変化するグラフに特徴的に表れる共通の変化をマイニングする手法を確立した.

(3) グラフ系列マイニング手法の係り受け解析への応用

近年, 係り受け解析は, 情報抽出, 機械翻訳, テキスト含意認識, 質問応答, オントロジー導出などに様々に応用される自然言語処理における基礎技術として注目を浴びている. 係り

受け解析手法は，状態遷移系に基づく手法，グラフ理論に基づく方法，文法に基づく方法に大別することができる．状態遷移に基づく方法は，その内部状態をグラフ，単語（あるいは文節）を頂点，係り受け関係を辺とするグラフで表すことができる．さらに初期状態から最終状態への過程において，各状態を表すグラフは変化するために，状態遷移の系列はグラフの系列により表現することが可能である．状態遷移に基づく係り受け解析器が，出力結果を誤る状態遷移系列の集合から共通する変化（状態遷移の共通性）を見出すことができれば，係り受け解析器が誤る原因の究明，新たな係り受け解析アルゴリズムのデザインなどに役に立つと考えられる．前述のグラフ系列マイニング手法により得られる共通パターンは，人間に可読であり，共通性を見出すという目的に合致する．我々は，状態遷移に基づく係り受け解析器の 1 つである，arc standard shift reduce 型の係り受け解析器について，日本語の係り受け解析の検証実験を行った．この検証実験では，新聞記事からなる京都大学テキストコーパスを分析対象とした．arc standard shift reduce 型の係り受け解析器を用いて係り受け解析を行い，係り受け解析器内の状態遷移の系列をデータ化した．さらに，グラフ系列マイニング手法を適用し，arc standard shift reduce 型の係り受け解析器が解析を誤る典型的な状態の遷移をパターンとしてマイニングした．検証実験の結果，日本語の文法上，妥当なパターンが得られた．また，これらのパターンに基づいて，グラフを書き換えることで，係り受け正答率を改善することができた．以上により，前述のグラフ系列マイニング手法の有用性を確認できた．

さらに，適用手法により得られるパターンが文法上，妥当であった場合，そのパターンに基づいて係り受け関係を修正することで誤った係り受け構造が得られる文は，人手で作成されたコーパスデータが誤りの可能性がある．従ってコーパスの改善にも役に立つと考えられる．

3，今後の展開

成果3に示した手法の原理は，日本語に限らず，また arc standard shift reduce 型の係り受け解析器に限らず適用できる手法であるので，他言語，あるいは arc standard shift reduce 型ではない状態遷移に基づく係り受け解析器へ適用可能である．さらに，これの手法は，係り受け解析に限らず，状態がグラフで表現される状態遷移系全般に適用可能であるため，さらに広い範囲の応用分野に適用可能であると考えられる．

また，さきがけ研究開始時点では，目的の 1 つとして掲げてはいなかったが，データストリームに対するリアルタイム分析も重要な課題の 1 つであると考えられる．Facebook におけるコメントや Titter におけるリツイートなど，人間関係ネットワークにおいて生成されるデータは，データストリームとして大量にインターネット上を流れている．これらの過去のデータによって構成される人間関係ネットワークの構造と現在のデータによって構成される人間関係ネットワークの構造が必ずしも一致するとは限らない．リアルタイムに起こっている構造の変化を検知，把握する分析を行うためには，本研究で開発したバッチ処理による分析手法ではなく，ストリーム処理による分析手法，それらのハイブリッド型の分析手法が必要である．今後，本研究での技術をさらに発展させ，多様で，大規模なデータの分析を，さらに“リアルタイム”に実行する計算基盤技術の確立が重要であると考える．

4，自己評価

この 3 年半のさきがけ研究の研究期間で，構造が変化するグラフデータを対象として，特徴

的な構造の変化をマイニングする手法を開発し，その有用性を示した．開発した手法は，後続の研究により開発された手法よりも汎用なデータに適用できる手法であり，様々な分野のデータに適用可能である．今後，データストリーム処理の概念を取り入れ，リアルタイム分析が可能な処理基盤を開発し，さらに本分野の研究を推進したい．

5　研究総括の見解

大規模データからグラフ構造を、その変化に着目して抽出するグラフマイニングアルゴリズムのメインストリームである手堅い研究である。社会的な応用を意識しながら、データマイニングの先端的な技術を追求していくことを期待していた。

この 3 年半のさきがけ研究の研究期間で、構造が変化するグラフデータを対象として、特徴的な構造の変化をマイニングする手法を開発し、その有用性を示している。開発した手法は、後続の研究により開発された手法よりも汎用なデータに適用できる手法であり、様々な分野のデータに適用可能であることを評価する。

今後、データストリーム処理の概念を取り入れ、リアルタイム分析が可能な処理基盤を開発し、さらに本分野の研究を推進して行くことを期待する。

6，主な研究成果リスト

（1）論文（原著論文）発表

1. Akihiro Inokuchi and Takashi Washio. A Fast Method to Mine Frequent Subsequences from Graph Sequence Data. Proc. of the 8th IEEE International Conference on Data Mining (ICDM 2008), pp 303-312, 2008.
2. Akihiro Inokuchi and Takashi Washio: GTRACE2: Improving Performance Using Labeled Union Graphs. Proc. of 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2010), pp. 178-188, 2010.
3. Akihiro Inokuchi and Takashi Washio. Mining Frequent Graph Sequence Patterns Induced by Vertices. Proc. of SIAM International Conference on Data Mining (SDM 2010), pp. 466-477, 2010.
4. Akihiro Inokuchi and Takashi Washio: GTRACE: Mining Frequent Subsequences from Graph Sequences. IEICE Transactions 93-D(10): pp. 2792-2804, 2010
5. 猪口　明博，頻出パターンマイニングのグラフ系列への適用，人工知能学会学会誌，掲載予定

（2）特許出願

該当なし

（3）その他の成果（主要な学会発表、受賞、著作物等）

- 猪口　明博，鷲尾隆，頂点により誘導される頻出グラフ系列パターンのマイニング，第 12 回 人工知能学会 データマイニングと統計数理研究会，2010
- Nguyen Duy Vinh, Akihiro Inokuchi, and Takashi Washio, Graph Classification Based on

Optimizing Graph Spectra. Proc. of International Conference on Discovery Science (DS2010), pp. 205-220, 2010.

- 生田 泰章, 猪口 明博, 鷲尾 隆, 逆探索法によるグラフ系列マイニングの高速化, 第 3 回データ工学と情報マネジメントに関するフォーラム, B10-2, 2011
- 猪口 明博, グラフ系列マイニング, 第 2 回 Latent Dynamics Workshop (招待講演)
- Akihiro Inokuchi, Hiroaki Ikuta, and Takashi Washio: GTRACE-RS: Efficient Graph Sequence Mining using Reverse Search CoRR arXiv: 1110.3879, 2011
- 猪口明博, 山岡 歩, 鷲尾 隆, 松本裕治, 浅原正幸, 岩立将和, 賀沢秀人, 係り受け解析における状態書き換え規則のマイニング, 第 1 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 予稿集, pp.2.33-2.41, 2011