様式C-19

科学研究費補助金研究成果報告書

平成 22 年 5 月 19 日現在

研究種目：若手研究（B）
研究期間：2008～2009
課題番号：20700136
研究課題名（和文）　時間変化するネットワーク構造データの局所特徴的パターンマイニング手法の開発

研究課題名（英文）　Development of methods for mining characteristic patterns from network structure changing with time

研究代表者
猪口　明博（INOKUCHI AKIHIRO）
大阪大学・産業科学研究所・助教
研究者番号：70452456

研究成果の概要（和文）：グラフ系列で表現可能な実世界の対象は多く存在する．例えば，ある時点での人間関係ネットワークは人が頂点，関係が辺であるグラフで表現できる．さらに，人がコミュニティ（ネットワーク）に参加，脱退することで頂点や辺が増減し，そのグラフの構造は時間とともに変化する．本研究では，変化するグラフ構造をグラフの系列として扱い，グラフ系列から特徴的な変化のパターンをマイニングする手法を研究・開発した．さらに，開発した手法を実装し，人工データと実世界データを用いて，計算効率の特徴を実験的に示し，提案手法の実用性を示した．

研究成果の概要（英文）：There are many real-world applications suitable to model objects by using graph sequences. For example, a human network is represented by a graph where each human and each relationship between two humans correspond to vertices and an edge, respectively. If a person joins or leaves a human community, the numbers of vertices and edges in the graph change with time. In this study, we developed efficient methods for mining characteristic patterns from graph sequences, and evaluated their efficiencies applying them to artificial and real-world datasets.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,700,000	510,000	2,210,000
２００９年度	1,600,000	480,000	2,080,000
総　計	3,300,000	990,000	4,290,000

研究分野：データマイニング
科研費の分科・細目：情報学・知能情報学
キーワード：人工知能，機械学習，アルゴリズム

１．研究開始当初の背景

(1) 膨大なデータから有用な，あるいは興味のあるパターンを知識として発掘するデータマイニングの研究が盛んに行われている．有用性は人それぞれ異なるので定義するのは難しいが，一般に多くの事例を説明できる知識は有用と考えられる．複数のアイテム集合のデータから頻出アイテム集合を列挙する Apriori アルゴリズムが提案されて以来，様々なデータ構造に対して頻出パターン列挙手法が提案されている．近年では，頂点間連結関係と頂点や辺ラベルの情報からなるグラフ構造に頻出する部分グラフパターンをマイニングする手法が提案されている．提

案されているグラフマイニング手法は実用上，非常に効率的であるが，部分グラフ同型問題が NP 完全であるため，より大きな部分グラフをマイニングするのに多くの計算時間を要する．従って，既存手法をグラフ系列のような複数グラフからなる大きなグラフに対して適用することは困難である．

しかしながら，グラフの系列によるモデル化が適している実世界の対象は多く存在する．図 1 は 4 状態，5 頂点 ID からなるグラフ系列を示している． 例えば，人間関係ネットワークは人が頂点，関係が辺であるグラフで表現でき，人がコミュニティ（ネットワーク）に参加，脱退することで頂点や辺が増減する．同様に，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が新規獲得されたり，欠落，突然変異するグラフの系列で表現できる．

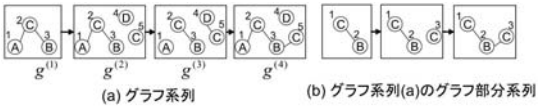

図 1: 観測グラフ系列とそのグラフ部分系列の例

本研究の開始時点で，グラフの頂点数，辺数が変化するグラフ系列を対象として，そこから頻出する部分グラフの変化を列挙する手法は確立されていなかった．そこで本研究では，次節に示す研究目的に基づいて，2 年間の研究を遂行した．

２．研究の目的

(1) 本研究の第 1 の目的は，観測されたグラフ系列の集合に含まれる頻出部分系列を列挙するために，グラフ系列を表現するための簡潔な表現方法を考案することである．そのために，観測グラフ系列中で連続する 2 つのグラフに着目し，その差異を変換規則で表す方法を考案し，変換規則を支配する許容性を理論的に示すことである．

(2)第 2 の目的は，観測グラフ系列を変換規則系列で表わし，そこから頻出変換部分系列(Frequent Transformation Subsequence)と呼ばれる頻出部分系列を列挙（マイニング）する手法 GTRACE (Graph TRAnsformation sequenCE mining)を提案することである．

(3)第 3 の目的は，グラフ系列中でどの頂点同士が関連するかを定義する和グラフを導入し，それに基づいて頂点の関連性を読み取れるグラフ部分系列のマイニング手法を提案することである．我々の興味は関連のある人，イベント，現象に着目することが多いので，系列中の頂点や辺の関連性を読み取れるグラフ部分系列のマイニングは重要である．

(4)第4の目的は,人工データを用いてGTRACEの計算効率の特徴を実験的に示し，実データを用いてGTRACEの実用性を示すことである．

３．研究の方法

本節では，本研究で提案した GTRACE の概略について述べる．その詳細は発表論文を参照されたい．

本研究で提案した GTRACE は，図 1(a)に示すグラフ系列の集合から，それらに頻出する図 1(b)のような系列を列挙する手法である．GTRACE が対象とするグラフ系列は，以下を満たすグラフの系列である．

- 系列中でグラフの頂点数や辺数が増減する．
- 系列中で頂点ラベルや辺ラベルが変わる．
- 観測グラフ系列の中の連続する 2 つのグラフ $g^{(j)}$ と $g^{(j+1)}$ 間でその構造のごく一部のみが変化する．
- 各グラフは疎グラフである．

例えば，一度に大半の人間や遺伝子が入れ替わることはなく，更に各時点では個々の人間や遺伝子は他の一部としか関係を持たない人間関係ネットワークや遺伝子ネットワークのように，実世界の多くのグラフ変化は，これらの仮定を満たしている．

グラフ系列中で連続する 2 つのグラフのごく一部が変化するという仮定より，各グラフ $g^{(j)}$ をその全頂点，及びその間の辺で直接表す方法は冗長である．部分系列を効率よく探索するためには，計算コストと空間コストを抑えるためのグラフ系列の簡潔な表現が必要となる．そこで本節では，GTRACE が用いるグラフ系列の表現形式を考案した．

表１．変換規則

変換規則	
頂点の追加 $vi^{(j)}_{[u,o]}$	ラベルが l，ID が u である頂点を $g^{(j)}$ へ追加し， $g^{(j+1)}$ へ変換
頂点の削除 $vd^{(j)}_{[u,o]}$	ラベルが l，ID が u である頂点を $g^{(j)}$ から削除し $g^{(j+1)}$ へ変換
頂点ラベルの変更 $vr^{(j)}_{[u,o]}$	ID が u である頂点のラベルを l に変更し， $g^{(j)}$ を $g^{(j+1)}$ へ変換
辺の追加 $ei^{(j)}_{[(u_1,u_2),o]}$	ID が u1 と u2 である頂点間にラベル l の辺を追加し， $g^{(j)}$ を $g^{(j+1)}$ へ変換
辺の削除 $ed^{(j,k)}_{[(u_1,u_2),o]}$	ID が u1 と u2 である頂点間から辺を削除し， $g^{(j)}$ を $g^{(j+1)}$ へ変換
辺ラベルの変更 $er^{(j)}_{[(u_1,u_2),o]}$	ID が u1 と u2 である頂点間の辺のラベルを l に変更し， $g^{(j)}$ を $g^{(j+1)}$ へ変換

例えば，図 2 の連続するグラフ $g^{(j)}$ と $g^{(j+1)}$ の間の変化は変換規則系列 $\langle vi^{(j)}_{[(1,A]} ed^{(j)}_{[(2,3),-]} \rangle$ で表わされる．この系列は ID が 1 でラベルが A である頂点が $g^{(j)}$ に追加(vi)され，ID が 2 と 3 である頂点間の辺が $g^{(j)}$ から削除(ed)された結果，$g^{(j)}$ が $g^{(j+1)}$ に変換されたことを意味している．このようにグラフの頂点や辺が多い場合でも，上記の仮定のもとでグラフ系列を簡潔に表現することが可能である．

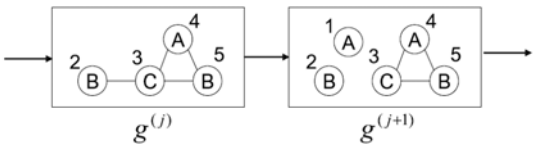

図 2. グラフ系列中の連続する 2 グラフ

上記に示す 6 つの変換規則を用いることにより，如何なるグラフ系列も線形時間で変換規則の系列に変換することが可能である．この系列を変換系列と呼ぶ．

本研究で対象とする問題は，n 個の変換系列が与えられたとき，σ’以上で頻繁に出現する部分変換系列を効率良く列挙することである．列挙される部分系列を頻出変換部分系列(FTS: Frequent Transformation Sub-sequence)と呼ぶ．膨大な量のグラフ系列から FTS を効率良く列挙するために，GTRACE という手法を考案した．GTRACE の特徴は，FrefixSpan という手法を拡張した手法であり，Pattern Growth 法を用いて，FTS となりえない変換部分系列を効率良く排除しながら FTS を列挙する手法である．

例えば，図 3 は 2 つのグラフ系列(n=2)をその変換規則に変換し，2 つの変換規則に共通して現れる FTS をマイニングする例を表している．ここでは，簡単化のため，n=2 の例を用いているが，実際は n が 1,000 以上となっても効率の良い手法を研究・開発するのが目的である．

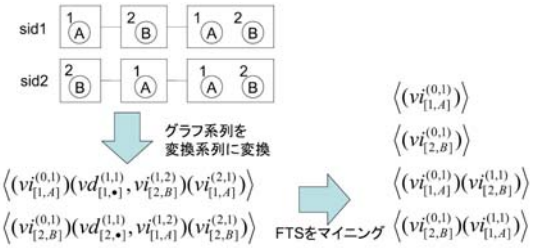

図 3. 2 グラフ系列からの FTS のマイニング

GTRACE は，実用性の観点から出力される系列中の頂点と辺が互いに関連がある(relevant)系列のみを列挙する．例えば，図 4 のグラフ系列では，ラベルが A で ID が 1 である頂点は，どの外部状態においても他の頂点と連結していないため，他の頂点と関連がないと考える．一方，頂点 2 と頂点 4 はどの外部状態においても直接は接続していないが，それらの頂点はラベル B をもつ頂点 3 と，1 番目の外部状態と 4 番目の外部状態でそれぞれ連結している．この場合，本稿では頂点 2 と 4 は頂点 3 を介して互いに関連があると考える．このように，図 4 における関連性のある系列の例として，頂点 2，3，4 を含み，頂点 1 を含まないものが考えられる．そこで，グラフ系列中の各グラフの ID に基づいて，各グラフが重なるようにして生成した和グラフを定義し，和グラフが連結である FTS のみ(rFTS: relevant FTS と呼ぶ)を出力する効率的なアルゴリズムを提案した．

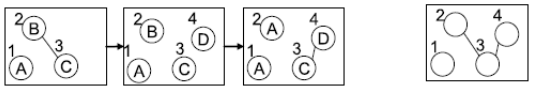

図 4. グラフ系列とその和グラフ

４．研究成果

提案手法の計算効率を検証するために，人工的に生成したデータ，および実世界データを用いて，様々な性能評価を行った．紙面の都合上，その 1 つを説明するが，その他の実験は，他の発表論文を参照されたい．

GTRACE の実用性を評価するために，エンロン社電子メールデータを用いた．このデータは，1998 年 11 月 15 日(日)から 2001 年 3 月 25 日(土)までの 123 週の 182 人の人間間の電子メールやり取りのデータである．182 人それぞれが固有の名前，すなわち，ユニーク ID を持ち，ある 2 人が 1 日の間に電子メールでコミュニケーションをとると 2 頂点間に辺を張り，ある 1 日に対応するグラフ $g^{(j)}$ を生成した．また，各頂点には，CEO，Director，Employee，Lawyer，Manager，President，Trader，Vice President のいずれかが頂点ラベルとして割り当てられている．各週を単位として各々 1 つのグラフ系列を生成した．すなわち，入力となるグラフ系列数は 123 である．

表 2 はランダムに選択した 100 個のユニーク ID から構成されるグラフ系列の集合を対象として，最小支持度 σ’を変化させたときの計算時間，導出された頻出部分系列数，1 頻出部分系列導出あたりの平均計算時間を示している．表 3 は，最小支持度を 50% とし，選択するユニーク ID 数を変化させたときの実験結果である．最小支持度を減少させたとき，あるいはユニーク ID 数を増加させたとき，GTRACE が導出する頻出部分系列の数が増加し，GTRACE の計算時間も増加する．表の 4 列目は 1 頻出部分系列あたりの平均計算時間である．１パターンあたりの計算時間はわず

か数ミリ秒程度であることがわかる．また，ここでは紙面の都合上，掲載していないが，計算時間はグラフ系列数の増加，すなわち n の増加に対して，比例するので，GTRACE を大規模なグラフ系列データにも適用可能である．

GTRACE の全計算時間を t1 とし，GTRACE で呼び出される複数回の PrefixSpan のうち，rFTS を出力しない PrefixSpan を実行するのに要する計算時間を t2 とする．表 2 の最後の列は，t1 に対する t2 の割合を表している．このような PrefixSpan の実行を避けるような探索空間の枝刈りが可能となれば，さらに高速に GTRACE を動作されることができる．この課題に対して，我々は GTRACE2 を開発し，今後開催される国際会議で発表予定である．GTRACE2 は GTRACE に比べ，10 倍から 1000 倍高速に動作する．

表 2 ． 実験結果 1

最小支持度 σ’	計算時間［秒］	列挙された FTS の数	1FTS あたりの計算時間［秒］	t2/t1
20%	73.5	3346	0.022	82.0%
21%	46.7	2827	0.017	79.6%
22%	25.3	2081	0.012	69.4%
23%	15.5	1754	0.008	58.0%
24%	5.0	1495	0.003	51.8%
25%	3.9	1257	0.003	0%
30%	1.1	524	0.002	0%
35%	0.3	196	0.001	0%
40%	0.2	106	0.002	0%

表 3. 実験結果 2

人の数	計算時間［sec］	列挙された FTS の数	1FTS あたりの計算時間［秒］
80	0.02	12	0.0013
100	0.14	26	0.0054
120	0.56	206	0.0027
140	0.81	254	0.0032
160	14.8	790	0.0187
182	87.1	1,376	0.0633

以下の変換規則系列は GTRACE によって列挙された FTS であり，図 5 はこの FTS を図示したものである．

$$\left\langle \mathrm{vi}^{(1,1)}_{[1,\mathrm{CEO}]}\mathrm{vi}^{(1,2)}_{[2,\mathrm{Vice\,Pre}]}\mathrm{vi}^{(1,3)}_{[3,\mathrm{CEO}]}\mathrm{ei}^{(2,1)}_{[(1,2),1]}\mathrm{ei}^{(2,2)}_{[(1,4),1]}\mathrm{ed}^{(3,1)}_{[1,3],1}\right\rangle$$

上記の系列で Vice Pre は Vice President を表している．この系列は，26 個のグラフ系列に出現していた．すなわち，123 週のうち 26 週でこのようなコミュニケーションがとられていた．図 5 では，ユニーク ID が 4 の頂点が $g^{(1)}$ で追加された（出現した）ように描かれているが，この FTS には頂点を追加する変換規則 $\mathrm{vi}^{(j,k)}_{[4,1]}$ が含まれていないために，実際はそれ以前の場合もありうる．また，同様の理由で，変換規則 $\mathrm{vi}^{(j,k)}_{[4,1]}$，あるいはこの頂点のラベルを変更する変換規則 $\mathrm{vr}^{(j,k)}_{[4,1]}$ が含まれていないため，この頂点のラベル（職位）はこの FTS からは特定できない．この FTS より，ユニーク ID が 1 の CEO がその他の 3 人を結びつけるハブの役割を果たしており，この CEO はユニーク ID が 3 の CEO からの情報をユニーク ID が 2 の Vice President や 4 に伝えている可能性があるといえる．

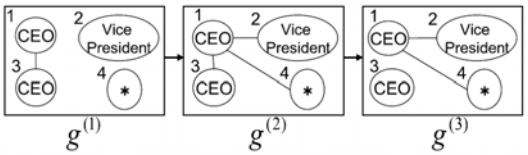

図 5. エンロンデータから列挙された FTS のグラフ表現

本研究では，本研究の開始時点で未確立であった，グラフの頂点数，辺数が変化するグラフ系列を対象として，そこから頻出する部分グラフの変化を列挙する手法 GTRACE を考案した．GTRACE の計算時間はグラフ系列数の増加，すなわち n の増加に対して，比例するので，GTRACE を大規模なグラフ系列データにも適用可能である．また，提案した手法を人工データ，実世界データに適用し，計算効率の特徴を実験的に示し，実データを用いて GTRACE の実用性を示した．

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）

〔学会発表〕（計 8 件）

① Akihiro Inokuchi and Takashi Washio. GTRACE2: Improving Performance Using Labeled Union Graphs, Proc. of The 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining, June, 23, 2010, to appear. Hyderabad International Convention Centre (Hyderabad, India)

② Akihiro Inokuchi and Takashi Washio: Mining Frequent Graph Sequence Patterns Induced by Vertices, Proc. of SIAM Data Mining Conference, pp.466-477, April, 29, 2010. Renaissance Columbus Downtown Hotel, (Columbus, Ohio, USA)

③ 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, pp.131-141, 2010年3月30日, 統計数理研究所（東京都）

④ 生田 泰章, 猪口 明博, 鷲尾 隆, 複雑時系列データのための OLAP システムの並列化手法, 第 23 回人工知能学会 全国大会, 1C1-4, 2009年6月18日, サンポートホール高松（香川県）

⑤ 猪口 明博, 鷲尾 隆, グラフ時系列データからの頻出部分系列マイニング手法の性能評価, 第 23 回人工知能学会 全国大会 2C2-1, 2009年6月17日, サンポートホール高松（香川県）

⑥ 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 pp.303-312, December 17, 2008, Palazzo dei Congressi, (Pisa, Italy)

⑦ Akihiro Inokuchi and Takashi Washio, Feasibility of Graph Sequence Mining based on Admissibility Constraints, The Third International Workshop on Data-Mining and Statistical Science, pp.1-4, September 25, 2008, Tokyo Institute of Technology (Tokyo)

⑧ 猪口 明博, 鷲尾 隆, グラフ系列マイニングのための表現制約とアルゴリズム ,第7回人工知能学会 データマイニングと統計数理研究会, pp.54-62, 2008年7月24日, 小樽市民センター（北海道）

〔図書〕（計 3 件）

① Sanjay Chawla, Takashi Washio, Shin-ichi Minato, Shusaku Tsumoto, Takashi Onoda, Seiji Yamada, and Akihiro Inokuchi, New Frontiers in Applied Data Mining, PAKDD 2008 International Workshops, Osaka, Japan, May 20-23, 2008. Revised Selected Papers Springer 2009.

② Takashi Washio, Einoshin Suzuki, Kai Ming Ting, and Akihiro Inokuchi, Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conference, PAKDD 2008, Osaka, Japan, May 20-23, 2008 Proceedings Springer 2008.

③ Ken Satoh, Akihiro Inokuchi, Katashi Nagao, Takahiro Kawamur: New Frontiers in Artificial Intelligence, JSAI 2007 Conference and Workshops, Miyazaki, Japan, June 18-22, 2007, Revised Selected Papers Springer 2008.

〔その他〕
ホームページ等
http://www.ar.sanken.osaka-u.ac.jp/~inokuchi/

６．研究組織

(1)研究代表者
猪口　明博（INOKUCHI AKIHIRO）
大阪大学・産業科学研究所・助教
研究者番号：70452456

(2)研究分担者
該当なし

(3)連携研究者
該当なし