2A4-3

データストリームに対する相関ルールを用いたコミュニティの時系列解析

Time-Series Analysis of Communities using Association Rule in Data Streams

山口 雄大[*1] 新美 礼彦[*2]
Takehiro Yamaguchi Ayahiko Niimi

[*1]公立はこだて未来大学大学院 システム情報科学研究科
Graduate School of Systems Information Sciences, Future University-Hakodate

[*2]公立はこだて未来大学 システム情報科学部
Systems Information Sciences, Future University-Hakodate

In this paper, we focus on the characteristic that huge data of the data stream changes as time eries, and we propose the technique for discovering the rule and the pattern of the change. Our approach is to express the data stream as a graph sequence, and to analyze the change in the graph structure as a change in the data stream. The proposed approach is extracting the change pattern of each community from graph sequence in consideration of the character of the data stream. The methodology is to extract the change pattern of the community from the graph sequence by focused the transition of the node that composes the community, and analyzing it as an association rule. We use SBM data for experiment, we extracted user community that marked same web pages for long term with high threshold of confidence rate.

1. はじめに

近年、新しいタイプの大規模データとしてデータストリームが注目されている。データストリームとは、「膨大な量のデータが高速なストリームを通じて、時間的に変化しながら終わりなく到着し続ける」という特性を持つ動的な大規模データである。金融や流通分野の取引記録やネットワーク監視システムの通信記録、オンラインニュースなどの多数の情報源から生成されるデータもこれに含まれる。現在、これらのデータからの有用な知識やパターンを見つける要求が高まっている。[有村 05]

このようなニーズを背景に、我々は、その「膨大なデータが時間的に変化する」という特性に着目し、そのデータ中で発生した、変化のシナリオを抽出する手法を提案した[Yamaguchi 09]。変化のシナリオとは、データ中において、何らかのトピックがいつ発生し、どのように発展したのかといった変化の過程と定義した。この研究では、対象とするデータストリームはグラフ構造を持ったデータが時間的に変化するデータセットとし、データストリームの時間的変化をグラフ系列として表現し、密な部分グラフ構造（以降、コミュニティとする）の変化をトピックスの変化として抽出するというものであった。データ自体の変化や、データ間の関係の変化を、グラフ上のノードやリンクの追加、削除によって表現し、より多くのデータと関係のあるデータ集合の変化を解析する手法を提案した。

コミュニティの変化を抽出するにあたり、問題となるのが、グラフ系列中の連続する二つのグラフ間で、如何にしてコミュニティを同定するかということである。コミュニティの同定とは、グラフ系列中のあるグラフに含まれるコミュニティが新たに形成されたものなのか、前のグラフに存在していたものと同一のものを指すのか、前のグラフにおける複数のコミュニティが合わさったものなのか、または前のグラフにおけるひとつのコミュニティが複数にわかれたうちのひとつなのかを特定することである。データストリームをグラフ系列で表した場合、連続する二つのグラフ間で、その構造の大部分が変化するため、全く同一の構成要素を持つコミュニティが存在する確率は極めて低く、その同定手法が問題となる。本稿では特に、この問題を解決するための手法を提案する。

連絡先: 山口雄大,
公立はこだて未来大学大学院システム情報科学研究科,
北海道函館市亀田中野町 116 番地 2,
Mail: g2109046@fun.ac.jp

2. 提案手法

本研究では、以下の 4 つのステップに従って、コミュニティの変化過程を抽出する。まず、(1) グラフ系列中の各グラフからコミュニティを抽出し、(2) グラフ系列中の各コミュニティを構成するノードの変遷を解析し、(3) 相関ルールを用いてコミュニティを同定し、(4) コミュニティの変化過程を抽出する。

2.1 グラフ系列からのコミュニティ抽出

本研究が定義するコミュニティとは、モジュラリティをネットワークの分割過程に用いて、モジュラリティ最大化を目指すアルゴリズムによって抽出される、リンク密度の高い部分グラフのことである。モジュラリティとは、分割されたネットワーク評価指標であり、全分割がどれだけネットワーク全体をバランスよくリンク高密度集団に分割したかを評価している。例えば、対象のネットワークデータ全体が $V_1, V_2, \ldots, V_L$ と L 個の重複しないコミュニティに分割された際、モジュラリティ Q は以下のように定義される。

$$Q = \sum_{l \in 1 \ldots L} Q_l = \sum_{l \in 1 \ldots L} (e_{ll} - a_l^2) \quad (1)$$

e_{ll} は V_l 内部のリンクの存在確率を意味し、a_l は、無向グラフであっても敢えて出・入リンクとして「リンク端」を分けて考えたとき、V_l 内にあるリンク端総数のネットワーク全体に対する存在確率を意味し、それぞれ以下の式で得られる。

$$e_{ll} = \frac{1}{2m} \sum_{i \in V_l} \sum_{j \in V_l} A(i,j) \quad (2)$$

$$a_l = \frac{1}{2m} \sum_{i \in V_l} \sum_{j \in V} A(i,j) \quad (3)$$

$A(i,j)$ は、ネットワークの隣接行列で、ノード i,j 間にリンクがあると 1、なければ 0 を返し、ネットワークのリンク端の総計は $2m$ である。つまり、e_{ll} は、各コミュニティ内部の密度がそれぞれ高いことを求め、a_l は全体をひとつのコミュニティにしている場合や、ランダムな分割に対して Q を下げる補正項として導入されている。

我々は、モジュラリティを最大化する手法のひとつである、*CNM(Clauset-Newman-Moore)* 法 [Clauset 04] を用いてコミュニティを抽出する。この手法の優れた特質として、圧倒的に計算量のオーダーが小さい点、パラメータ調整が不要な点が挙げられる。つまり、ネットワーク構造データを与えるだけで、分割結果を求めることができる [大和田 08, 湯田 08]。

2.2 構成ノードの変遷解析

上記のコミュニティ抽出により得られたコミュニティは、グラフ系列中のグラフ毎に抽出されたことになる。そこで、次のステップでグラフ系列中の隣り合うグラフにおいて、あるグラフのコミュニティが次のグラフのどのコミュニティと同じなのかを同定する。

グラフ系列中の連続する二つのグラフ間でコミュニティを同定するために、各コミュニティを構成する要素がそれぞれのグラフでどのコミュニティに属しているかを探索する。探索するのに際して、まず、抽出したコミュニティに対して、ユニーク ID を割り当てる。

割り当てる各 ID は C_i^k とする。これは、グラフ系列中の k 番目のグラフにおけるあるコミュニティの ID が i であることを表す。そして、各コミュニティに対して割り当てられた ID を用いて、各ノードが属しているコミュニティの変遷を探索する。探索結果は以下の表現を用いて、集計する。

$$ID(C_i^k, v_m) \rightarrow ID(C_j^{k+1}, v_m) \quad (4)$$

これは、グラフ系列中の k 番目のグラフにおいて、ID が i のコミュニティに属していたノード v_m が、$k+1$ 番目のグラフにおいて、ID が j のコミュニティに属している変遷を表している。この集計結果は、k 番目のグラフにおけるコミュニティ ID を条件部、$k+1$ 番目のグラフにおけるコミュニティ ID を結論部とすると、ノードのコミュニティ所属の変化を表した相関ルールととらえることができる。

しかし、コミュニティを構成する全てのノードに対して探索することは、大規模なデータセットに対しては、現実的ではない。実データでは、少数の一部のノードが多くのノードとつながっており次数が高いが、多くのノードは少数のノードとしかつながっておらず次数が低いというフリースケール性が見られることが多い。そこで、現実世界のデータにはフリースケール性があることが多いことを考慮して、全てのノードの変遷を探索せず、各コミュニティ内のハブノードとその周辺ノードのみに着目し、その変遷を探索する。ここでいうハブノードとは、コミュニティ内において最も高い次数を持つノードのことであり、ハブノードと l リンク以内に繋がっているノード集合を周辺ノードと定義する。l はコミュニティの大きさを決める閾値とし、グラフの規模によって設定するものとする。

2.3 相関ルールを用いたコミュニティの同定

コミュニティ内のハブノードと周辺ノードの探索結果を相関ルールとして捉え、それぞれに対して確信度を算出し、その値を用いてコミュニティを同定する。連続する二つのグラフにおいて、k 番目のグラフにおけるコミュニティ ID を条件部、$k+1$ 番目のグラフにおけるコミュニティ ID を結論部として、確信度 ($confidence$) を算出する。その算出式は以下のようになる。

$$confidence = \frac{Number\{ID(C_i^k, v_m) \rightarrow ID(C_j^{k+1}, v_m)\}}{Number\{ID(C_i^k, v_m)\}} \quad (5)$$

$Number\{ID(C_i^k, v_m) \rightarrow ID(C_j^{k+1}, v_m)\}$ とは、$ID(C_i^k, v_m) \rightarrow ID(C_j^{k+1}, v_m)$ という変遷を辿るノード数を表し、$Number\{ID(C_i^k, v_m)\}$ は、k 番目のグラフにおいて、ID が i のコミュニティに属しているノード数を表す。これによって算出された確信度に対して、閾値を設定し、それ以上の確信度をもつ相関ルールを採用し、それを用いてコミュニティを同定する。つまり、コミュニティを構成する中心的なノードの変遷を解析することで、探索コストを抑えながらも、連続する二つのグラフ間で強い関連性をもつコミュニティを見つけることができると考える。

2.4 コミュニティの変化を抽出

コミュニティの時系列変化を解析する先行研究として、ウェブコミュニティの全体的な発展過程を分析している [豊田 03] の研究と、成長するネットワークモデルを対象にコミュニティ構造の時間変化を観察する方法論を提案している [大和田 08] の研究が挙げられる。本研究では、コミュニティの変化の過程を、(1) 生成、(2) 消滅、(3) 統合、(4) 分離、(5) 維持、(6) 拡大・縮小の 6 つ定義する。これらの定義は、[豊田 03] と [大和田 08] におけるコミュニティの変化過程の定義を拡張したものである。

生成: グラフ g^k におけるコミュニティ C_i^k が、グラフ g^{k-1} におけるどのコミュニティにも属していない。

消滅: グラフ g^k におけるコミュニティ C_i^k が、グラフ g^{k+1} におけるどのコミュニティにも属していない。

統合: グラフ g^k におけるコミュニティ C_i^k が、グラフ g^{k-1} における複数のコミュニティと同じコミュニティを構成している。

分離: グラフ g^k におけるコミュニティ C_i^k が、グラフ g^{k+1} における複数のコミュニティと同じコミュニティを構成している。

維持: グラフ g^k におけるコミュニティ C_i^k が、グラフ g^{k+1} において、他のコミュニティと統合することなく、かつ複数のコミュニティに分裂することなく、単一のコミュニティとして成り立っている。

拡大・縮小： グラフ g^k におけるコミュニティ C_i^k が、グラフ g^{k+1} において、構成するノード数が増える・減る（統合・分裂・維持の過程には、構成ノードの拡大・縮小の可能性も含んでいる）

採用された相関ルールに対して、これらの変化過程を当てはめることで、コミュニティの変化を抽出する。最終的に抽出され

るコミュニティの変化は、生成から消滅するまでの変化過程である。例えば、以下のような変化過程である。

$$\text{生成} \rightarrow \text{維持・拡大} \rightarrow \text{分裂} \rightarrow \text{消滅} \quad (6)$$

これは、コミュニティが生成された後、維持・拡大、分裂の変化過程を経て、消滅するという変化の過程を示している。以上の提案手法を適用することで、このようなコミュニティの変化の過程が可能となる。

3. 評価実験

提案手法を評価するのに、ソーシャルブックマーク (以降、SBM とする) を用いた実験を行った。SBM とは、自分のブックマークをネット上に公開し、不特定多数の人間とそれらを共有するという Web サービスである。実験に用いた SBM データは、livedoor 社が提供している livedoor クリップの研究用データセットである [livedoor clip, EDGE]。このデータには、ユーザ ID、ブックマークページの URL、ブックマーク作成時刻、登録タグが含まれている。

ただし、実験では提案手法である閾値による周辺ノードに着目したコミュニティの同定までは行えなかったため、コミュニティの全ノードを対象とした探索となっている。

3.1 実験概要

本研究では、livedoor クリップがサービスを開始した 2006 年 6 月から 2008 年 9 月までの、約 2 万 5 千ユーザの約 150 万件のデータを解析対象とした。そして、SBM を利用するユーザをノードに、同じ Web ページをブックマークしているユーザ間の関係をリンクとしたグラフから成る系列を定義した。系列中の各グラフを作成する期間を 1 週間とし、作成したグラフ系列を図 1 に示す。コミュニティ抽出には *CNM(Clauset-Newman-Moore)* 法 [Clauset 04] を用い、コミュニティを同定する際の確信度の閾値を 0.9 に設定し、コミュニティの変化を抽出した。

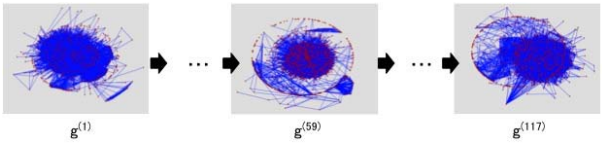

図 1: 解析対象のグラフ系列

3.2 実験結果とその考察

グラフ系列から抽出されたコミュニティの総数は 4,220 であり、これらのコミュニティの変化過程を解析した。探索されたノードの変遷総数は 11,540 であり、設定した閾値 0.9 以上に該当する変遷数は全体の 24% にあたる、2,777 であった。この変遷情報を用いて、コミュニティを同定し、その変化を抽出した。図 2 は、抽出したコミュニティの変化を表している。横軸が、生成から消滅するまでの変化過程の数を表し、縦軸がそれに該当する変化のシナリオ数を対数を用いて表している。まず、注目すべき点として、抽出した約 90% の変化パターンが生成後、その次のグラフで消滅しているという点である。これは、グラフ系列中で連続する二つのグラフ間において、その大部分が変化するという、解析対象のデータの性質がコミュニティの変化に現れているのではないかと考えられる。次に注目すべき点として、変化数が 2, 3 のいくつかの変化パターンに

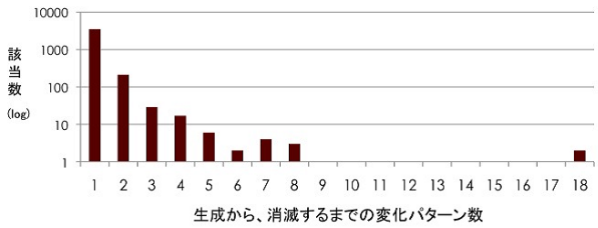

図 2: 抽出したコミュニティの変化数

おいて統合する変化過程が見られた。そして、それらの全てのコミュニティが、統合によって、構成要素が劇的に拡大した。しかし、統合・拡大後、それらのコミュニティは直ぐに消滅した。これは、ホットトピックスに多くのユーザが群がるという、SBM の特性が現れているのではないかと考えられる。最後の注目点として、同じ Web ページを 3ヶ月にわたってブックマークするユーザコミュニティが抽出された。抽出されたコミュニティは極めて小規模のものであったが、確信度の閾値を高く設定したことによって発見することができたユーザ間の関係ではないかと考えられる。

4. 結言

本稿では、データストリームに潜む変化のシナリオを見つけるための手法を提案した。そして、SBM を用いた評価実験において、確信度の閾値を高く設定したことによって、長期間にわたって、同じ Web ページをブックマークしているユーザコミュニティを抽出することができた。今後の展開として、今回報告できなかった周辺コミュニティを対象とした同定、および、l を変化させることによって得られる結果の分析を行うと共に、異なる確信度の閾値で抽出される変化パターンの解析や、コミュニティの変化に強い影響を与えるような外的要因の解析を考えている。

参考文献

[Yamaguchi 09] Takehiro Yamaguchi, Ayahiko Niimi: Community Graph Sequence with Sequence Data of Network Structured Data. 5th International Workshop on Computational Intelligence and Applications 2009 (IWCIA 2009), Hiroshima, Hiroshima, Japan, IEEE Systems, Man & Cybernetics Society, ISSN 1883-3977: pp.196–201, 2009.

[有村 05] 有村博紀: 大規模データストリームのためのマイニング技術の動向, 電子情報通信学会論文誌, D-I J88-D-I(3), pp.563-575, 2005.

[豊田 03] 豊田正史, 喜連川優: 日本におけるウェブコミュニティの発展過程, 日本データベース学会 letters Vol.2, No.1, pp.35-38, 2003.

[大和田 08] 大和田純, 吉井伸一郎, 古川正志: 成長ネットワークにおけるコミュニティ構造推移の観察, 情報処理学会論文誌, Vol.49, No.2, pp.765-773, 2008.

[湯田 08] 湯田聰夫: コミュニティ抽出法とその展望, オペレーションズ・リサーチ:経営の科学, Vol.53, No.9, pp.529-535, 2008.

[Clauset 04] A.Clauset, M.E.J. Newman, and C. Moore: Finding community structure in very large networks, Physical Review E, Vol.70, p.066111, 2004.

[livedoor clip] livedoor clip, http://clip.livedoor.com/

[EDGE] EDGE Datasets, http://labs.edge.jp/datasets/