科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１６　日現在

機関番号：　３２６０１
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００１８１

研究課題名（和文）大規模社会ネットワークにおける情報伝搬履歴に基づく多重コミュニティ抽出

研究課題名（英文）Extraction of multiple communities based on information diffusion results on a large social network

研究代表者
大原　剛三（Ohara, Kouzou）

青山学院大学・理工学部・准教授

研究者番号：３０２９４１２７

交付決定額（研究期間全体）：（直接経費）　　3,200,000円、（間接経費）　　960,000円

研究成果の概要（和文）：本研究では，代表的なマイクロブログサービスであるTwitter上の社会ネットワークを対象に，タグ情報，投稿記事中の特徴語，トピック推定の結果を利用した情報拡散ネットワーク抽出，およびグラフマイニング技術を利用した情報拡散ネットワークの融合のための方法を提案し，部分的に重複する興味の対象が異なる複数のコミュニティの融合体（多重コミュニティ）を抽出した．また，その特徴分析のために，情報拡散時の速度が変化する点を精度よく検出する手法を提案した．

研究成果の概要（英文）：In this work, we extracted multiple communities that overlap with each other, but have interest in different things from a large social network on Twitter, a notable microblogging service. To this end, we first extracted information diffusion networks by means of tags and characteristic keywords in articles, as well as results of a topic estimation method (LDA: Latent Dirichlet Allocation). Then, those networks are integrated by means of a graph mining technique that can find frequent patterns from multiple graphs. Namely, information diffusion networks are integrated if they share common substructures whose frequency is equal to or greater than a given threshold. Furthermore, we devised a method of accurately detecting change points in information diffusion sequences around which diffusion speed has changed in order to investigate the features of resulting communities.

研究分野：　総合領域

科研費の分科・細目：　情報学・知能情報学

キーワード：　社会ネットワーク　コミュニティ抽出　グラフマイニング　データマイニング　情報工学

１．研究開始当初の背景

近年，インターネットアクセスの高速化やパーソナルコンピュータ，携帯電話・スマートフォンの高機能化などを背景に，Mixi や Twitter に代表されるソーシャルネットワーキングサービス（SNS）が急速に普及している．特に，Twitter のようなマイクロブログは記事の長さが制限される反面，手軽に記事を投稿できることから，コミュニケーションツールとして爆発的に普及しており，新たなビジネスモデル構築基盤として社会への影響も大きくなっている．実際のところ，このような SNS を介してインターネット上に構成される大規模社会ネットワーク上では，多種多様な情報がいわゆる口コミにより伝播しており，ある種の情報をより多くの人に伝える，もしくは風評被害などを迅速に阻止するために，ネットワーク構成員の情報発信者としての影響力を分析する技術への要求は高まっている．

従来，そのような社会ネットワークの分析に関する研究は社会学の研究者を中心になされてきた．しかし，近年のインターネット上の社会ネットワークの普及は，大規模なネットワークとその上での豊富な情報伝播記録の取得を可能にし，その結果，情報工学を始めとする様々な分野の研究者が社会ネットワーク上の情報伝播分析に関する研究に参入するに至っている．研究代表者自身も，これまでに共同研究者と共に数理モデルを用いた影響度解析などを行ってきた．

一方，SNS を介して構成される利用者間の大規模参照関係ネットワークは１つの巨大な社会ネットワークを構成していると同時に，それぞれ異なる対人関係に基づく大小様々な数多くのコミュニティの融合体でもあることに注意を払う必要がある．実際，研究代表者は，これまでに携わった社会ネットワーク上の影響度解析に関する研究を通して，コミュニティごとに流通するトピックの性質が異なる場合があり，またその内容によって伝播速度や影響範囲も異なり得ることを確認している．すなわち，社会ネットワーク上の影響度をより正確に分析するためには，そこに内在するコミュニティを把握しておくことが重要であるといえる．

２．研究の目的

以上のような背景から，本研究では，SNS 上で展開される大規模社会ネットワークを対象に，そこから部分的に重複し得る複数のコミュニティ（多重コミュニティ）を抽出し，その特徴を分析する技術を確立することを目的とする．具体的には，代表的なマイクロブログである Twitter を対象に，そこから得られる表層的な利用者間相互参照関係ネットワークを１つの巨大な社会ネットワークとみなし，そこからのコミュニティ抽出を試みる．その為に，本研究では，実際に情報が伝わった時系列情報（情報拡散系列）に基づきネットワーク全体から切り出した部分ネットワーク（情報拡散ネットワーク）が特定の興味を共有するコミュニティの一部に相当することに着目する．言い換えるなら，十分な量の情報拡散ネットワークを正確に切り出すことができれば，それらを融合することで多重性を考慮しつつ，社会ネットワーク中のコミュニティ構造を再現することが可能となると考えられる．

本研究では，以上の点を踏まえ具体的には以下の技術目標の達成を目指す．

（1）投稿記事集合からの情報拡散系列抽出法の実現

（2）コミュニティ構造を再現する情報拡散ネットワーク融合法の実現

（3）抽出したコミュニティの特徴解析法の実現

３．研究の方法

上記 3 つの技術目標のうち（1）に関しては，SNS からは任意の会員が投稿した記事の単なる時系列情報のみが得られるため，そこからネットワーク中の参照関係を考慮した実際の情報拡散系列を自動抽出することは必ずしも容易ではない．そのため，本研究では最初のステップとして，対象とする Twitter の特性を考慮し，投稿記事中に現れる特定のタグ情報を手掛かりとして情報伝搬系列を再現し，その次のステップとして伝搬する情報コンテンツ自身に基づいたより汎用性の高い情報伝搬系列抽出法の実現を目指す．その上で，2 番目のステップで得られるコンテンツ解析結果も利用して，上記技術目標(3)の達成を目指す．

一方，技術目標（2）に関しては，個々の情報拡散ネットワークの共通性を考慮するために，グラフマイニングにおける多頻度共通部分グラフ列挙アルゴリズムを応用した効率的な手法の実現を目指す．

４．研究成果

<u>（1）投稿記事集合からの情報拡散系列抽出法の実現に関する成果</u>

本研究で対象とした代表的なマイクロブログである Twitter では，特定のトピックに関する記事を投稿する場合にはそのトピックを表すキーワードの先頭に記号#を付したハッシュタグが用いられる．また，他人の投稿をそのまま転送形式で利用する場合には，記号“RT”などのタグを投稿記事中に記述する慣習がある．これらのタグを手掛かりとしつつ，さらに，文中の潜在トピックを推定する技術である潜在的ディリクレ配分法（LDA: Latent Dirichlet Allocation）を適用して得られた潜在トピックを利用し，トピックごとに投稿記事の時系列リストである情報拡散系列を抽出し，それらと相互参照関係ネットワーク中のユーザ間の接続関係から情報拡散ネットワークを抽出した．“RT”を投稿記事の連鎖の指標とし，ハッシュタグ，ツイ

ート中の特徴語，潜在トピックをトピック同定に用いた．特徴語に関しては，単語 2-gram まで利用している．実際に収集した約 94,000 ユーザによる利用者間相互参照関係ネットワークを対象に実験したところ，一定の精度で情報拡散系列の抽出が可能であることを確認した．その一方で，その系列長は比較的短いものが多くなる傾向が見られた．これは，実際の情報拡散では，単一記事の連鎖的投稿は一部の反響の大きい記事を除いて，ごく小さいコミュニティ内でとどまることを示している．ただし，投稿記事に対するトピック推定精度が必ずしも十分ではなかったことにも起因しており，その点に関しては今後の改善が必要である．

これらの情報拡散系列の抽出技術は，情報拡散現象の精緻な分析には必須であり，その意味においてここで得られた知見は重要である．

（2）コミュニティ構造を再現する情報拡散ネットワーク融合法の実現に関する成果

複数のグラフ構造から頻出する共通部分グラフを抽出するグラフマイニング手法を利用し，抽出した情報拡散ネットワークを融合することで，共通する興味をもつ人物から構成されるコミュニティ，およびそれらを融合した多重コミュニティを抽出するシステムを実現した．情報拡散系列から抽出される個々のコミュニティは比較的小規模であるものの，幾つかは重複するノードを有する．そのため，まず類似トピックに関して抽出された小規模な情報拡散ネットワークのうち，低い頻度閾値の下で共通部分グラフをもつもの同士を融合することで，同一対象に興味をもつ一定サイズのネットワークを抽出できることを確認した．そして，その結果得られたネットワークに対して，より高い頻度閾値を設定して共通部分グラフを抽出し，それを共有するネットワークを融合することで，興味の対象が異なるコミュニティが一定の強さで融合している多重コミュニティが抽出できることを確認した．

情報拡散現象の分析では，コミュニティを 1 つの単位として考えることは重要であり，ここでの成果はその分析に資するものである．今後は，他の既存手法で抽出されるコミュニティ構造との詳細な比較を進める必要がある．

（3）抽出したコミュニティの特徴解析法の実現に関する成果

これまでに実装した手法により抽出されたコミュニティ内のツイートに対して，LDA を適用し，コミュニティにおける複数の興味対象を特徴づける特徴ベクトルを生成し，その内容を分析した．その結果，個々の情報拡散ネットワークを抽出する際に利用したトピックの上位概念（「政治」など）と解釈され得るいくつかの潜在トピックの生起確率が高くなる傾向を確認した．このような多重コミュニティにおける興味の対象の重なり具合は，情報拡散における話題のドリフト（遷移）を解析する上で重要な役割を果たす．ただし，LDA の適用対象となる単語の選定により結果が影響を受けるため，その選定が重要となる．本研究では，いわゆる崩れた表記が多用される SNS 上のコンテンツから意味のある単語群を選定するための多くの知見を得ることができた．

一方，タグ情報に基づいて抽出した情報拡散系列を対象とした変化点検出にも取り組み，情報拡散過程の特徴分析を行った．ここでは，情報拡散現象をモデル化する確率モデルを仮定し，観測した情報拡散系列から，情報の拡散速度を規定するモデルパラメータの値が変化する変化点を精度よく抽出する方法を提案した．提案手法を東日本大震災前後のツイートデータに適用した結果，放送局などの公共アカウントが発信する情報は大きなイベントの前後でも拡散されやすい傾向があること，個人の発信する情報は通常時はごく身近なコミュニティ内でしか拡散されないが地震などの大きなイベントに関するものはその範囲を超えて拡散され得ること，検出した変化点付近の投稿記事内容は同一のものが集中することからバースト的に拡散したトピックを効率よく特定できることなどを確認した．このような変化点検出技術は，コミュニティ構造と合わせて利用することにより，より精緻な情報拡散現象の分析を可能とするものである．今後は，前提としている情報拡散モデルが実際の情報拡散現象をどれだけよくモデル化できているかの検証と，その結果に応じたモデルの洗練が必要である．

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）

〔雑誌論文〕（計 1 件）

① 大原剛三，斉藤和己，木村昌弘，元田浩，情報拡散モデルに基づくツイート系列からのバースト期間検出，日本データベース学会論文誌，査読有，Vol.11，No.2，2012，pp.25-30

〔学会発表〕（計 6 件）

① Kazumi Saito, Kouzou Ohara, Masahiro Kimura, and Hiroshi Motoda, Detecting Changes in Content and Posting Time Distributions in Social Media, the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), 2013 年 8 月 28 日，カナダ・ナイアガラフォールズ

② 大原 剛三，斉藤 和巳，木村 昌弘，元田浩，Twitter 上の情報拡散系列からの変化点検出，第 27 回人工知能学会全国大会

（JSAI2013），2013 年 6 月 6 日，富山国際会議場

③ 大原 剛三，斉藤 和巳，木村 昌弘，元田 浩，社会ネットワークの構造的特徴量と情報拡散モデルにおける期待影響度の関係について，人工知能学会第 97 回知識ベースシステム研究会（SIG-KBS），2012 年 11 月 15 日，慶應大学日吉キャンパス来往舎

６．研究組織

(1)研究代表者

大原　剛三　（OHARA, Kouzou）
青山学院大学・理工学部・准教授
研究者番号：30294127