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

平成　２９　年　　６　月　１２　日現在

機関番号：　１４７０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０３４５

研究課題名（和文）ソーシャルメディアにおける時空間コミュニティの抽出

研究課題名（英文）Temporal-Spatial Community Extraction in Social Media

研究代表者
風間　一洋（Kazama, Kazuhiro）

和歌山大学・システム工学部・教授

研究者番号：６０６４７２０４

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：ソーシャルメディア上のユーザのコミュニケーションによって生まれる議論話題を検出，分析する手法を研究した．
まず，時系列情報を伴うユーザ間のコミュニケーションのバーストを検出するために，Twitter上のリプライ関係から時間の閾値を設けて作成した双対グラフから，コア部抽出を行う双対SR 法を提案した．
さらに，ツイート空間における議論話題の時系列変化を分析するために，トピックモデルであるLDAで，単語ではなく単語の関係性（単語2-gram）のbag-of-wordsをトピックごとに分類することで，時系列変化を表すトピック系列と，トピック内の単語の関係を表すトピックグラフの可視化手法を提案した．

研究成果の概要（英文）：We study methods to detect and analyze discussion topics created by user communication on social media..
In order to detect temporal-spatial bursts in user communication, we propose dual-graph SR (Spectral Relaxation) method to extract core potions of a dual graph, which is created from replies on Twitter with specified time interval threshold.
Furthermore, we propose a method to classify topics for bag-of-words, which is created with word 2-grams instead of words, by using a generative topic model such as latent Dirichlet allocation (LDA) and visualize the results as topic sequences, which means time series variation of topics, or a topic graph, which means word relationships in each topic.

研究分野：　Webマイニング

キーワード：　時空間コミュニティ　コミュニティ抽出　ネットワーク分析　トピックモデル　Twitter

1．研究開始当初の背景

Twitter，Facebook，LinkedIn などのサービスの普及により，そのようなオンラインサービス上に構築したソーシャルネットワークを経由して，多種多様な最新情報を効率良く送受信できるようになった．
このような情報交換は，1) 各ユーザが非同期におこなう個人的なつぶやき，2) 現実世界や仮想世界のイベントと連動して生じる一過性の盛り上がり，3) 多数のユーザ間が相互にコミュニケーションすることで生み出される**議論話題**（Discussion topics）の 3 種類に大きく分類できる．

Twitter について考えると，1)が一番量が多いが，社会への影響は少ない．

2)は TV 番組のクライマックスや地震に関して多くのユーザが同期して発言したり，あるユーザの発言をリツイートを用いてフォロワーに転送するもので，瞬間的に大量のツイートを誘発するものの，発言内容がある実世界のイベントに簡単に言及しているだけのことが多い．さらに，情報を充分精査せずに転送するので，デマ情報も多く含まれる．

これに対して，3)は，主に全体の約 2～4 割を占めるリプライで交換・伝播される．Twitter のリプライは発言元と発言先のフォロワーも閲覧できることから，フォローネットワーク上の 2 次近傍ユーザも含む緩やかな情報拡散を引き起こし，インターネット上の合意形成や口コミの拡散などの重要な社会現象に関係する．

例えば，ビッグデータのプライバシー保護の問題に関して，Twitter 上で関連分野の専門家が直接話し合って問題点や改善策について議論した結果を受けて企業側が方針を変更するような，ソーシャルメディア上の**議論話題**が実世界に影響を与え，変える動きが実際に起こっている．

このような創造的な活動は，記者が専門家に個別にインタビューした意見を集約・整理して一方的に読者に伝える形態のマスメディアには存在し得ないことから，情報学に留まらず，経営学，社会科学，さらには心理学などの学術分野においても重要な研究対象である．

2．研究の目的

本研究では，多くのユーザ間で大規模に展開される主要な議論話題の検出，議論話題の内容の分析，さらに議論話題に関する影響拡散のモデリングに関する手法の確立を課題とする．

議論話題の検出は，Twitter の大規模なソーシャルネットワーク上のメッセージ交換により，ある期間だけ形成される時空間コミュニティの抽出問題として定式化できる．

Twitter における議論は，ネットワーク上で近いユーザの間で行われ，情報伝播の進行と共に範囲が拡大し，特にその一部で濃密な議論が行われ，やがて終了する．なお，このような議論は Twitter 上で同時多発的に発生するが，ネットワーク上で遠いユーザ群の議論は，必ずしも最終的に融合するとは限らない．つまり，このような**議論話題**の抽出では，時間と空間の両方を考慮しなければならない．

時空間コミュニティの空間的な側面に限れば，ネットワーク構造からのコミュニティ抽出法に関する研究が存在する．例えば，Clauset らの CNM 法では，ネットワーク構造を粗な部分のエッジを切断して複数のノード集合に分割する．また，Palla らの k-clique community 法では，ノード同士が密結合したクリーク（clique）から構成されるサブネットワークを抽出する．ただし，これらの手法は単純無向グラフしか処理できず，コミュニケーションの密度や時間のような概念は存在しない．

時間的な側面に限れば，テキストストリームからのバースト検出の研究が存在する．例えば，Kleinberg は，テキストストリームを無限状態オートマトンでモデル化し，電子メールのやりとりにおいて注目する単語の出現頻度が急増する期間を抽出する手法を提案した．また，Deepayan らは，Twitter におけるスポーツに関するサブイベントの移り変わりを，隠れマルコフモデルを用いて抽出した．ただし，これらの研究では巨大な人間関係における位置は完全に無視されるので，ソーシャルネットワーク上で独立に多数発生する**議論話題**を個別に検出することはできない上に，生成過程を個別に分析できない．

本研究では，Twitter のユーザ間のタイムスタンプ付きのコミュニケーション履歴から，情報拡散や意見形成に関する確率的な数理モデルとネットワーク上で密結合するコア部抽出法を有機的に結合することで，時空間コミュニティを抽出する技術を確立する．また，抽出された時空間コミュニティに含まれるツイートの内容を分析することで，重要な**議論話題**の生成過程や変遷し，さらに複数の時空間コミュニティ間の関連を分析する技術を確立する．さらに，時空間コミュニティを内容と共に可視化し，重要な**議論話題**の進行を概観することを可能にする．

最後に，本研究で構築するモデルや手法の評価には，機械学習やデータマイニング分野などで標準的に採用されている，コミュニケーションがネットワーク上で将来的にどのように展開するかなどの予測性能を重要な指標として用いる．

ただし，このような定量的な評価指標だけでなく，知識発見の観点より，抽出する議論話題の重要性とともに，ユーザ間の影響拡散モデルに対する解釈の可能性についても積極的に評価する点も特色とする．

3．研究の方法

(1) 情報拡散の数理モデルとネットワーク

構造からのコア部抽出の融合

時空間コミュニティ抽出法を，情報拡散や意見形成に関する確率的な数理モデルとネットワーク上で密結合するコア部抽出法を用いて実現する．

まず，ユーザu_iからv_iに時刻t_iにコンテンツc_iで送信したメッセージ$m_i = (u_i, v_i, c_i, t_i)$の集合を$M$とした時の 2 つのメッセージ$m_i, m_j \in M$の関係を，我々が提案した情報拡散モデルを用いて定量化する．

いま，$v_i = u_j$かつ$t_i < t_j$の条件下における情報拡散の成功確率$p(m_i, m_j)$は，コンテンツ類似度$q(c_i, c_j)$と情報伝達時間遅れ$r(t_i, t_j)$の積により，次のように定義する．

$$p(m_i, m_j) \propto q(c_i, c_j) \times r(t_i, t_j)$$

最初は，コンテンツ類似度$q(c_i, c_j)$としては，一般的なc_i, c_jの特徴ベクトルのコサイン類似度を，情報伝達時間遅れ$r(t_i, t_j)$には時間差$t_j - t_i$に対する指数分布などの標準的な確率モデルを用いる．

さらに，一般にユーザをノード，メッセージをエッジとしてモデル化するが，本研究では，メッセージをノード，ユーザをエッジとした双対グラフ，すなわち二つのメッセージm_i, m_j間に上述した情報拡散の成功確率$p(m_i, m_j)$を重みとしてもつエッジで構成されるネットワークを考える．

あるユーザ群の間の活発なコミュニケーションは，大きな重みが付与されたリンクで密結合するサブネットワークとして創発し，このようなネットワークのコア部を時空間コミュニティと見なすことができると考えられる．

コア部抽出には，単純無向グラフからのコア部抽出法である SR(Spectral Relaxation) 法を，隣接行列の固有ベクトル計算とベクトル要素の量子化のアイデアを維持したままで重み付き有向グラフを扱えるように拡張した抽出法を用いて Twitter 上でのユーザ間のリプライメッセージから構成されるネットワークに適用し，幾つかの興味深いコア部を抽出できることを確認している．

この手法を出発点に，より優れた時空間コミュニティ抽出法を探求する．

評価データとしては，2012 年 3 月 14 日から 2013 年 3 月 14 日の期間に Twitter 上でリプライされた 11,500,369 ユーザの 1,649,048,139 ツイートを利用する予定である．

(2) ツイート群における議論話題の変遷の分析手法の開発

時空間コミュニティ抽出が利用できるようになる前段階として，ツイートアーカイブから現実に発生したイベント特有の単語を抽出し，それらがどのように変遷するかを分析する手法を確立する．

これは，我々が以前提案した東日本大震災前後のツイートから震災に関連した用語を抜き出して，その遷移確率から因果関係を求める技術を元に開発する．

ただし，時空間分析の時間的な面だけしか着目していないので，時空間コミュニティが融合した形で抽出されてしまうことから，複数の話題が混在することが考えられるので，LDA や DTM などの潜在的トピック分類法を用いてイベント特有の単語をトピックごとに分類した後に処理することも検討する．

さらに，イベント特有の語と同時にツイート内に共起する共起語の推移を調べ，どのような因果関係で話題が推移したかを分析する．

(3) 時空間コミュニティ抽出法の性能評価・向上

まず，実現した時空間コミュニティ抽出法の総合的な性能を評価する．例えば，機械学習やデータマイニング分野などで標準的に採用されているような，コミュニケーションがネットワーク上で将来的にどのように展開するかなどの予測性能をパラメータ推定の評価指標として用いる．

さらに，予測通りにはいかない実データに対する時空間コミュニティ抽出法の実性能を向上させるために，性能評価によって発見されたボトルネック部分の改良と，細部に関する見直しをおこなう．

まず，コンテンツ類似度としてコサイン類似度，情報伝達時間遅れとして時間差に対する指数分布より優れた結果をもたらす尺度を検討する．

さらに，抽出された時空間コミュニティは情報拡散の成功率に関して固有のパラメータを持つことが想定されるので，より精緻な時空間コミュニティが抽出できるように最尤推定で求めることを試みる．

(4) 時空間コミュニティ内の議論話題の変化の分析

時空間コミュニティ抽出法が利用できるようになり次第，抽出された時空間コミュニティに含まれるツイート群を対象に，前年度に確立した分析技術を用いて時空間コミュニティの議論話題の内容の変遷の分析を試みる．

さらに，時空間コミュニティ抽出法の処理結果の定性的な理解が容易になるように，単一の時空間コミュニティ内の変遷を可視化する技術を開発する．

また，Twitter 上のソーシャルネットワーク上で，2 人のユーザが同一トピックに関してほぼ同時にコミュニケーションを開始したとしても，必ずしも同一の時空間コミュニティとして抽出されるとは限らないと予測される．

そこで，抽出された時空間コミュニティ間の関連性を分析することで，このような現象の発生の有無を実証する．

実際には，時空間コミュニティ間の意味的な関連性は，単語ベクトルのコサイン類似度

などの指標を用いる．

これにより，個々の時空間コミュニティだけでなく，Twitter 空間内における同時多発的な時空間コミュニティの説明付けが可能な分析システムを構築する．

４．研究成果

(1) 双対 SR 法による議論話題の抽出

時系列情報を伴うユーザ間のコミュニケーションのバースト検出を行う手法として，閾値を設けた双対グラフを生成し，MDSR 法の応用によりコア抽出を行う双対 SR 法を提案した．

詳細には，ユーザ集合$U = \{u, v, \ldots\}$について，ユーザ$u \in U$からユーザ$v \in U$へ時刻tに送られたメッセージを$(u, v; t)$と表すと，全メッセージ集合Dは次のように表記できる．

$$D = \{(u_1, v_1; t_1), (u_2, v_2; t_2), \ldots\}$$

以下の分析では，このDの各要素をノードとする．いま，各メッセージの間に，ユーザvを受信者とするメッセージからユーザvを送信者とするメッセージへのリンクが存在すると考え，受信した情報をさらに送信したと判断するための時間の閾値Δtを設けることで，リンク集合$E\big((u, v; t)\big)$を次のように定義する．

$$E\big((u, v; t)\big) = \big\{\big((u, v; t), (v, w; t')\big) \big| t < t' < t + \Delta t\big\}$$

なお，このときのvを特に中心ユーザと呼ぶこととする．

この$E((u, v; t))$をD内の全てのメッセージについて考えることで，全リンク集合$E = \{E((u, v; t)) \mid (u, v; t) \subset D\}$が得られる．

本研究で着目している議論話題は，短い時間間隔内でのコミュニケーションから生まれると推測できるため，あるメッセージを受け取ってから別のメッセージを送信するまでに掛かった時間が充分短いことを保証できるように，閾値Δtを設定する．このようにして得られる双対グラフ$G\Delta t = (D, E)$について，コア抽出手法を適用し，時空間バーストを検出する．

東日本大震災前後における Twitter のリプライをデータセットとした評価実験により，同一の有名アカウントから発せられた異なる話題を，双対 SR 法は正しく分けて抽出することを確認した．また，双対グラフを生成する際の閾値の設定によって，優先的に抽出される中心ユーザに変化が生じることも確認した．

(2) 回遊行動モデルの構築

議論話題変遷モデルの研究進展に伴い，観光地等における回遊者の将来行動予測を妥当な精度で可能にする回遊行動モデルの構築した．具体的には，従来研究と同様，回遊者の行動プロセスが Levy Flight に従うと仮定し，観光地間の距離や人気度などの属性に対するパラメータを導入することで，回遊行動モデルを構築した．本研究の特色として，モデルの解析的特性からではなく，統計的機械学習でパラメータを推定するこれにより，ある程度複雑で多様なパラメータを組み込むことが可能となり，従来研究より柔軟なモデルの構築が可能なことである．
オンライン写真共有サイト Flickr に投稿された位置情報付き写真群を学習データとして用い，回遊者の行動予測性能とパラメータ推定精度の安定性を用いて評価し，提案モデルの妥当性を検証したところ，実際の回遊行動との一致度からモデルの有効性・妥当性を確認出来た．

(3) LDA を用いた単語 2-gram による議論話題の時系列変化の分析手法

ツイート空間における議論話題の時系列変化を分析する手法に関しては，トピックモデルである LDA により，単語ではなく単語の関係性（単語 2-gram）を分類することで，トピック間の時系列変化と，特定のトピックの内容をトピックグラフとしてわかりやすく可視化する手法を提案した．

既存研究においても，同様のシステムが試作されていたが，新聞記事などの整ったデータや，小規模のツイートデータしか扱っていないことが多かったが，本研究の相違点は (1)現実の大規模データに適用可能であること．LDA によるトピック抽出を高速に実施できるように，並列処理可能な PLDA を用いた．(2) スペルミスなどのノイズが大量に含まれる実データを処理可能にしていること．LDA のような確率ベースの手法では，これらはノイズとなり抽出結果を著しく悪化させることが多いが，多面的なノイズ低減手段を用いることで，妥当な結果が得られるように抽出性能を改善している．(3) 単語の関係性を対象にすることで，新語としての複合語が辞書に存在しなくても抽出可能になったと共に，トピックの内容をわかりやすく簡略に表示することが可能になった．

(4)ジオタグ付きツイートからの複数ユーザによる共同利用経路の抽出手法

研究の進展と共に，ツイート空間の分析を実空間と結びつけて，実空間上のコア部抽出をおこなう試みとして，ツイートに付与されているジオタグ（位置情報）を用いて，実空間上に展開されている，複数のユーザが同じ目的のために集団行動しているような経路を自動的に発見する手法を提案した．

例えば，既存研究においては，GPS を対象に装着させることで，短い間隔で細粒度の経路を測定できることを前提として作られており，例えば東日本大震災時に Google が公開した，震災直後に通行可能な道路を示す「通れたマップ」では，普段から観測用に用いているプローブカー（GPS と通信機能を持つ車）のデータを使用したが，ツイートデータを用いるだけで可能になるとしたら，特別な機器を用いずに，さまざまな応用が可能になると考えられる．

ただし，単一ユーザのツイート数は散発的であり，それだけでは経路の予測は困難であることから，複数のユーザによって生み出される人の動きをコア部として抽出することを試みた．

実際には，実空間をグリッド状に分割して，画像認識技術の一種である Hough 変換を応用することで抽出した断片的なエッジを，複数レベルの接続・再接続と，補完・スムージングを行うことで，電車路線などの公共交通機関の経路が少量のデータでも比較的高い精度で抽出できることを示すと共に，例えば桜の開花時期においては，大阪造幣局の（桜の）通り抜けなどの，ユーザが敷地内の桜並木の鑑賞という同一の目的で通過する経路の抽出が可能であることを示した．

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）

〔雑誌論文〕（計１件）

1，谷 直樹，風間 一洋，榊 剛史，吉田 光男，斉藤 和巳：ジオタグ付きツイートを用いた交通路の抽出法，情報処理学会論文誌：データベース，査読有，Vol．10，No．2，2017．

〔学会発表〕（計２２件）

1，久保 侑哉，風間 一洋，鳥海 不二夫，斉藤 和巳：東日本大震災時のツイートの単語 2-gram に基づくトピックの可視化，第 31 回人工知能学会全国大会，愛知県名古屋市，3P2-NFC-00b-1in1，2017．

2，鈴木 優伽，斉藤 和巳，風間 一洋：群れモデルに基づく意見形成ダイナミクスの分析，第 15 回情報科学技術フォーラム（FIT2016），富山県富山市，2016．

3，岩崎 一輝，鈴木 優伽，斉藤 和巳：ユーザ行動データによるサンプリング計算解の評価，第 15 回情報科学技術フォーラム（FIT2016），富山県富山市，2016．

4，白澤 穂香，斉藤 和巳：ノード間の実距離に基づく近接中心性と媒介中心性の特性評価，第 15 回情報科学技術フォーラム（FIT2016），富山県富山市，2016．

5，鈴木 優伽，斉藤 和巳，風間 一洋：群れモデルに基づくオピニオン形成の変化点検出，第 13 回 ネットワーク生態学シンポジウム（NETECO2016），千葉県木更津市，2016．

6，鈴木 優伽，斉藤 和巳，風間 一洋：スポット情報を考慮した複合データ分類モデル，第 30 回人工知能学会全国大会（JSAI2016），福岡県北九州市，2016．

7，鈴木 優伽，斉藤 和巳，風間 一洋：スポット情報組み込みモデルによる回遊行動データの分類，情報処理学会第 78 回全国大会，神奈川県横浜市，2016．

8，鈴木 優伽，斉藤 和巳，風間 一洋：重み付き回遊行動中心性による抽出スポットの安定性評価，第 2 回とうかい観光情報学研究会，愛知県名古屋市，2016．

9，小林 えり，中里 主哉，斉藤 和巳，風間 一洋，吉田 光男：観光イベントにおける位置情報ツイートによるユーザ行動分析，第 2 回とうかい観光情報学研究会，愛知県名古屋市，2016．

10，鈴木 優伽，斉藤 和巳，風間 一洋：回遊行動中心性によるスポット抽出，研究報告データベースシステム（DBS），Vol．2015-DBS-162，No．17，pp．1-8，東京都江東区，2015．

11，鈴木 優伽，斉藤 和巳，風間 一洋：最尤推定による回遊行動モデリング，第 14 回情報科学技術フォーラム（FIT2015），F-005，愛媛県松山市，2015．

12，鈴木 優伽，斉藤 和巳，風間 一洋：最尤推定にもとづく回遊行動統計モデリング，第 11 回ネットワークが創発する知能研究会（JWEIN2015），東京都千代田区，2015．

13，鈴木 優伽，斉藤 和巳，風間 一洋：分割データによる回遊行動変化の検出，第 12 回ネットワーク生態学シンポジウム， 静岡県伊東市，2015．

14，北田 剛士，風間 一洋，榊 剛史，鳥海 不二夫，栗原 聡，篠田 孝祐，野田 五十樹，斉藤 和巳：東日本大震災時のツイートのトピック系列の可視化と分析，第 29 回人工知能学会全国大会，2B3-NFC-02a-1，北海道函館市，2015．

15，鈴木 優伽，斉藤 和巳：ハブ・オーソリティモデルによる主要スポット代表ユーザ抽出法，第 29 回人工知能学会全国大会，4C1-4，北海道函館市，2015．

16，谷 直樹，風間 一洋，榊 剛史，吉田 光男，斉藤 和巳：移動速度条件を考慮したジオタグ付きツイートからの交通路の抽出と分析，ARG 第 4 回 Web インテリジェンスとインタラクション研究会，WI2-2014-04，大阪府豊中市，2015．

17，鈴木 優伽，伏見 卓恭，斉藤 和巳，風間 一洋：回遊行動モデルに基づく重要観光スポット抽出法，情報処理学会第 77 回全国大会，1R-03，京都府京都市，2015．

18，北田 剛士，風間 一洋，榊 剛史，鳥海 不二夫，栗原 聡，篠田 孝祐，野田 五十樹，斉藤 和巳：Twitter のトピック変遷の可視化法の提案，第 7 回データ工学と情報マネジメントに関するフォーラム（DEIM2015），E2-6，福島県郡山市，2015．

19，加藤 翔子，斉藤 和巳，風間 一洋：固有ベクトル法による類似文書抽出，第 6 回テキストマイニング・シンポジウム，信学技報 114(444)，pp．11-16，電子情報通信学会，大阪府大阪市，2015．

20，藤野 まり菜，加藤 翔子，斉藤 和巳，風間 一洋：多項分布変化点検出法による Twitter 上のユーザ動向分析，第 13 回情報科学技術フォーラム（FIT2014），F-014，茨城県つくば市，2014．

21，加藤 翔子，斉藤 和巳，風間 一洋：双対 SR 法による Twitter データの時空間バースト検出，第 13 回情報科学技術フォーラム (FIT2014)，F-013，茨城県つくば市，2014.
22，加藤 翔子，斉藤 和巳，風間 一洋：双対 SR 法による時空間バースト検出，第 10 回ネットワークが創発する知能研究会 (JWEIN 2014)，東京都調布市，2014.

〔図書〕（計０件）

〔産業財産権〕

○出願状況（計０件）

○取得状況（計０件）

〔その他〕

６．研究組織

(1)研究代表者

風間　一洋（KAZAMA，Kazuhiro）
和歌山大学・システム工学部・教授
研究者番号：６０６４７２０４

(2)研究分担者

斉藤　和巳（SAITO，Kazumi）
静岡県立大学・経営情報学科・教授
研究者番号：８０３７９５４４

(3)連携研究者

(4)研究協力者