時系列テキストにおける クエリ依存の局所的なEmerging Topic の抽出

遠藤　結城† 　戸田　浩之† 　鷲崎　誠司††

† 日本電信電話株式会社 NTT サービスエボリューション研究所　〒 239-0847 神奈川県横須賀市光の丘 1-1
†† 日本電信電話株式会社 NTT メディアインテリジェンス研究所　〒 239-0847 神奈川県横須賀市光の丘 1-1
E-mail: †{endo.yuki,toda.hiroyuki,suzaki.seiji}@lab.ntt.co.jp

あらまし　ソーシャルメディアにおける時系列テキストを分析することで, 世の中で盛り上がっている話題 (Emerging topic) を特定する研究が多く行われている. しかし既存手法の多くは, 与えられたデータセットにおいて大域的な Emerging topic を特定することが前提であるため, 任意の情報と関連する局所的な Emerging topic を特定するのが困難であった. この問題を解決するために, 本研究では関連文書の絞り込みと Non-negative Matrix Factorization (NMF) にもとづく Emerging topic 抽出の枠組みを提案する. 提案手法では膨大なテキストストリームにおいて, クエリをもとに Coverage 優先で関連文書を高速に絞り込む. 広く絞り込まれたデータにおいてはノイズとなる非関連文書が存在しうるため, クエリと関連性の高い Emerging topic を抽出するために, 従来の時間依存 NMF をさらにクエリ依存に拡張する. 評価実験を通して, 提案手法がベースラインと比較して効率的かつ高精度に局所的な Emerging topic を抽出できることを示す.

キーワード　Emerging topic, ローカルクラスタリング, トピックモデル, NMF, クエリ依存

1. はじめに

IT 技術の目覚ましい進歩は, 我々の社会に情報の爆発的な増加をもたらしている. そのような状況の中で, 多くの人々が多様かつ膨大な情報を発信できるソーシャルメディアが注目を集めている. Twitter や Facebook に代表されるソーシャルメディアにおいては, 人々が今何を感じているか, 何をしているかといった情報が時々刻々と発信されている. ソーシャルメディアにおける情報は時間の経過と共に変化していくことから, 我々の社会の状況を反映するある種の指標と捉えることができる. このソーシャルメディアを分析することにより, 企業の評判や広告の影響を評価するなどマーケティングへの活用が注目されている. さらには, ソーシャルメディアにおけるトレンドトピックにもとづいて関連映像を推薦する Trend-aware recommendation [17] が提案されるなど, ソーシャルメディアの分析は幅広いコンテンツの流通や個人向けサービスへの応用も期待されている.

ソーシャルメディアの分析を効果的な応用につなげるには, ユーザが知りたい情報をすぐに分析できることが重要である. 例えば企業のマーケターは, 自身の企業や商品の評判は言うまでもなく, 自身の企業と関連する周辺の企業や商品の評判について可能な限り鮮度の高い情報を得て, 比較検討することで戦略を立案したいと考える. また Trend-aware recommendation などへの応用を考えてもユーザの好みは多様であることから, スポーツやアニメなどユーザの好みにもとづいた任意の情報と関連する世の中の話題を分析できることが望まれる. つまり, ソーシャルメディアにおける情報を広く浅く分析するだけでなく, 狭く深く分析することで応用先が広がると考えられる.

ソーシャルメディアにおける情報の中でも, 社会の状況を強く反映するものとして, 今多くのユーザが注目し盛り上がっている話題 (Emerging topic) があげられる. この Emerging topic を分析する研究がこれまで多く取り組まれてきた [6], [12], [15], [20], [23], [25], [28]. これらの手法は時系列テキストを入力することで, 単語の集合として表現された Emerging topic を推定することができる. しかしながら, 与えられたデータセットにおいて大域的な Emerging topic を特定することが前提であるため, 任意の情報と関連する局所的な Emerging topic を特定するのが困難であるという問題があった. そのため, 多様な情報が発信されているソーシャルメディアからは, ユーザが望む Emerging topic を得られるとは限らない. 例えばソーシャルメディア全体において, 政治に関する話題の盛り上がりが支配的である場合, マーケターや消費者が特定の商品や TV の放送内容などと関連する Emerging topic を知りたくとも, これらの手法では政治に関するトピックに他のトピックが埋もれてしまう. 加えてトピックモデルに代表されるこれら

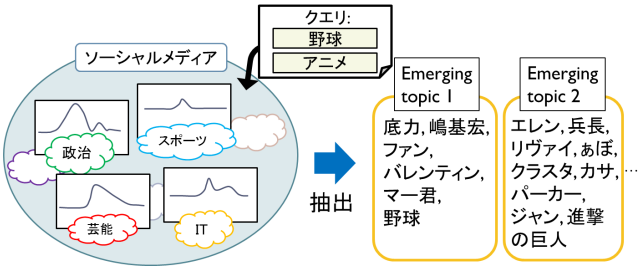

図 1　クエリ依存の局所的な Emerging topic の抽出例. 従来の手法では多様な話題が発信されているソーシャルメディアから, ユーザが望む Emerging topic を得られるとは限らない. これに対し提案手法ではクエリ依存で, 特定の情報と関連性の近い Emerging topic を精度良く抽出する.

トピック抽出の手法は, データ量に依存して大きな処理時間を要するため, 大規模な計算機環境を利用できなければ, 膨大なソーシャルメディアから鮮度の高い情報を抽出するのは難しい. これに対し Chen ら [5] はソーシャルメディアから, 企業や大学など特定の組織に関する Emerging topic を検出する手法を提案した. しかし, 組織に関するテキストを集めるための分類器を構築する際に, 組織のアカウント情報などを用いた正解データを用意する必要があることに加え, 分類器の構築にも時間を要する. 理想的には多様な情報に関する Emerging topic を効率的に特定できる, より汎用的なアプローチが期待される.

そこで本稿では図 1 に示すように, 任意の情報と関連する Emerging topic を効率良く抽出する枠組みを提案する. 前述のとおり関連研究の多くのアプローチは, 特定の情報に関する Emerging topic の抽出を想定していないことに加え, データ量に依存して多くの時間を要するという問題があった. これに対し本研究のアプローチは, まず膨大な文書データから関連文書を集めることでデータ量を削減してから, 削減したデータに対してトピックモデルを適用し Emerging topic を抽出する 2 段階のステップからなる. 1 段階目のステップにおいては, データ量にそれほど依存しない方法を用いることで高速に計算する. この際, 汎用性や効率性の観点から分類器は用いずに, ユーザが指定するキーワードの集合からなるクエリのみにもとづく局所的なクラスタリングにより, Coverage 優先で関連文書を網羅的に収集する. 1 段階目のステップは関連文書を広く高速に収集できる一方で, 収集された文書にクエリと無関係のノイズとなる文書が混在する可能性がある. そこで 2 段階目のステップにおけるトピックモデルとして, 新たな制約の付与による拡張が容易な Non-negative Matrix Factorization (NMF) 採用し, これにもとづく従来の時間依存 NMF をさらにクエリ依存に拡張する. これにより, クエリと関連性の高い Emerging topic をより精度良く抽出する.

本研究の貢献をまとめると以下のようになる.

- クエリを用いた特定の情報に関する文書データの収集と Emerging topic 抽出のフレームワークの提案
- クエリ依存および時間依存 NMF の提案
- 特定の情報と関連する Emerging topic を抽出するタスクについて Twitter データセットを用いた実験による提案手法の有効性の検証

本稿の構成は以下のとおりである. まず 2. 節で関連研究について述べ, 3. 節で提案手法の概観を説明する. 次に 4. 節で関連文書の収集方法について述べ, 5. 節でトピックの抽出手法について述べる. 6. 節では提案手法を評価し, 最後に 7. 節で本稿をまとめる.

2. 関 連 研 究

ソーシャルメディアの時系列テキストにおけるトピック抽出の手法は, マーケティングの分野のみならず地震などの災害検知 [18] や選挙活動の反応予測 [21], 政治や経済の動向調査 [11] など様々な目的で用いられている. 最近では再生数の増加が予測される映像の推薦 [17] や映像配信における効率的なデータ配置 [24] など, マルチメディア分野への応用も期待されている. 時系列テキストを分析しトピックを特定するアプローチは, キーワードベース, 確率的なトピックモデルおよび非確率なトピックモデルの三つに大別される.

キーワードベースのアプローチとして, Weng と Lee は [25] 単語発生頻度の時間変動の相関をもとに, 同じように盛り上がっている単語をクラスタリングすることで Emerging topic を検出した. その際ウェーブレット解析を用いて時系列データからノイズとなる高周波成分を除去することで検出精度を向上させた. 彼らの手法では単語単体で盛り上がりを検出しているのに対し, Li ら [15] は Wikipedia のアンカーテキストなどの情報を用いて, 複数で意味を持つ単語の組み合わせ (セグメント) にまとめ, セグメント単位で盛り上がりを検出した.

確率的なトピックモデルとして Probabilistic Latent Semantic Indexing (PLSI) [10] や Latent Dirichlet Allocation (LDA) [2] が知られている. これらを拡張することで時系列文書を扱えるトピックモデルがいくつか提案されている [3], [6], [22], [23]. Blei らの Dynamic Topic Model や [3] や Wang らの Continuous Time Dynamic Topic Models [22] は, マルコフ性にもとづいて時間的に遷移するトピックをモデル化したものである. Wang ら [23] は過去のトピックの時間遷移にもとづいて, 未来のトピックを予測するモデルとして TM-LDA を提案した. これらはトピックの時間遷移を考慮しているが, 盛り上がりは考慮されていない. Diao ら [6] はマイクロブログのコーパス全体で時間に依存して変化するトピックと, 各ユーザで時間に依存しないトピックがあるという仮定にもとづきトピックを抽出するモデルを提案し, 抽出したトピックをバースト検出のアルゴリズムに適用することで, Emerging topic を抽出した.

非確率的なトピックモデルとして, Dictionary Learning や Non-negative Matrix Factorization (NMF) [14] にもとづいた Emerging topic の抽出手法が提案されている [12], [20]. Kasiviswanathan ら [12] は Dictionary learning の目的関数の時間変動量にもとづく新規文書の検出と, 検出した文書をクラスタリングし Emergint topic を推定する二段階のアプローチを提案した. さらに彼らはこのアプローチの最適化を, ストリームデータに対して効率的に行う方法を示した. Saha と Sindhwani [20] はこれを発展させ, NMF による Emerging topic の推定に応用した. 彼らは NMF にトピックの盛り上がりや遷移を考慮した制約を加えることで, Emerging topic と滑らかに遷移する Evolving topic の推定手法を示した.

しかしながら, これら全てのアプローチでは特定の情報と関連する詳細な Emerging topic を抽出することを想定していない. ソーシャルメディアには様々な分野の話題が存在するため, データ全体に手法を適用した場合, 最も数が多く盛り上がっている支配的なトピックが優先的に抽出されてしまう. 抽出するトピックの数を増やせば, 細かな Emerging topic を抽出できる可能性もあるが, 処理時間やメモリ利用量の観点から, 現実的な方法ではない. オンラインでデータを処理する既存手法 [9] であっても, 膨大な時系列テキストを実用的な時間で処理するのは容易ではない. Zimmerman ら [28] や Xie と Xing [26] は

文書データをクラスタリングし, 個々のクラスタをさらにクラスタリングする 2 段階のアプローチにより, 大局的なトピックとそれをさらに詳細化したトピックを検出する手法を提案した. しかしここで抽出されるトピックは, データ全体で支配的なトピックを詳細化したものであるため, 詳細化されたトピックも特定の情報に関係するものであるとは限らない. またソーシャルメディアにおける膨大なテキストデータから, Emerging topic を抽出するには多くの時間を要する.

これに対し Chen ら [5] はソーシャルメディアから, 企業や大学など特定の組織に関する Emerging topic を検出する手法を提案した. この手法では組織に関する公式アカウントやキーワード検索結果から得られた文書を教師データとして構築した分類器を用いて, 他の関連する文書を分類し収集している. さらに収集したデータをインクリメンタルにクラスタリングし, そのクラスタにおける文書数やユーザ数の変化に応じて分類器により Emerging topic か否かを判定する. 彼らの研究目的は本研究と近いが, この手法は関連データを収集するために, 各々の組織について分類器を構築する必要がある. しかしソーシャルのメディアの情報は刻一刻と変化するため, 一度構築した分類器では時間の経過に伴い関連文書の収集精度が低下すると予想される. さらに新たな対象に関する文書を収集する度に, 組織に関するユーザアカウントなどの教師データを集め分類器を構築する手間を要する. 提案手法は, 教師データなしでクエリのみを用いて関連データを広く効率的に集め, さらにクエリと近い Emerging topic を時間依存およびクエリ依存の NMF にもとづいて抽出する. これにより, 特定の組織だけでなく, 多様な情報と関連する Emerging topic を効率的かつ精度良く抽出できるより汎用的な手法の確立を目指す.

2.1 時間依存 NMF

本節では提案手法の背景技術となる時間依存 NMF について簡潔に述べる. 一般にトピックモデルとして用いられる NMF [27] は, bag-of-words モデルを用いて表現された文書の特徴ベクトルを複数合わせた行列を入力としたとき, 各文書に対するトピックの分布である文書トピック行列, 各トピックに対する単語の分布であるトピック単語行列を出力する. これらの行列は入力文書における単語の共起性にもとづいて推定される.

この NMF を拡張することで, 時系列データにおいて Emerging topic を抽出する手法が Saha と Sindhwani によって提案されている [20]. この手法は, ある時間区間における文書データに対して NMF を適用する. あるタイムステップを t , 時間区間を表すタイムウィンドウのサイズを w , このときの文書数を m , 単語数を n としたときの文書群の特徴量を $\mathbf{X}(t-w+1:t) \in \mathbb{R}_{\geqq 0}^{m\times n}$ ($\mathbb{R}_{\geqq 0}^{m\times n}$ は正の実数からなる $m \times n$ の行列集合) とすると

$$\mathbf{X}(t-w+1:t) \approx \mathbf{WH} \tag{1}$$

として近似される文書トピック行列 $\mathbf{W} \in \mathbb{R}_{\geqq 0}^{m\times k}$ およびトピック単語行列 $\mathbf{H} \in \mathbb{R}_{\geqq 0}^{k\times n}$ を推定する. ここで k はトピック数である. このとき, $\mathbf{W}$ および $\mathbf{H}$ の各要素は非負値であるという制約をもつ. 以降は表記の簡単化のため, $\mathbf{X}(t-w+1:t)$ を $\mathbf{X}$ と記載する.

Saha と Sindhwani は式 (1) の行列分解に対して, トピックの強度の時間変動を考慮する正則化を組み込むことで, 盛り上がっているトピック (Emerging topic) を精度良く抽出している. なおトピックの強度とは, 特定のトピックに関連する文書の量を表す. 具体的には, トピックの強度の時間変化量が小さいほどペナルティを課す制約を次式の正則化項として加える.

$$\min_{\mathbf{W}^{em}} \sum_{\mathbf{w}_j \in \mathbf{W}^{em}} \sum_{i=1}^{T-1} c_i \max(0, \nu - (\mathbf{DFSw}_j)_i)^2 \tag{2}$$

ここで, $\mathbf{W}^{em} \in \mathbb{R}_{\geqq 0}^{m\times k_{em}}$ は列成分が Emerging topic に該当する $\mathbf{W}$ の部分行列であり, $\mathbf{w}_j$ は j 番目の列ベクトルである. また k_{em} は Emerging topic の数である. $\mathbf{S}$ は同じタイムステップに存在する文書群のトピック分布の和を計算する $T \times m$ の行列であり, ある時刻を示す行において, ある列の文書が投稿されていれば, その要素の値は 1 を取る. ここで T はタイムステップの総数を表す. $\mathbf{F}$ はトピックの強度の時間的変動を平滑化しノイズを除去するための $T \times T$ の行列であり, Hodrick-Prescott Filter [7] と呼ばれている. $\mathbf{D}$ は勾配作用素として働く $(T-1) \times T$ の行列であり, 例えば $\mathbf{D}_{i,i} = -1$ および $\mathbf{D}_{i,i+1} = 1$ のような値を取る. これによりあるステップと一つ前のステップとのトピックの強度の差分を計算する. この正則化項はこれらトピック強度の差が小さいほど, 定数 $\boldsymbol{\nu}$ によりペナルティを課すように設計されている. また, c_i はタイムステップ i 毎の重みを表しており, 古い情報ほどペナルティを小さくするように設定される.

なお Saha と Sindhwani は $\mathbf{W}^{em}$ にあたる Emerging topic に加えて, 時間の経過と共に徐々に遷移する Evolving topic も考慮しているが, 本研究では任意の情報と関連する Emerging topic の抽出に焦点を当てる. Evolving topic の考慮は提案手法と競合するものではなく, 提案手法を Evolving topic を考慮するために拡張することも可能である.

3. 提案手法の概観

ソーシャルメディアから任意の情報と関連する Emerging topic を抽出する本研究のアプローチは 2 段階のステップからなる. 1 ステップ目においては膨大な文書データから, クエリにもとづいて関連文書を Coverage 優先で高速に絞り込みデータ量を削減する. この際広く収集されたデータにおいては, ノイズとなる非関連文書が混在する可能性がある. そこで 2 ステップ目として, 絞り込まれたデータに対してクエリとより関連性の高い Emerging topic を抽出するトピックモデルを提案する.

各ステップの処理の流れについて, 図 2 を用いて説明する. まず 1 段階目のステップにおいては, ソーシャルメディアから発信されるテキストストリーム (ソーシャルストリーム) を入力とし, 文書の関係を表すグラフを生成する. ここで文書とは, 例えばユーザが投稿する一つのテキストメッセージのことを表す. このグラフは文書が入力される度にオンラインで更新される. 次にこのグラフと入力されたクエリを用いて局所的にクラスタリングを行い, 特定のクエリと関連する文書と接続関係の

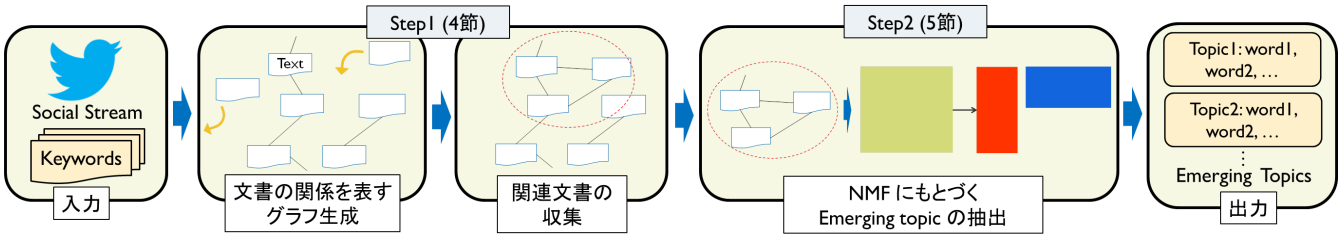

図 2　提案手法の処理の流れ．まずソーシャルストリームを入力とし，オンラインで文書の関係を表すグラフを生成・更新する．次にユーザによって指定されたクエリとなる特定のキーワード群をもとに，クエリを含む文書のノードと関係の近いノードの文書を絞り込む．絞り込んだ文書に対して時間とクエリを考慮した NMF を適用し，特定の情報と関連する Emerging topic を抽出する．

近いノードの文書をグループ化することで，クエリと関連性の近い文書を絞り込む．この際のクエリは，ユーザが任意の情報を示すキーワード群として指定する．グラフの生成およびクエリにもとづく関連文書の収集については 4. 節で示す．2 段階目のステップにおいては，削減したデータに対しトピックモデルを適用し，Emerging topic を抽出する．トピックモデルとしては，制約の付与による拡張が容易な NMF を採用し，2.1 節で述べた従来の時間依存 NMF をさらにクエリ依存に拡張する．NMF にもとづく Emerging topic の抽出については 5. 節で述べる．

4. 関連文書の収集

膨大な量のソーシャルストリームから時間をかけずに関連文書を広く収集する．そのために高速に局所的なデータをクラスタリングできるアルゴリズムとして，理論計算機科学の分野において提案されているローカルクラスタリング [1] を文書データの収集に応用する．それに伴い，まず文書の関係性を表すグラフをオンラインで構築する．文書の関係性を定義するため本稿では，特徴的な単語が互いに共起している文書は関連があるという仮定を置く．以下，4.1 節ではグラフの生成，4.2 ではローカルクラスタリングを用いた関連文書の収集について述べる．

4.1 グラフ生成にもとづく文書の関係性構築

ソーシャルストリームに対して，文書との関係性を構築するためにグラフをオンラインで生成する．生成するグラフを無向グラフ $G = (V, E)$ としたとき，V を現在の分析対象とする文書集合 D における個々の文書 d_i をノードとするノード集合，E をノード d_i をつなぐエッジ集合とする．ソーシャルストリームとして新たな文書 d_i が入力されると，文書のノードがグラフ G のノード集合 V に追加される．このとき d_i は bag-of-words モデルを用いて単語の集合として表現されている．また同時に文書が投稿された時刻 d_i^t を持つ．

文書間の関係性は，特徴的な単語の共起数によって定義する．具体的には，新しい文書 d_i とグラフ内の文書 $d_r \in V$ における特徴的な単語の共起数が規定値以上のときに，これらは関連しているとし各ノードをエッジで接続する．この際 d_r は全てのノードを探索する必要はなく，単語をキーとした各ノード群へのハッシュテーブルを作成することで d_i の単語数に応じた探索を行い，ノード数にほとんど依存しない時間で関連文書のノードを探すことができる．また単語の特徴度は予め用意した文書コーパスから計算した df により定義し，df が規定値よりも小さい単語，つまり多くの文書に表れない単語を特徴的とする．

最後に新たに追加された文書 d_i のノードに対して，予め与えたタイムウィンドウ w よりも古いノードを削除する．すなわち，ノード集合 V において最も古いノード $d_o \in V$ を取得し，$d_i^t - d_o^t > w$ である場合にノード d_o に接続しているエッジおよびノード d_o を削除する．この処理は，$d_i^t - d_o^t > w$ を満たすノード d^o がなくなるまで繰り替えし行う．

4.2 グラフからの関連文書の収集

前節で生成したグラフからクエリと関連する文書を高速に絞り込むため，ローカルクラスタリングのアルゴリズム [1] を応用する．ローカルクラスタリングはシードとなるノードから，Conductance と呼ばれる指標をもとに局所的にクラスタリングする手法である．Conductance はサブグラフにおけるノード間を結合するエッジの密度と，サブグラフとサブグラフ外のノードとを結合するエッジの疎性をもとに決まる値である．サブグラフ S における Conductance ϕ は次式によって定義される．

$$\phi = \partial(S)/\mu(S) \tag{3}$$

ここで $\partial(S)$ はサブグラフ S のノードとサブグラフ外のノードとをつなぐエッジの数，$\mu(S)$ は S のノードの次数の総和である．Conductance はグラフ全体の情報を考慮する指標ではないが，特定のノードと局所的に類似した要素をクラスタリングできる．加えて計算量がグラフ全体のサイズにさほど依存しない特徴がある．

具体的なアルゴリズムを簡単に説明する．最初にシードとなるノードを決める．次にシードを起点にランダムウォークにもとづいて隣接するノードに移動しながら，付近のノードを一つのサブグラフにまとめる処理を繰り返す．繰り返し中に Conductance がしきい値よりも小さい場合に処理を打ち切り，その時点でのサブグラフをクラスタリング結果として出力する．

本研究ではクエリと関連する文書を集めるため，ローカルクラスタリングのシードとなるノードを，クエリを含む文書のノードとする．またクエリが複数ある場合は，各クエリについてシードをそれぞれ選択し，それぞれのシードについてローカルクラスタリングを行い，得られたクラスタを結合する．このクラスタにおいてグループ化された文書集合をクエリと関連するものとみなし，次節における Emerging topic の抽出に用いる．

5. Emerging topic の抽出

図 3 に NMF による Emerging topic の抽出処理の概要を示す. ローカルクラスタリングで絞り込んだ, ある時間区間の文書データに対して NMF を適用する. 2.1 節において述べた時間依存 NMF と同様に入力となる文書単語行列において, 行に対応する文書はタイムステップごとに区切られ, 各タイムステップ毎のトピックの盛り上がりの変化をもとに Emerging topic を抽出する. この際, ノイズとなる非関連文書を含む文書データから, 特定の情報と関連する Emerging topic を精度良く抽出するために時間依存 NMF を拡張する. 次節では, NMF の拡張に応じた目的関数を設定する.

なお本研究では NMF を適用する文書の特徴量として, 自然言語処理の分野において一般的に多く用いられる TFIDF [19] を利用する. TFIDF にもとづく行列 $\mathbf{X}$ の特徴量は次の通りである.

$$\mathbf{X}_{i,j} = \frac{\log_2(1 + TF(i,j)) \times IDF(j)}{C_i} \tag{4}$$

ここで, $\mathbf{X}_{i,j}$ は行 i 列 j における行列 $\mathbf{X}$ の要素, $TF(i,j)$ は文書 i における単語 j の出現頻度であり, $IDF(j)$ は全文書数に対して単語 j が出現する文書数の割合の逆数の対数をとった値を表す. また C_i は $\mathbf{X}$ の各行ベクトルを正規化する定数である.

5.1 目的関数の設定

クエリと関連性の高い Emerging topic を抽出するために, 各トピックにおいてクエリと関連して盛り上がっている内容と, それ以外の内容とに分かれることを狙い NMF の目的関数を設定する. 具体的には時間依存制約およびクエリ依存制約をもとに, 以下の目的関数を最小化する問題として定式化される.

データ依存性: 一般的に NMF は式 (1) における左辺と右辺の二乗誤差や KL Divergence が最小になるように $\mathbf{WH}$ を求める. すなわち $\mathbf{WH}$ が $\mathbf{X}$ に近づくように解を求める. したがって, 文書データに依存する項は

$$E_d = ||\mathbf{X} - \mathbf{WH}||_F^2 \tag{5}$$

と定義される. ここで $||\cdot||_F$ はフロベニウスノルムを表す.

時間依存性: Emerging topic を抽出するために, Saha と Sindhwani [20] の制約式 (2) にもとづく時間依存の制約を与える. この際クエリと関連する Emerging topic を抽出するために, クエリと関係の近いトピックの強度の時間変化量が小さいほどペナルティを与えるように修正した正則化項 E_t を定義する.

$$E_t = \sum_{\mathbf{w}_j \in \mathbf{W}^{qem}} \sum_{i=1}^{T-1} c_i \max(0, \nu - (\mathbf{DFSw}_j)_i)^2 \tag{6}$$

ここで, $\mathbf{W}^{qem} \in \mathbb{R}_{\geqq 0}^{m \times k_{qem}}$ は列成分がクエリと関連する Emerging topic に該当する $\mathbf{W}$ の部分行列であり, $\mathbf{w}_j$ はその j 列目のベクトルである. また k_{qem} はクエリと関連する Emerging topic の数である. 後述のクエリ依存制約を与えるトピックに対してのみ時間依存制約を与えることで, 盛り上がっているトピックの中でもクエリと関連するものを抽出する.

クエリ依存性: クエリと関係の近いトピックについてのみ盛り上がりを検出するために, 特定のトピックにクエリが出現しない場合ペナルティを与える項 E_q を定義する.

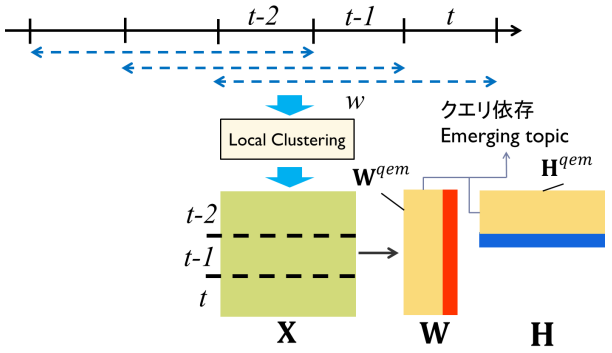

図 3 時系列文書データから, 時間およびクエリ依存 NMF による Emerging topic 抽出処理の概要. タイムウィンドウ w の区間においてローカルクラスタリングで絞り込んだ関連文書から, NMF の入力となる行列 $\mathbf{X}$ を計算する. この際 $\mathbf{X}$ は複数のタイムステップの文書の特徴量から構成される. $\mathbf{X}$ を NMF により $\mathbf{W}$ と $\mathbf{H}$ に分解する. これらは提案手法の制約により, クエリ依存かつ時間依存なトピックに関するもの ($\mathbf{W}^{qem}$, $\mathbf{H}^{qem}$) と, クエリ非依存かつ時間非依存なトピックに関するものに分けられる.

$$E_q = \sum_{i}^{k_{qem}} \sum_{j}^{n} \max(0, \mathbf{Q}_{i,j} - \mathbf{H}_{i,j}^{qem})^2 \tag{7}$$

ここで $\mathbf{H}^{qem} \in \mathbb{R}_{\geqq 0}^{k_{qem} \times n}$ は行成分がクエリと関連する Emerging topic に該当する $\mathbf{H}$ の部分行列である. $\mathbf{Q}$ は $k_{qem} \times n$ の行列であり, クエリに該当する単語の列成分に定数 η を, それ以外の成分には 0 を持つ. E_q は $\mathbf{H}^{qem}$ の各要素が $\mathbf{Q}$ よりも大きくなるように制約を与える. これによりクエリの単語や共起性の近い単語を, 特定のトピック中に表れやすくする効果が期待できる. E_q と E_t の両方の制約を考慮することで, クエリとの関連性の近い Emerging topic を抽出する.

5.2 最 適 化

前節で定義した目的関数を最小化するように, 文書トピック行列 $\mathbf{W}$ およびトピック単語行列 $\mathbf{H}$ を最適化する. データ依存項, 時間依存項およびクエリ依存項, パラメータの非負性を考慮すると, 最終的な目的関数および求める $\mathbf{W}$ と $\mathbf{H}$ は次のようになる.

$$E = E_d + \lambda_t E_t + \lambda_q E_q \tag{8}$$

$$\mathbf{W}, \mathbf{H} = \underset{\mathbf{W},\mathbf{H} \geqq 0}{\arg\min}\, E \tag{9}$$

ここで λ_t および λ_q はそれぞれの項により与える制約の重みを表すパラメータである. 式 (8) は非凸な関数であることから, $\mathbf{W}$ または $\mathbf{H}$ を固定した凸関数を交互に最適化して解を求める方法が一般的である. 最適化の方法は交互最小二乗法や最急降下法, アクティブセット法 [13] など様々な手法が提案されている. 本研究では大きな行列の処理も想定されることから, 省メモリで計算可能な L-BFGS-B [4] 法を用いて実験を行った. 目的関数は非凸な関数であるため大域的な最適解を求めることは難しい. そのため $\mathbf{W}$ や $\mathbf{H}$ を最適化する際には, いくつかのパターンによるランダムな値で初期化を行い, 目的関数が最も小さくなったものを最終的な推定値とした.

Algorithm 1 Query dependent emerging topic extraction

Input: text graph G, query terms Q,
topic size k, k_{qem}, hyperparameters ν, η, λ_t, λ_q

Output: $\mathbf{W} \in \mathbb{R}_{\geqq 0}^{m \times k}$, $\mathbf{H} \in \mathbb{R}_{\geqq 0}^{k \times n}$

$D^Q \leftarrow LocalClustering(Q)$
$\mathbf{X}(t-w-1:t) \leftarrow CalculateFeatures(D^Q)$
$\mathbf{Q} \leftarrow CalculateQueryDependentMatrix(Q)$
Initialize $\mathbf{W}$ and $\mathbf{H}$ to random distribution
while not(converge) **do**
  $\mathbf{W} \leftarrow \arg\min_{W \geqq 0} E$
  $\mathbf{H} \leftarrow \arg\min_{H \geqq 0} E$
  Check convergence based on relative change in objective value $E < \epsilon$
end while

最後に, 本研究のクエリ依存 Emerging topic 抽出アルゴリズムを Algorithm 1 にまとめる.

6. 評　　価

提案手法の有効性を以下の三つの観点から検証する.

（1） 関連文書収集の精度
（2） 特定の情報と関連する Emerging topic 抽出の精度
（3） 処理時間

関連文書収集の精度とは, 4. 節で述べた方法により, クエリと関連する文書をどれだけ集められるかという観点である. また, 収集したデータに対して 5. 節で述べた時間依存かつクエリ依存のトピック抽出により, どれだけ特定の情報と関連する Emerging topic を精度良く抽出できるかを評価する. 処理時間については, 関連データをローカルクラスタリングで高速に絞り込むことで, Emerging topic 抽出の枠組みにおいて処理時間にどの程度の違いが生じるかを検証する. 以下本実験に用いたデータセットについて説明した後, それぞれの評価について述べる.

6.1 データセット

今回はトレンド抽出の情報源としてソーシャルメディアの中でも即時性が高く, 情報量の多いマイクロブログの Twitter を用いる. 関連する文書の収集や Emerging topic の抽出について提案手法の精度を評価するために, Twitter API を用いて収集された日本語 Tweet に対して, 予め想定したキーワードと関連するツイートにその内容を表すラベルの付与を行う. 今回はオリンピックをキーワードと想定し, 関連データとして 2012/7/30 から 2012/8/5 までのロンドンオリンピックに関連するツイートを用いることとした. 本実験では, 人手によりオリンピックと関係の深いハッシュタグをラベルとして選び, そのラベルの付与されたツイートを関連データとした. オリンピック関連するハッシュタグとしては, 例えば# 柔道や# 体操などを選択した. 簡単のため各 Tweet に複数の関連ハッシュタグが付与されているものは, データセットから除外した. 一方, オリンピックと関係のないツイートとして, オリンピックと関係のないハッシュタグが付与されたツイートを用いることとした. これらオリンピックの関連ツイートと非関連ツイートを混在させたデータセットに対して, 提案手法でどれだけ精度よくオリンピックに関連する Emerging topic を抽出できるかを評価する. これらのデータセットは, オリンピックに関連するものが 57,306 ツイート, 関連しないものが 329,238 ツイートであった. なお, 日本語の文章を単語単位に分割するために, 形態素解析エンジンの MeCab(注1) を用いた.

6.2 関連文書の収集精度

提案手法により, どれだけ正確かつ網羅的に関連文書を集められるか否かを検証する. 前述のデータセットを用いて「オリンピック」をクエリとして関連データを収集した際に, オリンピックと関係するラベルが付与された Tweet の収集精度を適合率 (*Precision*), 収集量を再現率 (*Recall*) によって評価する.

本実験においては, 提案手法を含む下記の手法を比較する.

- KS: キーワードサーチによりクエリを含む文書を取得.
- LC@θ: オンラインで更新される文書グラフに対してローカルクラスタリングを適用. θ は Conductance を基準に, クラスタリングアルゴリズムのイテレーションを打ち切るしきい値. 文書データをグラフに保持するタイムウィンドウ w およびローカルクラスタリングを適用する間隔は 1 時間とした.

表 1 に実験結果を示す. KS による結果は *Precision* が 0.7 であり, オリンピックと関連する文書をある程度正確に集められているものの, *Recall* が 0.019 であり全然網羅的に集められていないことがわかる. これに対し, LC による結果は *Precision* は KS と同程度の値であり, *Recall* も比較的高い値であることから, ある程度関連する文書を比較的多く集められている. このときクラスタリングを打ち切る基準となる Conductance のしきい値 θ が小さいほど, より広い範囲で関連データを収集するため *Recall* が上昇し *Precision* が低下する. パラメータによる結果の変動はあるものの, 文書グラフのローカルクラスタリングによる本研究のアプローチは, 単純なキーワードサーチよりも高精度に関連データを収集できていることを確かめられた. この絞り込まれたデータにおいて NMF を適用することで, 特定の情報と関連する Emerging topic を効率的に抽出できると考えられる.

表 1 データ収集の精度. P は *Precision*, R は *Recall*, F1 は $F-measure$, TP は True positive, FP は False positive, FN は False negative を表す.

Method	P	R	F1	TP	FP	FN
KS	0.73	0.020	0.039	1138	419	56168
LC@0.6	**0.83**	0.45	0.58	25882	5326	31424
LC@0.2	0.64	**0.62**	**0.63**	35664	20153	21642

6.3 Emerging topic の抽出精度

特定の情報と関連する Emerging topic の抽出精度を検証するために, まず評価指標について簡潔に述べる. 本実験では Saha と Sindhwani [20] も用いた指標である $Jensen-ShannonDivergence$ (JSD) を利用して, 真のトピックにおける単語の分布と, 推定されたトピックにおける単語の分布を比較

(注1)：http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

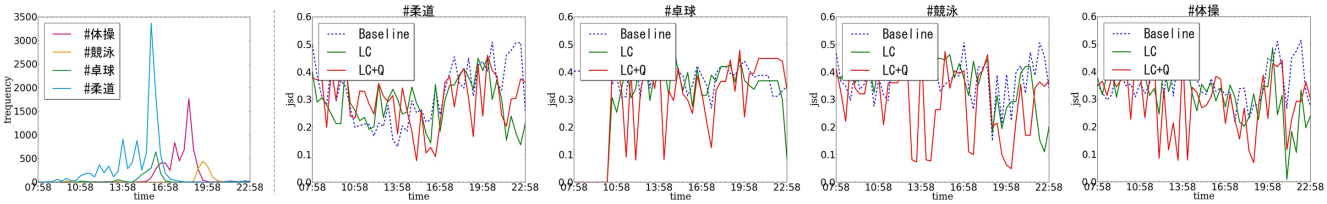

図 4　ある時間帯におけるオリンピックと関連するツイート量の時間変化 (左) と対応する時刻におけるトピックの抽出精度 (右の四つ). 抽出精度は各ラベルに関する真のトピックと推定されたトピックとの JSD にもとづいており, 値が小さいほど高精度であることを表す. 左図において数が増加し盛り上がっているラベルについて, 該当する時刻の JSD が小さいほど, Emerging topic を精度良く抽出できていることになる.

する. 真のトピックは特定のラベルが付与された文書は同じトピックであるという仮定のもと, トピックモデルを用いてあらかじめ計算しておく. 今回はオリンピックに関連する# 体操や# 卓球などのラベルが付与された各々のツイートから, 真の# 体操トピックや# 卓球トピックを計算する. これら真のトピックと, 様々なラベルのツイートが混在したデータセットに各手法を適用し得られたトピックとの JSD を比較する. 各々のラベルの付与されたツイート数が増加し盛り上がっている時点で, 対応するトピックに関する JSD が減少しているほど, Emerging topic の抽出精度が高いと考えることができる. ページ数の都合上詳細は省くため, 評価値の算出などは Saha と Sindhwani の文献 [20] を参照されたい.

前述の JSD による評価指標を用いて, 下記の Baseline と提案手法を比較する.

- Baseline: Emerging topic を抽出するための時間依存制約 [20] を付与した NMF を全データに対して適用.
- LC: 時間依存 NMF をローカルクラスタリングにより絞り込んだ関連データに対して適用.
- LC+Q: 時間依存かつクエリ依存 NMF をローカルクラスタリングにより絞り込んだ関連データに対して適用.

本実験において用いたパラメータは, タイムウィンドウ w を 1 時間としタイムステップを 20 分とした. また, ローカルクラスタリングの Conductance のしきい値は 0.2, NMF におけるパラメータは $k = 5$, $k_{qem} = 4$, $\nu = 20$, $\eta = 1$, $\lambda_t = 10$ および $\lambda_q = 1000$ とそれぞれ経験的に設定した.

図 4 および表 2 に評価結果を示す. 図 4 において, 左はある時間帯における各ラベル付きツイート量の時間変化であり, 右の四つは対応する時刻において各手法を適用して得られたトピックとラベル毎の真のトピックとの JSD を示している. また表 2 は各ラベルに関する話題が最も盛り上がっている時刻の JSD を示している. 各ラベルに関する話題が最も盛り上がっている時刻は, 単純に図 4 左において各ラベルのツイート量が最も多い時刻とした. Baseline の結果においては, 各ラベルのツイートが盛り上がっている時刻において, 顕著な JSD の減少がみられない箇所 (例えば# 競泳の 20 時前後や# 体操の 18 時前後など) が存在する. これはオリンピックと関係のないトピックが混在しているデータセットの影響を受けているため, オリンピックと関連する Emerging topic を精度よく抽出できていないと考えられる. 一方 LC は, Step1 で「オリンピック」

表 2　各ラベルが最も盛り上がった時刻において各手法で得られたトピックと真のトピックとの JSD および全ラベルの JSD の平均.

Method	# 柔道	# 卓球	# 競泳	# 体操	全ラベルの平均
Baseline	0.212	0.353	0.212	0.228	0.253
LC	0.178	0.275	0.195	0.235	0.221
LC+Q	**0.106**	**0.250**	**0.100**	**0.113**	**0.142**

をクエリとしたローカルクラスタリングにより関連データに絞ることで, 表 2 からも例えば# 競泳について Emerging topic の抽出精度が向上しているのがわかる. しかしながら, LC だけではデータを十分に絞りきれない場合もあり, データ量を削減することで高速にトピックを抽出できるものの, 表 2 の体操に示すように精度が向上しない場合も多くみられる. これに対し LC+Q は, Step2 でさらにクエリ依存制約を与えた時間依存 NMF を適用することで, さらなる精度向上が確認できる.

6.4　処理時間と定性的な結果

提案手法は Python 言語で実装し, NMF の L-BFGS-B 法による最適化にはライブラリとして Numpy, SciPy を用いた. 実装したプログラムは CentOS 5.5, Intel Xeon5650 2.66GHz 6Core HT x 2, 48GB RAM の環境で実行した.

図 5 にある時刻におけるツイートに Baseline と提案手法 (LC+Q) を適用した場合の処理時間および抽出された Emerging topic の例を示す. NMF はデータ量に依存して計算量が大きく増大してしまうため, ベースラインは 1388 秒の時間を要している. 一方提案手法は関連データを絞り込む処理に 0.8 秒, NMF の処理に 65.4 秒の時間を要しており, ベースラインと比較して約 20 倍の速度で処理できている. 本研究における処理時間の削減のポイントは, データ量にほぼ依存しないアルゴリズムであるローカルクラスタリングにより, データ量を削減してから NMF を適用していることであり, NMF の最適化による処理速度は考慮していない. そのため NMF を並列に計算する手法 [16] などを用いることで, さらなる高速化が期待できる.

抽出されたトピックについては, Baseline がデータ全体で支配的なトピックを抽出してしまっているのに対し, 提案手法はオリンピックに関係するトピックを抽出できていることがわかる. 図 5 に抽出した全てのトピックを示していないが, Baseline によって抽出されたどのトピックも, オリンピックに関連する内容と該当するものはなかった. また今回は定量的評価のために, ラベルとして使える多様なハッシュタグがあるオリンピックの開催時期のデータセットを用いたが, 他のデータセットに

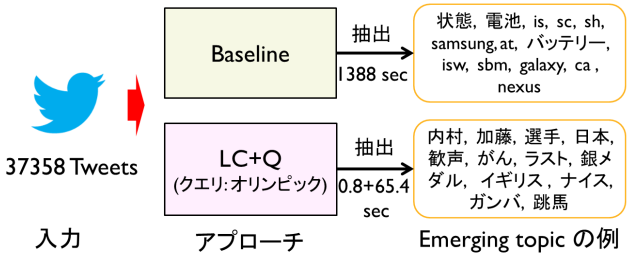

図 5　提案手法 (LC+Q) と Baseline による処理時間および抽出される Emerging topic の例.

提案手法を適用することで, 図 1 に示すような Emerging topic を抽出するできることも確認した.

7. おわりに

本稿ではソーシャルメディアから任意の情報と関連する Emerging topic を抽出する枠組みを提案した. 多様な情報と関連する Emerging topic を抽出する汎用的なアプローチを実現するため, 提案手法では教師データを用いずにクエリのみを用いて, クエリと関連する Emerging topic を精度良く抽出することを課題とした. 提案手法では, 一段階目のステップとして文書の関係性を表すグラフをオンラインで生成し, グラフにローカルクラスタリングを適用することでクエリと関連する文書を高速に収集する. これにより, 特定の情報と関連するデータを広く集めると同時に, 次ステップにおけるトピックモデルの処理速度を大幅に削減できる. この際広く絞り込まれた文書データにおけるノイズである非関連文書を考慮するため, さらに NMF を用いた Emerging topic 抽出手法をクエリ依存に拡張した. 評価実験では特定の情報と関連する Emerging topic の抽出精度および処理時間の観点で, ローカルクラスタリングによるデータの絞り込みと, 時間依存かつクエリ依存 NMF による枠組みの有効性を示した.

現在の手法は Emerging topic の抽出精度がクエリの粒度やトピック数などに依存してしまうと考えられる. そこで今後の課題として, トピック数を自動で適切な値に決定するノンパラメトリックなアプローチ [10] などを応用することで, さらなる精度向上を目指すことがあげられる. また高速化について, 今回は関連文書の収集についてのみに注目しているため, NMF を並列に最適化するアルゴリズムを提案手法の枠組みに組み込むことで, 更なる高速化が期待できる. また提案手法を Trend-aware recommendation [17] などに応用することも興味深い.

文　献

[1] Andersen, R. and Peres, Y.: Finding Sparse Cuts Locally Using Evolving Sets, *Proc. of STOC'09*, pp.235-244, 2009.

[2] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, In *Journal of Machine Learning Research*, 3, pp.993-1022, 2003.

[3] Blei, D. M. and Lafferty, J. D:Dynamic topic models, In *Proc. of ICML'06*, pp.113-120, 2006.

[4] Byrd, R. H., Lu, P. and Nocedal, J.: A Limited Memory Algorithm for Bound Constrained Optimization, SIAM Journal on Scientific and Statistical Computing, 16, 5, pp.1190-1208, 1995.

[5] Chen, Y., Amiri, H., Li, Z. and Chua, T.-S.: Emerging topic detection for organizations from microblogs, In *Proc. of SIGIR'13*, pp.43-52, 2013.

[6] Diao, Q., Jiang, J., Zhu, F., Lim, E.-P.: Finding bursty topics from microblogs, In *Proc. of ACL'12*, pp.536-444, 2012.

[7] Hodrick ,R. J. and Prescott, E. C.: Postwar U.S. Business Cycles: An Empirical Investigation, Journal of Money, Credit, and Banking, 29, pp.1-16, 1997.

[8] Hoffman, M. D., Blei, D. M. and Cook, P. R.: Bayesian Nonparametric Matrix Factorization for Recorded Music, In *Proc. of ICML'10*, pp.439-446, 2010.

[9] Hoffman, M. D., Blei, D. M. and Bach, F. R.: Online learning for latent dirichlet allocation, In *Proc. of NIPS'10*, pp.856-864, 2010.

[10] Hoffman, T.: Probabilistic latent semantic analysis, In *Proc. of SIGIR'99*, pp.50-57, 1999.

[11] Jin, X., Gallagher A. C., Cao L., Luo, J. and Han, J.: The wisdom of social multimedia: using flickr for prediction and forecast, In *Proc. of ACM MM'10*, pp.1235-1244, 2010.

[12] Kasiviswanathan, S. P., Melville, P., Banerjee, A. and Vikas, S.: Emerging topic detection using dictionary learning, In *Proc. of CIKM'11*, pp.745-754, 2011.

[13] Kim, H. and Park, H.: Non-Negative Matrix Factorization Based on Alternating Non-Negativity Constrained Least Squares and Active Set Method, SIAM Journal on Matrix Analysis and Applications, 30, 2, pp.713-730, 2008.

[14] Lee, D. D. and Seung, H. S.: Learning the parts of objects by non-negative matrix factorization, Nature, 401, pp.788-791, 1999.

[15] Li, C., Sun, A. and Datta, A.: Twevent: segment-based event detection from tweets, In *Proc. of CIKM'12*, pp.155-164, 2012.

[16] Liu, C., Yang, H.-C., Fan, J., He, L.-W., Wang, Y.-M.: Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce, In *Proc. of WWW'10*, pp.681-690, 2010.

[17] Roy, S. D., Mei, T., Zeng, W. and Li, S.: SocialTransfer: Cross-Domain Transfer Learning from Social Streams for Media Applications. In *Proc. of ACM MM'12*, pp.649-658, 2012.

[18] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of WWW'10*, pp.851-860, 2010.

[19] Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval, Information Processing and Management: an International Journal, 24, 5, pp.513-523, 1988.

[20] Saha, A. and Sindhwani, V.: Learning evolving and Emerging topics in social media: a dynamic nmf approach with temporal regularization, In *Proc. of WSDM'12*, pp.692-702, 2012.

[21] Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welpe, I. M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proc. of ICWSM'10*, pp.178-185, 2010.

[22] Wang, C., Blei, D. B. and Heckerman, D.: Continuous Time Dynamic Topic Models, CoRR, abs/1206.3298, 2012.

[23] Wang, Y., Agichetein, E. and Benzi, M.: TM-LDA: efficient online modeling of latent topic transitions in social media, In *Proc. of KDD'12*, pp.123-131, 2012.

[24] Wang, Z et al: Propagation-based social-aware replication for social video contents, *In Proc. ACM MM'12*, pp.29-38, 2012.

[25] Weng, J. and Lee, B.-S.: Event Detection in Twitter, In *Proc. of ICWSM'11*, pp.401-408, 2011.

[26] Xie, P. and Xing, E. P.: Integrating Document Clustering and Topic Modeling, CoRR, abs/1309.6874, 2013.

[27] Xu, W., Liu, X. and Gong, Y.: Document clustering based on non-negative matrix factorization, In *Proc. of SIGIR'03*, pp.267-273, 2003.

[28] Zimmermann, M., Ntoutsi, I. and Siddiqui, Z. F., Spiliopoulou, M. and Kriegel, H.-P.: Discovering global and local bursts in a stream of news, In *Proc. of SAC'12*, pp.807-812, 2012.