バースト現象検出のための Tweet Pooling による潜在トピック推移の抽出

福山　怜史† 　若林　 啓††

† 筑波大学大学院 図書館情報メディア研究科　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s1721691@s.tsukuba.ac.jp, ††kwakaba@slis.tsukuba.ac.jp

あらまし　Twitter において流行している話題を抽出することは，現実世界で流行している話題を抽出する上で重要なアプローチである．Twitter における流行に関連のある話題を抽出するためには，話題の抽出とその話題の出現頻度の推移を捉える必要がある．このためのアプローチとして，Latent Dirichlet Allocation (LDA) に対してトピックの時間発展を考慮した推論アルゴリズムを利用する手法が提案されている．しかし，バースト現象のような突発的かつ短期間にのみ発生する流行現象は，時間発展を考慮した推論アルゴリズムによる恩恵が得られないため，通常の LDA を用いた手法の方が計算効率性が高いと考えられる．本研究では，時間発展を考慮したアルゴリズムを使用せずに，Tweet Pooling の拡張によって．潜在的なトピックの推移の抽出を行う手法を提案する．提案手法では，ハッシュタグと単位時間ごとに Tweet Pooling を行なった擬似文書を用いて，LDA による学習を行う．学習後は，各時刻における擬似文書のトピックの割合とツイート数の積からトピックの推定ツイート数を計算し，その値からバースト現象の検出を行う．本稿では，2012 年 8 月 1 日から 8 月 7 日までのツイートデータからトピックとその推定ツイート数を抽出し，トピックごとのバースト検出を行った．この結果，提案手法によってバースト検出されたトピックの一部が現実世界で発生した事象と対応し，事象の発生時刻とトピックのバースト現象が一致していたことを報告する．

キーワード　Twitter，バースト現象，潜在トピック推移，LDA，Tweet Pooling，ハッシュタグ

1. はじめに

Twitter は，現実世界で起こった事象に対してユーザがリアルタイムにツイートを投稿する性質から，現実世界を知覚するセンサとしての利用が期待されている．例えば，2011 年 3 月 11 日に発生した東日本大震災では，東京都において，地震発生から 1 時間以内に毎分 1,200 件以上のツイートが投稿されたことが報告されている [1]．また，近年では，Twitter からの評判情報抽出 [2] や病気の流行予測 [3] といった手法の有効性が確認されており，Twitter ユーザが現実世界の事象に対して敏感に反応することが分かる．

このような特徴から，Twitter の投稿の傾向を分析することで，現実世界で起きた出来事や流行している話題を抽出する手法が研究されている [4] [5] [6] [7] [8]．これらの手法では，局所的な時間で話題の出現頻度が急激に増加する “バースト現象” を検出することで，出来事や流行の抽出を行う．このようなバースト現象が発生している状態を本稿では，“バーストしている” と表記する．このようにバーストしている話題を抽出し，それを分析することは，現実世界の事象の調査や，特定の人物や商品の評判分析などにおいて有用である．

バースト現象を検出するためには，特定の話題に対応するツイート数の推移データを用意する必要がある．Twitter のような大量のテキストデータにおける話題とその話題の推移を抽出する手法として，Latent Dirichlet Allocation (LDA) [9] を時系列変化に対応させた Dynamic topic model (DTM) [10] がある．LDA とは，単一の文書には複数のトピックが含まれることを仮定したトピックモデルにおいて，文書中の単語のトピックをベイズ推定によって求める確率的生成モデルである．この LDA に対して，トピックごとの単語分布およびトピック分布の時間発展を捉えるモデルが DTM である．Koike ら [8] は，DTM を用いて，大量のニュース記事やツイートからトピックとそれらの時間的推移を抽出し，個々のトピックにおいてニュース記事やツイートがバーストしているか検出する手法を提案している．

このように Twitter のような時間情報が付与されたテキストデータに対して潜在的なトピックの変化を推定する場合，DTM を用いた手法は有効であると考えられる．一方で，DTM では LDA に対して新たにパラメータを加えるため，モデルが複雑化し，計算量が増加する可能性が考えられる．LDA には，モデル特有の性質に依存した高速な推論アルゴリズムとして確率的変分ベイズ法 [11] があり，トピック数の増加に対して計算時間が劣線形である．Koike らの手法をツイートデータに適用する場合，ツイートデータに含まれるトピック数は，ニュース記事と比較してより多くの潜在トピック数が存在する可能性があるため，トピック数の影響を受けない LDA の確率的変分ベイズ法を用いることが望ましい．

本研究では，ツイートデータから潜在的なトピックの変化を推定する際に，時間発展を考慮したアルゴリズムを使用しない手法を提案する．提案手法では，Tweet Pooling の手法の拡張によって，潜在的なトピックの推移の抽出を行う．Tweet Pooling とは，LDA の手法自体を変更せずに，類似性のある複数のツイートを結合した擬似文書を作成する方法である．本研究では，Tweet Pooling の一つである Hashtag Pooling によって作成され

た擬似文書に対して，単位時間ごとに擬似文書を分割した新たな擬似文書を用いて，LDA の学習を行う．学習後はそれぞれの擬似文書におけるトピックの推論を行い，それぞれの擬似文書に対応するトピックの割合とツイート数の積を計算して，その積を擬似文書内におけるトピックの推定ツイート数とみなす．そして，各時刻におけるそれぞれのトピックの出現数を合計した結果を得る．この結果得られたトピックの推移データに対して，バースト検出を行うことにより，トピックのバースト時刻を求めることができる．以上の手法により，DTM を用いることなく潜在的なトピックの推移とバースト時刻を得ることができるため，LDA と同程度の計算コストで Twitter における話題の獲得と話題のバースト検出を行うことができる．

本稿では，提案手法を用いて，2012 年 8 月 1 日から 8 月 7 日に発生した現実世界の事象に対応するトピックの抽出とそれらのトピックの推移の抽出とバースト検出を行った．この結果，提案手法によってバースト検出されたトピックの一部が現実世界で発生した事象と対応し，事象の発生時刻とトピックのバースト現象が一致していたことを報告する．

2. 関 連 研 究

本研究に関連して，バースト現象に関する研究と潜在トピック推移の抽出に関する研究，マイクロブログからのトピック抽出に関する研究がある．

2.1 バースト現象に関する研究

まずバースト検出において著名な手法の一つである Kleinberg のバースト解析 [12] がある．この手法は，時系列データに対して定常状態とバースト状態を確率的オートマトンによってモデル化したもので，それぞれの状態において異なるパラメータを持つポアソン分布を仮定している．しかし，Twitter のような過去の情報が現在の情報に影響を与えるようなメディアにおいては，ツイート数のような時系列データがポアソン分布に従っているケースは稀であるため，このような仮定のない柔軟な手法が適当だと考えられる．

次に，Twitter におけるバースト現象の分析を行った研究として，水沼ら [4] の研究がある．水沼らは，Twitter におけるバーストの特徴を分析し，その特徴ごとにバーストの類型化を行なっている．水沼らはツイートのバースト検出手法を選択するために，バースト現象を出現頻度の外れ値とみなし，外れ値検知手法である ROKU [13], 3σ 法, 増山の検定 [14], MAD 法 [15] [16] の比較を行なっている．この比較では，1 度のバースト現象を複数のイベントとみなすことがないか，4 手法で共通して得られる確信度の高いバースト現象がどれぐらい多く検出できるかを評価している．この結果，3σ 法が，バースト平均継続時間が最も長く，また 4 手法がいずれもバースト現象として検出したハッシュタグの集合との一致率が最も高いことから，4 手法の中で最も理想的な手法であるとしている．3σ 法は，データの分布を仮定しない異常検知の手法であることから，本研究ではバースト現象の検出手法に 3σ 法を用いることとする．

2.2 潜在トピック推移の抽出に関する研究

潜在トピック推移の抽出に関する研究として，Blei らによる Dynamic topic model(DTM) [10] が提案されている．DTM では，LDA に対して，トピックごとの単語分布およびトピック分布に時間マルコフ性を導入することによって，それらの時間発展を捉えることを目的としている．時間発展させるパラメータは，トピックの流行り廃りを制御する α とトピックごとの単語分布を表す β であり，一時刻分のパラメータ遷移は正規分布を用いてモデル化されている．Blei らは，DTM によって，科学誌 Science に掲載された論文の時系列トピック解析を行い，異なる科学的テーマを含んだトピックおよびそのトピックで用いられる単語の傾向が確認できることを報告している．

Iwata らは，複数の時間スケールでのトピックの発展を解析する Multiscale Dynamic Topic Model(MDTM) [17] を提案している．MDTM では，同一トピックにおいても出現する単語は時間スケールによって異なる性質を考慮したトピックモデルである．DTM と比較した場合，MDTM で考慮するスケール数を 1 つとした場合がこれと対応している．実験の結果，MDTM はスケール数を増加させることにより，モデルの平均パープレキシティが減少していることを確認し，複数の時間スケールの分布を考慮することの重要性を報告している．DTM や MDTM はオンライン学習可能なモデルであるため，モデルの長期的な利用を行うケースに適している．本研究では，Twitter における話題のバースト現象の検出を目的としているため，DTM や MDTM のような話題の時間発展を捉えることを目的とした手法は不適当であると考え，通常の LDA における前処理と後処理を工夫することによって目的の達成を図っている．

2.3 マイクロブログからのトピック抽出に関する研究

個々のツイートにおいて文字数の制限によって語彙が豊富でない問題を解決するために，Mehrotra ら [18] は Tweet Pooling を提案している．Tweet Pooling は，関連性のあるツイートを結合した擬似文書を作成し，これらの文書を LDA の入力文書とする手法である．代表的な Tweet Pooling は，ユーザ，単位時間，バーストした語，ハッシュタグごとにツイートを分類し，それぞれのグループでツイートを結合する手法である．大量のツイートから話題の抽出を行う場合，擬似文書は特定の話題を示している必要があるため，特定の話題を示すタグであるハッシュタグごとにツイートを結合することが望ましい．Mehrotra らは，実験によりハッシュタグごとにツイートを結合して擬似文書を作成する Hashtag Pooling が，クラスタリングおよびトピックの一貫性という観点において最も良い結果であったと報告している．

また関連した手法として，Tsur ら [19] の提案する事象ごとにハッシュタグをクラスタリングする手法がある．ここでは，ハッシュタグごとに当該のハッシュタグを含むツイートを全て結合した擬似文書を作成し，その擬似文書をクラスタリングする手法を提案している．Tsur らは，特徴ベクトルの作成に TF-IDF ベクトルやハッシュタグの共起ベクトルを用いており，クラスタリング手法には k-means 法を用いている．このほか，井上ら [20] が，Tsur らが英語で行なったハッシュタグのクラスタリングが日本語でも適用できること，福山ら [21] はバースト現象が観測されたハッシュタグでクラスタリングを行うことに

よってバーストしている話題に関連するハッシュタグクラスタが獲得できることを報告している．

本研究では，ツイートをハッシュタグごとに分類し，結合する Hashtag Pooling を手法の一部に採用し，Hashtag Pooling された擬似文書をさらに単位時間ごとに分割した文書を LDA に学習させる．

3. 提案手法

本研究では，ツイートデータをハッシュタグおよび単位時間ごとに分類して結合させた擬似文書を LDA に学習させ，学習後は各擬似文書に対してトピック分布の推論を行う．次に，各擬似文書に対応するツイート数とトピックの分布の積を各擬似文書におけるトピックの推定ツイート数とする．最後に，各時刻におけるトピックごとの推定ツイート数を合計したものを当該時刻におけるトピックの推定ツイート数とする．この結果得られた時系列データに対して，3σ 法を適用することによってバースト現象の検出およびバースト現象が発生した時刻の獲得を行う．以上のアプローチによって，Twitter における潜在的トピックの推移を時系列形式の推定ツイート数として獲得する．この推定ツイート数の時系列データからバースト検出を行い，バースト時刻を求める．提案手法では，LDA の推論アルゴリズムとして確率的変分ベイズ法 [11] を用いる．

3.1 擬似文書の作成

時刻 t におけるハッシュタグの集合を H_t とする．ここでは，H_t による擬似文書を LDA の学習データとする．

トピックモデルの学習データに用いるハッシュタグの擬似文書の表現方法として，共起する単語の Bag of words を用いる．期間 T における時刻を t とする．ある時刻 t におけるハッシュタグ $h_t \in H_t$ が出現するツイートの集合を D_{h_t} とする．ツイート $d_{h_t} \in D_{h_t}$ に語彙 w_i が出現する頻度を $tf(w_i, d_{h_t})$ と表すと，時刻 t におけるハッシュタグ h_t と共起する単語の頻度は以下のように定義される．

$$tf(w_i, h_t) = \sum_{d_h \in D_{h_t}} tf(w_i, d_{h_t}) \tag{1}$$

以上の方法で全てのハッシュタグにおける語の出現頻度を LDA の学習データとする．

3.2 単位時間における推定ツイート数の計算

擬似文書に対応する LDA によって推論されたトピック $k \in K$ の分布を $\boldsymbol{\theta}_{h_t} = (\theta_{h_t,1}, \theta_{h_t,2}, \cdots \theta_{h_t,k}, \cdots, \theta_{h_t,|K|})$ とする．h_t におけるトピック k の推定ツイート数 N_{k,h_t} は (2) によって計算される．

$$N_{h_t,k} = N_{h_t} \cdot \theta_{h_t,k} \tag{2}$$

計算された推定ツイート数 N_{k,h_t} を基に，時刻 t におけるトピック k の推定ツイート数 $N_{t,k}$ を (3) によって計算する．

$$N_{t,k} = \sum_{h_t \in H_t} N_{h_t,k} \tag{3}$$

この計算の結果得られたトピック k の時系列データ $N_k = (N_{1,k}, N_{2,k}, \cdots, N_{t,k}, \cdots N_{|T|,k})$ を期間 T におけるトピック k の推定ツイート数とする．

3.3 トピックのバースト検出とバースト時刻の獲得

3.2 節の方法によって得られた時系列データ $N_k = (N_{1,k}, N_{2,k}, \cdots, N_{t,k}, \cdots N_{|T|,k})$ に対して，以下の条件式を満たす時刻 t をバースト時刻とする．

$$N_{t,k} > E[N_{k\setminus t}] + 3\sqrt{V[N_{k\setminus t}]} \tag{4}$$

本研究では，時系列の単位を 1 日とし，当該時刻 t 以外のデータから平均と標準偏差を求め，3σ 法による閾値を計算する．ここで時系列の単位を 1 日にした理由は，ツイート全体の投稿数が増加しやすい時間帯では誤ってバースト検出される可能性があるためであり，この問題を考慮する現状の制約としてこの単位を設定している．次に 3σ 法によるバースト検出では，過去の出現頻度が 0 の場合，推定ツイート数が 1 ツイートでも観測されればバースト検出されてしまう問題がある．例えば，N_k=(0, 0, 0, 0, 0) の時系列データが与えられた場合，推定ツイート数が 1 ツイートであってもバースト現象と検出される．したがって，本研究ではバースト現象の検出の際に，トピックの推定ツイート数の下限を設ける．予備実験を行なった結果，直感的に下限が低すぎず十分なトピックが検出された推定ツイート数の下限 100 を設ける．

4. 評価実験

本研究では，Koike らの手法と同様にツイートデータから抽出したトピックの中から，バーストしているトピックの検出を行なっている．バースト現象は現実世界の事象に対してリアルタイムに発生すると考えられるため，提案手法によってバースト現象が検出されたトピックと現実世界で発生している事象が同定できることを確認する必要がある．そこで，実際のツイートデータに対して提案手法で得られたバーストしているトピックが，当該期間にて発生した現実世界の事象と共起しているか確認を行った．

4.1 LDA の設定

本研究の検証では，特徴量ベクトルの次元数が膨大になることを防ぎ，かつ，特定の話題を表す傾向の強い品詞に限定することを目的として，LDA の学習データである擬似文章および擬似文書に出現する語彙に以下の制約を設ける．

（1） 語彙の品詞は固有名詞，普通名詞，サ変接続の名詞のみ

（2） 語彙の文字列長は，漢字は 1 字以上，ひらがな，カタカナ，数字，記号では 2 字以上

（3） リツイートを示す「RT」，URL，リプライを示す「@ユーザ名」の文字列は無視

（4） 任意の時刻において (1) から (3) の条件を満たす語彙を含むツイート数が 10 以上のハッシュタグ，10 未満の場合は当該時刻においてそのハッシュタグは考慮しない

各パラメータの設定については，トピック数が 10,000，イテレーション回数が 1,000 回，ミニバッチサイズが 1,000 とする．

4.2 実験データ及び環境

4.1 節で示した条件を満たすハッシュタグは，2012 年 8 月 1 日から 2012 年 8 月 7 日の間に存在したハッシュタグ 965,369

表 1　提案手法によって得られるトピックの例

トピック	α	単語 | $\phi_{k,w}$							
1	0.009	体操	0.319	内村	0.208	演技	0.043	競技	0.031
2	0.008	北島	0.127	競泳	0.116	介	0.079	康	0.071
3	0.007	広島	0.472	野田	0.087	黙祷	0.041	投下	0.031
4	0.004	エジプト	0.128	ブラジル	0.126	永井	0.094	吉田	0.067
5	0.003	浜田	0.149	死去	0.141	幸一	0.135	衆議院	0.087

表 2　トピックの推定ツイート数の推移とバースト時刻

トピック	1 日	2 日	3 日	4 日	5 日	6 日	7 日	バースト時刻 (日)
1	4,412	24,011	3,793	1,981	4,950	2,843	2,244	8 月 2 日
2	4,685	11,708	5,156	1,351	9,403	1,757	2,463	8 月 2 日
3	830	849	1,247	1,134	1,370	6,602	1,603	8 月 6 日
4	1,117	5,333	1,141	37,554	5,506	1,058	4,061	8 月 4 日
5	272	841	599	332	10,954	1,119	541	8 月 5 日

種類中 40,851 種類，各ハッシュタグのコーパスに用いるツイートはハッシュタグと同様の期間に発生した 26,639,495 ツイート中 17,294,548 ツイート，擬似文書数は 117,547 であった．また計算時間は，ツイートデータの学習で 101 分，合計 143 分であった．

実験環境は，OS が Ubuntu 16.04，CPU が Intel Xeon E5-2630 (2.40GHz)8core 2 機，また，実装は Python および Java で行った．形態素解析器は MeCab [22]，LDA は Mimno ら [11] の確率的変分ベイズ法を Java によって実装した．

5. 実 験 結 果

実験データから提案手法によって得られたトピックの一部を表 1 に示す．表 1 に示すトピックは当該期間においてバースト現象が検出されたトピックであり，同様にバースト現象が検出されたトピックは 10,000 トピック中 647 トピックであった．α は各文書におけるトピック分布の事前分布であるディリクレ分布のハイパーパラメータであり，全体で比較した際のトピックの出現のしやすさとみなすことができる．また $\phi_{k,w}$ は，トピック k における語 w の出現する確率であり，この確率の高い上位 4 語を表 1 に示している．

次に，表 1 のトピックの推定ツイート数の推移とバースト時刻を表 2 に示す．バースト時刻においては，いずれのトピックも突出した推定ツイート数となっており，この時刻においてトピックでバースト現象が発生していることがわかる．

この結果をもとに，バースト時刻において当該トピックに関する推定ツイート数の多い上位 4 つのハッシュタグを表 3 に示す．トピック 1,2,4 では，共通してオリンピックに関連するハッシュタグである “オリンピック” や “olympic” が割り当てられている．またトピック 3 では，“広島” や “原爆” など単語としても共起しやすい語がハッシュタグが割り当てられている．トピック 5 では，$\phi_{k,w}$ の高い語とは異なり，ニュースに関連するハッシュタグが割り当てられている．以上の結果から，トピックに対応する単語とハッシュタグは同様の傾向がある語の他に異なる傾向を示すことがわかる．この性質を利用することによって，単語とハッシュタグによる多面的なトピックの解釈が可能となると考えられる．

実験結果から，トピックと関連性があると考えられる現実世界の事象を表 4 に示す．このように提案手法では，バーストしているトピックの情報として，トピック内で出現しやすい単語，バーストした時刻，トピックが割り当てられたハッシュタグといった多様な情報を，図 1 に示すことによって，事象の特定を行うことができる．したがって Koike らや Tsur ら，井上ら，福山らの提案する手法と比較して，トピックに関する情報が豊富でありトピックの解釈がユーザにとって容易になると考えられる．

表 3　バースト時刻における各ハッシュタグの推定ツイート数

トピック	ハッシュタグ | 推定ツイート数							
1	体操	5,440	オリンピック	4,258	olympic	975	mitazo	964
2	オリンピック	3,888	競泳	3,430	olympic	502	水泳	308
3	広島	472	原爆	374	twitr	374	IWJ_HIROSHIMA2	266
4	daihyo	6.880	オリンピック	5,522	サッカー	4,560	なでしこ	2,426
5	nhk_news	4,771	2ch	627	news	615	niconews	467

表 4　トピックと対応する現実世界の事象

トピック	事象	発生時刻 (日)
1	ロンドン五輪男子体操個人総合	8 月 2 日
2	ロンドン五輪男子平泳ぎ北島康介氏 4 位入賞	8 月 2 日
3	広島原爆の日 平和記念式典に野田佳彦氏参列	8 月 6 日
4	ロンドン五輪男子サッカー準々決勝 日本対エジプト戦	8 月 4 日
5	浜田幸一氏逝去	8 月 5 日

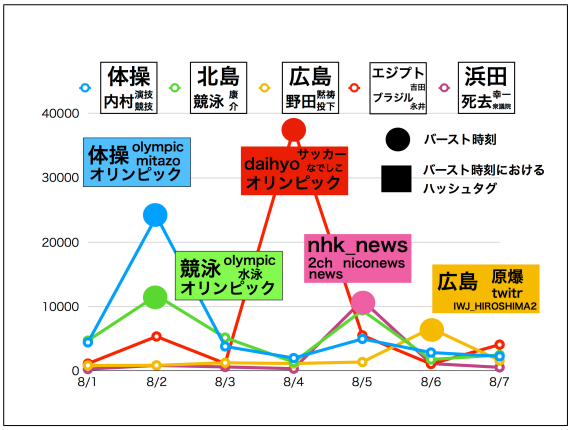

図 1　ハッシュタグと Twitter で流行している話題の関係

6. お わ り に

本研究では，LDA のアルゴリズムを変更せずに，Tweet Pooling の拡張によって，潜在的なトピックの推移の抽出を行う手法の考案を行った．実際に提案手法を用いて，2012 年 8 月 1 日から 8 月 7 日に発生した現実世界の事象に対応するトピックの抽出とそれらのトピックの推移の抽出とバースト検出を行った．この結果，LDA の高速な推論アルゴリズムが利用できるとともに，トピックに出現しやすい語，バースト時刻，トピックが割り当てられたハッシュタグといったトピックに関する豊富な情報から，容易に現実世界の事象との関係性を分析できることがわかった．

今後の課題として，提案手法の有効性を一部のトピックの主観的評価のみで行っている点が挙げられる．そのため，全てのトピックに対して，現実世界の事象と対応しているか確認を行う必要がある．また提案手法のベースラインとなる推論アルゴリズムを DTM に変更した場合のトピックおよび処理時間の比

較も行う必要がある．

謝　　辞

本研究の一部は，JSPS 科研費（課題番号 16H02904）および筑波大学図書館情報メディア系プロジェクト研究の助成によって行われた．

文　　献

[1] Harry Wallop. Japan earthquake:how twitter and facebook helped. *The Telegraph*, 3 2011.

[2] 芥子育雄, 鈴木優, 吉野幸一郎, 大原一人, 向井理朗, and 中村哲. 単語・パラグラフの分散表現を用いた twitter からの日本語評判情報抽出. In **第 *8* 回データ工学と情報マネジメントに関するフォーラム**, 2016.

[3] 荒牧英治, 増川佐知子, and 森田瑞樹. 事実性判定を用いたインフルエンザ流行予測. In **研究報告音声言語情報処理**, volume 2011-SLP-86, pages 1–8, 2011.

[4] 水沼友宏. Twitter におけるバーストの生起要因と類型化に関する分析. Master's thesis, 筑波大学, 2014.

[5] Diao, Qiming, Jiang, Jing, Zhu, Feida, and Ee-Peng Lim. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 536–544, 2012.

[6] Y. Du, W. Wu, Y. He, and N. Liu. Microblog bursty feature detection based on dynamics model. In *International Conference on Systems and Informatics*, pages 2304–2308, 2012.

[7] DONG Guozhong, LI Ruiguang, YANG Wu, WANG Wei, GONG Liangyi, SHEN Guowei, YU Miao, and LV Jiguang. Microblog burst keywords detection based on social trust and dynamics model. *Chinese Journal of Electronics*, 23(4), 2014.

[8] Daichi Koike, Yusuke Takahashi, Takahito Utsuro, Masaharu Yoshioka, and Noriko Kando. Time series topic modeling and bursty topic detection of correlated news and twitter. In *International Joint Conference on Natural Language Processing*, pages 14–18, 2013.

[9] Michael I. Jordan David M. Blei, Andrew Y. Ng. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[10] David M.Blei and John D.Lafferty. Dynamic topic models. In *ICML '06 Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

[11] D. Mimno, M.D. Hohhman, and D.M. Blei. Sparse stochastic inference for latent dirichlet allocation. In *the 29th International Conference on Machine Learning*, 2012.

[12] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373 – 397, 2003.

[13] Yuji Nakai Shimizu Tohru Kentaro Shimizu Koji Kadota, Ye Jiazhen. Roku: An improved method for the detection of tissue-specific expression patterns. *BMC Bioinformatics*, 2006.

[14] 石川栄介. 棄却検定の比較表. **岩手大学学芸学部研究年報**, 15(2):1–7, 1960.

[15] Smeeton Niqel A Sprent Peter. Applied nonparametric statistical methods. *Chapman and Hall*, page 480, 1993.

[16] Sprent Peter. Data driven statistical methods. *Chapman and Hall*, page 406, 1997.

[17] Tomoharu Iwata, Takeshi Yamada, Yasushi Sakurai, and Naonori Ueda. Online multiscale dynamic topic models. In *KDD '10 Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 663–672, 2010.

[18] Wray Buntine Lexing Xie Rishabh Mehrotra, Scott Sanner. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *SIGIR'13*, 2013.

[19] Oren Tsur, Adi Littman, and Ari Rappoport. Efficient clustering of short messages into general domains. In *International Conference on Weblogs and Social Media (ICWSM)*, 2013.

[20] 井上優作 and 若林啓. 表記の多様性を考慮したハッシュタグ推薦. In **第 *14* 回日本データベース学会年次大会**, 2016.

[21] 福山 怜史 and 若林啓. バースト現象を考慮したハッシュタグのクラスタリング手法の提案. **研究報告情報基礎とアクセス技術**（*IFAT*）, pages 1–6, 2017.

[22] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis,. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 230–237, 2004.