人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第7回)
SIG-AM-07-05

# Twitterにおけるトピック遷移分析システムの提案

# A Proposal of a Topic Transition Analysis System for Tweets

田中克明 [1*]

[1] 一橋大学情報基盤センター
[1] Center for Information and Communication Technology, Hitotsubashi University

**Abstract:** In this paper, we propose an interactive system to represent the transition of topics extracted from documents that are generated in chronological order, such as tweets. Many of methods, extracting and visualizing topic transitions in documents generated along the time series aim to show an overview. We implement a system, reorganizing and visualizing topic transitions based on keywords designated by a user, providing interfaces to read the original documents for user to support analyzing topic transitions.

## 1 はじめに

本研究では，Twitter など時間経過に伴い生成される文書集合にふくまれる時間経過に沿ったトピックの遷移を，インタラクティブに提示する仕組みを提案する．時系列に沿って生成される文書からそこに含まれる内容を抽出し可視化する手法の多くは，表示をユーザが目視することにより全体的な概要を理解することの支援を目的とする．それに対し，本稿で提案するシステムは，トピック遷移をそのまま提示するだけではなく，ユーザが指示した単語などの情報に基づき，トピック遷移の一部分を抽出し提示し，さらにトピックに含まれる文書の詳細をユーザが確認可能とすることにより，ユーザの興味に応じ，時間経過に沿ったトピック遷移の分析を支援するシステムを実装，提案する．

本稿では Twitter から取得したツイートからトピックを抽出するために，Probabilistic Latent Semantics Indexing（pLSI）[Hofmann 99] を用いた．pLSI により，文書からトピック $z$，文書中にあらわれる単語 $w$ についてトピックごとの生起確率 $p(w|z)$，各ツイート $d$ のトピックにおける生起確率 $p(d|z)$ などを求めることができる．これらの確率を用いて，実装したシステムでは，ユーザが興味を持った単語の生起確率が大きいトピックからなるトピック遷移を表示，そこから個別のトピックを選択し，トピック内での生起確率が高い単語や含まれるツイートの確認を可能とした．これにより，ツイート本文などを確認した後に，新たに興味をひかれた単語やツイートを指定し，それらを含むトピックの推移をあらためて確認するなどインタラクティブな分析が行える．

*連絡先： 一橋大学情報基盤センター
〒 186-8601 東京都国立市中 2-1
E-mail: sigam07@katsuaki-tanaka.net

## 2 関連研究

### 2.1 トピック遷移の抽出と可視化

本研究において提案するシステムの分析対象である，時間経過に沿ったトピック遷移の抽出のための手法は，トピックモデルに基づく Dynamic Topic Models[Blei 06] などが挙げられる．これらの手法では，時系列に沿って時区間を設け，その区間に対し一定数のトピックを抽出する．また，k-means を拡張し古い情報を忘却するモデルを取り入れたクラスタリング手法 [長谷川 07] の研究もなされている．

抽出したトピックの可視化は，特徴語を並べる，トピック出現確率の推移をグラフ化するなど以外に，全体の傾向を把握しやすいように可視化を行う，Themeriver [Havre 02] や Alluvial Diagram [Rosvall 10] などが研究されている．可視化結果は静的なものに限らず，一部を選択し強調表示するなどの操作が可能なものもある．

トピックの遷移を操作するためには，遷移をトピックとトピック間のリンクからなるグラフ構造とし，Gephi [Bastian 09] などのグラフ構造可視化ツールを用いる方法が考えられる．これにより，遷移の構造を可視化すると同時に，ノード（トピック）の表示・非表示，グラフ構造の変形などの操作を行うことができる．しかし，トピックに含まれる単語ごとの出現確率に応じた操作など，トピックの抽出過程で得られたデータを活かし，グラフの要素であるノードに対し細かな操作を行うためには，グラフの元となるデータの再生成が必要であり，グラフ可視化ツールは，文書・単語からなるトピック遷移のインタラクティブな操作には不十分である．

### 2.2 Twitter データの分析

Twitter のデータに対する分析は，ユーザ間の関係に関する研究 [風間 10][Cha 10]，時系列データとして一時的な増大などに着目した研究 [Sakaki 10][水沼 13]，タイムラインからのツイート間の構造抽出に関する研究 [松尾 14] などがなされている．

また，本稿で扱うデータと同様に,「人工知能」を含むツイートに対し，特徴語の推移，ツイートしたユーザに関する分析などが行われている [鳥海 14]．

## 3 時系列トピック遷移の抽出

提案システムが取り扱う時間の経過に沿ったトピックの遷移をツイート群から以下の手法により抽出した．なお，提案するシステムでは，一定の時区間（区間数 $N$）ごと各自区間における $K$ 個のトピック $z_{n,k}(n = 1,2,\ldots,N,\ k = 1,2,\ldots,K)$ と，トピックの生起確率 $p(z_{n,k})$，各トピックにおけるツイート $d_i$ の出現確率 $p(d_i|z_{n,k})$，単語 $w_m$ の生起確率 $p(w_m|z_{n,k})$ を利用する．これらを求めるために，筆者が人工衛星の設計議事録からのタスク抽出に用いた手法 [Tanaka 11] を改良してトピック遷移の抽出を行った．

### 3.1 前処理

トピック抽出の前に，処理対象とするツイートを，Twitter REST API の search/tweets により収集する．同 API で収集できるツイートは過去約 1 週間分に限定され，長期間にわたり収集するために，定期的な API 呼び出しを行った．得られたツイート群からは，タイムラインでのツイートの扱いを模して公式リツイートを除去した．また各ツイートからは，URL，リツイートまたは引用ツイートを示す「RT」「QT」に続くテキストを取り除いた．これらを MeCab[1]を用いて形態素解析し，名詞および未知語と分類された語とその出現回数を求め，各ツイートに対応する単語ベクトル $d_i$ を得た．

### 3.2 トピックの抽出

処理対象とするツイートのツイートされた時刻に着目し，最も古いものと最も新しいものの間を $N$ の区間に分割，各区間の終了時刻 $t_n$ をもとめる．ここでは，$N = 50$ とした．処理対象とするツイートのうち $t_n(n = 0,1,2,\ldots,N)$ 以下の時刻を持つツイートにより，ツイート集合 $D_n$ を設定し，pLSI により $K$ 個のトピック $z_{n,k}$ を抽出した．ツイート $d_i$ に対して pLSI により求められた $p(z_{n,k})$，$p(d_i|z_{n,k})$ を用いて，

$$\arg\max_k p(d_i, z_{n,k}) = \arg\max_k p(z_{n,k})p(d_i|z_{n,k}).$$

をとる $k$ を持つクラスタ $C_{n,k}$ へ，排他的なクラスタリングをあわせて行った．以後，$z_{n,k}$ あるいは対応する $C_{n,k}$ を，$t_n$ におけるトピックとして扱う．提案システムではトピックの遷移全体の俯瞰ではなく，その中でユーザが着目した部分を扱うことであるため，トピック数 $K$ は大きめにとった．

### 3.3 古いツイートの忘却

新しいツイートと関連を持たないツイートは，古い内容であり，時間の経過に従い徐々に忘れ去られていくと考えられる．そこで，$D_n$ からトピック抽出を行う前に，古いツイートの重みを徐々に減らす忘却の仕組みをもうけた．

古いトピックとは，トピック $z_{n,k}$ に対し，$p(d_i|z_{n,k})$ が大きいものから順に見ていき $\sum_{i\in C_{n,k}} p(d_i|z_{n,k}) \leq S(S = 0.2)$ の間に存在する $d_i$ において，ツイートされた時刻が $t_(n-1)$ より小さい，すなわち新しいツイートを含まないトピックを指すこととした．一方，ツイートされた時刻が $t_n$ より大きい，すなわち新しいツイートを含めば，$z_{n,k}$ を新しいトピックとみなす．

新しいトピックに含まれない $d_i$ に対し，$D_{n+1}$ からトピック抽出を行う際に $R(R \leq 1)$ を乗じ，古いツイート $d_i$ が徐々に忘れ去られるようにした．

### 3.4 トピック間遷移の設定

抽出したトピック間の類似度を以下の $sim(C_{n,i}, C_{n+1,j})$ と定義し，表示時に閾値 $T$ 以上の類似度を持つ $C_{n,i}, C_{n+1,j}$ に対し，リンクを設けた．

$$sim(C_{n,i}, C_{n+1,j}) = \frac{|C_{n,i} \cap C_{n+1,j}|}{|C_{n,i}|}. \quad (1)$$

クラスタなど，複数の要素からなる集合の類似度は，次の Jaccard 係数により求めることが多い．

$$Jaccard(C_{n,i}, C_{n+1,j}) = \frac{|C_{n,i} \cap C_{n+1,j}|}{|C_{n,i} \cup C_{n+1,j}|}. \quad (2)$$

ツイートは時間がたつほど数が増えるため，$C_{n,i}$ に比べ $C_{n+1,j}$ の方が要素数が多いと考えれ，$C_{n+1,j}$ の要素数が大きいと Jaccard 係数は小さな値を示し，類似度が低く判定されるため，(1) を類似度として用いる．

[1] http://mecab.sourceforge.net/

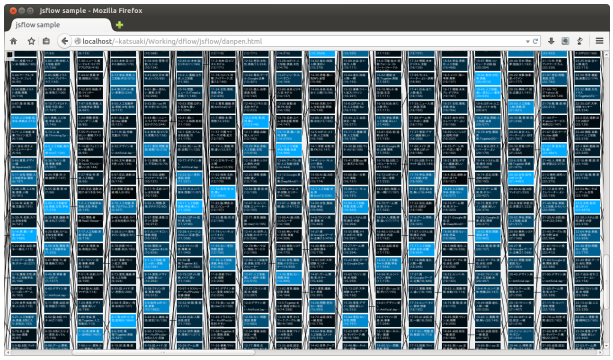


図 1: トピック遷移表示例

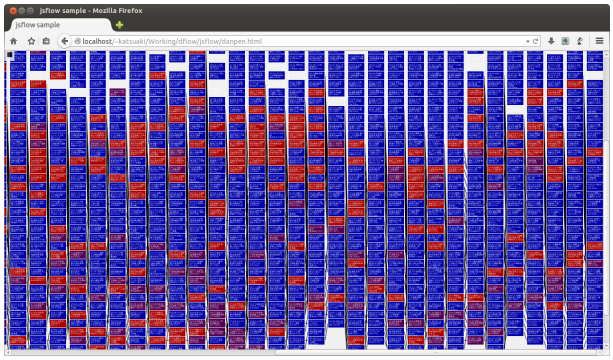


図 2: 「人工知能」(青)「表紙」(赤) を指定した例

### 3.5 トピック遷移の表示

ここでは，トピックをノード，トピック間の類似度が閾値以上のものをリンクとして得たグラフ構造を，時間を横軸にとり表示した．ラベルには，各トピック $z_{n,k}$ において $p(w_m|z_{n,k})$ が大きい語を選択した．表示例を図 1 に示す．

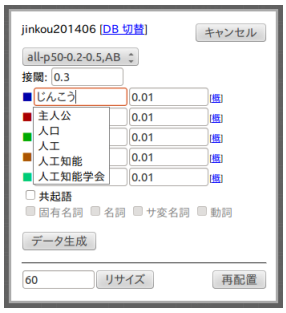
## 4 システム概要

ここから，本研究で提案するシステムで実装した，トピック遷移分析システムの各機能について述べる．

### 4.1 単語の生起確率によるトピック遷移の選択

3.5 にて述べたトピックの遷移全体の表示に対し，本システムのユーザがキーワード $w$ と閾値を指定することにより，キーワードの生起確率 $p(w|z_{n,k})$ が閾値以上のトピック $z_{n,k}$ を選択し，指定された色により表示する．すなわち，トピック遷移のうちキーワードに関連する部分を抽出して表示する．

ひとつのキーワードを指定すれば，そのキーワードを含むトピックを，複数のキーワードを指定すれば，各



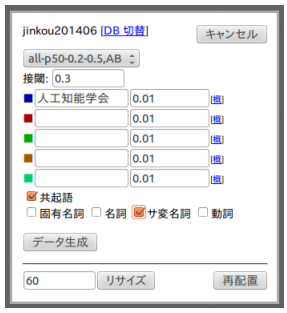




図 3: キーワード入力支援例 図 4: 単語ラベルの指定例

キーワードにまつわるトピックの移り変わりを表示することが可能である．図 2 に例を示す．キーワードの生起確率閾値の設定には，後述する単語出現状況の表示における $p(w|_z n,k)$ の推移が参考になる．

### 4.2 ラベル語の指定

ラベルとしてキーワードと同じツイートに含まれる単語，すなわち共起する単語を選択することを指定すると同時に，形態素解析時に得られた単語の品詞を指定することができるようにした．画面例を図 4 に示す．

キーワードとして文書群に含まれる何らかの「着目対象」を指定すると，着目対象に対してどのような議論が行われていたかを表示できる．同時に，ラベルとして表示する語の品詞として，サ変名詞 (「〜する」と「する」を続けられる名詞) を指定すると，着目対象に対して行われていた行為を抽出できる．これにより，ある対象への作業の一覧を確認することができる．また，時間経過に沿ったトピック抽出を経ているため，同じタイミングで並行して行われていた事象を分離することが可能である．

### 4.3 キーワード入力支援

ユーザがキーワードの入力を行う際，キーストロークを含む単語を文書に含まれる単語リストから取得，再構成用のキーワード候補として表示する仕組みを設けた (図 3)．

入力支援を行うことにより，文書中に確実に存在する単語を確実に入力できるようにすることを目指した．一方，キーワード入力支援を行わない場合，ユーザが，表記の揺れなど含まれる単語を把握した上でキーワードを指定する必要が生じる．また，入力支援により，例えば「人工知能」と「人工知能学会」の両方が単語として本システムに認識されている場合，両者を候補として同時にユーザに表示することにより，語の違いを意識してキーワードを指定する必要性を示せる．

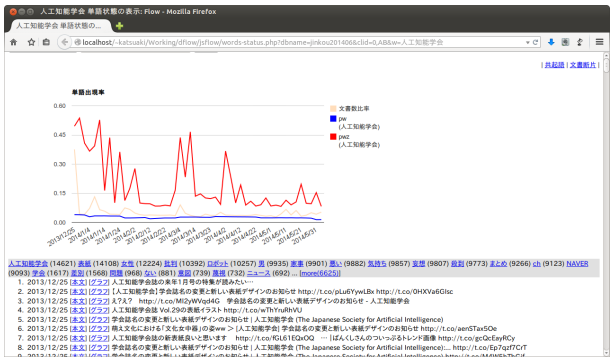


図 5: 単語出現状況の表示例

図 6: トピック詳細の表示例

## 4.4 単語出現状況の表示

キーワードの指定画面から，キーワードとして設定しトピックの選択を行う前に，キーワード候補である単語のトピック遷移内での出現状況を表示させられるようにした．図 5 に例を示す．単語の出現状況表として表示するのは，単語 $w$ について pLSI により求められた $\max p(w|z_{n,k})$ と $p(w)$ の推移を示すグラフ，ツイート内に共起するその他の単語，単語を含むツイートの一覧である．

本表示における単語の出現確率の遷移を示すグラフは，4.1 に述べたトピックの選択表示のためのキーワードと閾値となる $p(w|z_{n,k})$ を設定する支援となる．また，ツイート内で共起する単語の表示を行うことで，複数の単語を指定する場合の 2 番目以降のキーワードの選択を支援することも目指した．

## 4.5 トピック詳細の表示

$z_{n,k}$ において，$p(d_i|z_{n,k})$ の大きい $d_i$ 順，あるいはツイートされて時刻が新しい順に，ツイートを表示する．また，$p(w_m|z_{n,k})$ が大きい順に単語 $w_m$ も表示する．図 6 に例を示す．これにより，トピックの詳細を把握することができる．また，各ツイートについて，Twitter

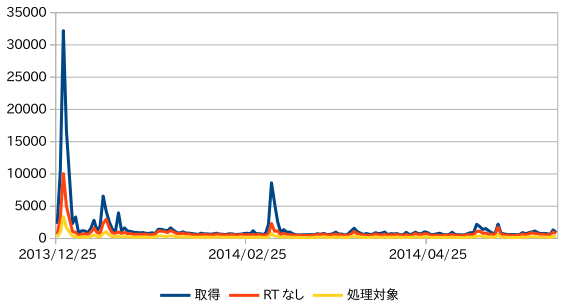



図 7: 取得ツイート・処理対象ツイート数の日次推移

Web クライアント上により実際のツイートを参照できるようにした．

## 4.6 ツイートを含むトピックの表示

単語の詳細表示画面，トピック詳細の表示画面には，単語を含むツイート，トピックに含まれるツイートの一覧が表示される．ここから，ツイート $d_i$ を指定し，$p(d_i|z_{n,k})$ が大きい上位 100 個の $z_{n,k}$ のトピックを選択，表示する機能をもうけた．これにより，指定したツイートがトピック遷移の中でどの期間にわたって主に出現し，どのようなトピックへ含まれているかを確認できるようにする．

# 5 「人工知能」を含むツイートにおける利用事例

処理対象の例として,「人工知能」を検索クエリとして Twitter API により収集，2013 年 12 月 25 日 19 時付近からから 2014 年 6 月 6 日 18 時付近（どちらも JST）までの 235,979 ツイートを得た．これらより 3.1 に述べたように公式公式リツイートを除去した 131,522 から，以後の処理では処理量を減らすために，約 $\frac{1}{3}$ にあたる 43,862 ツイートをランダムに選択した．選択されたツイートに 3 以後のトピック遷移抽出処理を行い，以後の事例確認に用いた．Twitter より取得したツイート数と処理対象としたツイート数などの日ごとの推移を図 7 に示す．

## 5.1 トピック遷移の選択とラベル語指定

「人工知能学会」「表紙」の 2 つの単語を指定してトピックの抽出を行うと，両者が混じり合いながらツイートが続いている様子がわかる．このうち,「表紙」が含まれないトピックの一部を確認すると，人工知能学会

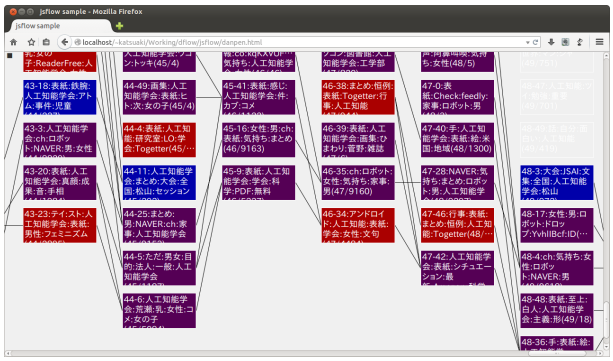


図 8: 「人工知能学系」(青)「表紙」(赤)を指定しラベルを共起する名詞にした例

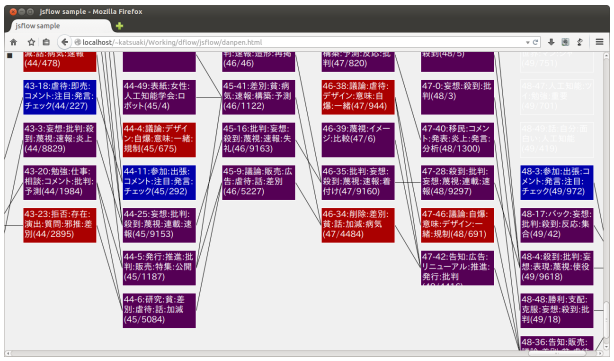


図 9: 「人工知能学会」(青)「表紙」(赤)を指定しラベルを共起するサ変名詞にした例

全国大会について述べているツイートを含むトピックであった．トピックの遷移表示において，ラベル語を共起する名詞とした場合を図 8 に，共起するサ変名詞とした場合を図 9 に示す．名詞をラベルにすると，どのような事象があったかを確認でき，サ変名詞をラベルにすると，どのような意図の記録としてツイートされているかを確認することが，おおよそ可能である．

### 5.2 トピックとツイートの参照

「人工知能」を含むツイートを分析した研究 [鳥海 14] にて，BBC などにて人工知能学会誌表紙が取り上げられた旨の記述があることから,「BBC」について確認した．はじめに図 3 のキーワード入力画面にて「BBC」を入力しようとしたところ，4.3 のキーワード入力支援により「BBC」が候補として表示され，ツイートに現れ単語として認識されていることがわかった．続いて 4.4 の単語の状況表示より,「BBC」のトピック遷移中での出現確率の推移を確認した．これに基づき，トピック中に「BBC」が出現すると判断する閾値を設定，4.1 のトピック選択を実行する．選択表示されたトピックの詳細を 4.5 のトピック詳細表示により表示することにより,「BBC」を含むトピックに含まれるツイートを確認することができる．この際，図 6 にも示したトピックより,「AFP」も表紙に関わる報道を海外向けに行なっていることがわかった.「BBC」同様に「AFP」について確認を行うと，AFP がいくつかの国にニュース配信を行ったことに触れたツイートを発見できた．

## 6 考察

本稿では，大量のツイートに対し，トピック抽出により得られたトピックの遷移をユーザの指示するキーワードに基づき提示する機能，それらトピックに含まれるツイートや単語の詳細を確認する機能などを持った分析システムを提案，実装した．本システムでは，ユーザが興味をひかれた事象について，代表的な単語などにより全体を俯瞰するだけではなく，個々のツイートに含まれる内容を読み込むことが可能である．これにより，ユーザがはじめに興味をひかれた事象の詳細を確認するうちに，あらたな事象に興味を持ち，トピックの遷移全体での新たな興味対象の位置づけを確認し詳細を読み込むという行為を繰り返し，ネットサーフィンに似たような形で，トピックの遷移を確認していくことができる．

既存のトピックの遷移抽出や抽出結果の提示手法は，抽出対象とした文書集合全体におけるトピック遷移の位置づけの提示を主な目的としている．また，対象として，報道記事や論文を扱っており，結果を見る側が処理され提示される文書集合の内容に対し，ある程度の知識を持っていることが暗黙の前提になっていると考えられる．例えば，今回取り扱った「人工知能」を含むツイートにおいて，学会誌の表紙について議論が起きたことを知っているため,「家事」「批判」などの特徴語の表記で何が議論されているかわかが、知らなければ人工知能と「家事を批判すること？」の関係は類推しづらい．

一方，本稿で提案するシステムでは，複数のトピックが提示され，含まれるツイートをひとつずつ確認することが可能であり，興味を持った部分から詳細を読み進めることにより，内容に関する前提知識がなくても，理解できる文から読みはじめることができる．

このように，Twitter 上の情報の理解を支援することが可能ではあるが，研究を進めるためは，どのような内容についてどの程度の支援が可能であるか，評価を行なう必要がある．

また，分析対象としたツイートを確認すると，それぞれある事象について感想などの形で言及が多く，現実世界のコピーとしての情報が多い．そのため，時間経過に沿って抽出されたトピックを確認すると，新しく雑誌が発行されたなど起こった事象は反映されてい

るが，Twitter 内部で時間の経過に沿って進んだ議論を見つけることができない．Twitter 自体のもつ性質にも依存するが，Twitter 上で，言及のあとどうするかという「議論」が起きていないわけではない．例えば，キーワードを含むリツイートに続いて投稿されたツイートは，リツイート内容についての意見を含んでいる可能性がある．キーワードを指定した検索では，言及以外のツイートを収集することが出来ないため，検索結果の前後のツイート，リプライ関係にあるツイートなども含めて収集し，分析対象とするか判断する必要があろう．

現在の提案システムの実装では，どのトピックを参照したかなどの履歴が残らないため，意図せず繰り返し同じトピックを参照してしまうなど，操作上の不都合がある．トピックやツイートにブックマークをつける，メモを付記するなどの機能，キーワードの生起確率によりトピックを選択する際にキーワードを含まないトピックを選択する機能，トピック選択時のデータ生成速度，トピックを示すグラフ上のノードの色付け方法など，改善すべき点が多い．

## 7 おわりに

本稿では，長期間の大量のツイートに対し，そこに含まれるトピックの遷移をユーザの興味に基づいて表示しつつ，ツイート本文までユーザが読み進めることができる仕組みを実装した．提案したシステムにより，概要を眺めるのでもツイートをすべて読むのでもなく，ツイートの拾い読みを支援するような形で，面白そうな部分を渡り歩くことが可能である．今後，分析対象とするデータと分析システムの整備を行いつつ，提案システムによる分析でどのような事柄の理解が可能か，評価を進めたい．

## 参考文献

[Bastian 09] Bastian, M., Heymann, S., and Jacomy, M.: Gephi: an Open Source Software for Exploring and Manipulating Networks, in *Proceedings of Third International AAAI Conference on Weblogs and Social Media*, pp. 361–362 (2009)

[Blei 06] Blei, D. M. and Lafferty, J. D.: Dynamic Topic Models, in *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120 (2006)

[Cha 10] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K.: Measuring User Influence in Twitter: The Million Follower Fallacy, in *Proceedings of Fourth International AAAI Conference on Weblogs and Social Media*, pp. 10–17 (2010)

[Havre 02] Havre, S., Hetzler, E., Whitney, P., and Nowell, L.: Themeriver: Visualizing Thematic Changes in Large Document Collections, *Visualization and Computer Graphics, IEEE Transactions on*, Vol. 8, No. 1, pp. 9–20 (2002)

[Hofmann 99] Hofmann, T.: Probabilistic Latent Semantic Indexing, in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57 (1999)

[Rosvall 10] Rosvall, M. and Bergstrom, C. T.: Mapping Change in Large Networks, *PloS one*, Vol. 5, No. 1, p. e8694 (2010)

[Sakaki 10] Sakaki, T., Okazaki, M., and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, in *Proceedings of the 19th international conference on World wide web*, pp. 851–860 (2010)

[Tanaka 11] Tanaka, K. and Hori, K.: Extracting Tasks in Design Process Records, in *Proceedings of Eighth International Joint Conference on Computer Science and Software Engineering*, pp. 373–378 (2011)

[松尾 14] 松尾 哉太, 新妻 弘崇, 太田 学：Twitter タイムラインの話題の可視化の一手法, 第 6 回データ工学と情報マネジメントに関するフォーラム (2014)

[水沼 13] 水沼 友宏, 池内 淳, 山本 修平, 山口 裕太郎, 佐藤 哲司, 島田 諭：Twitter におけるバーストの生起要因と類型化に関する分析, 情報社会学会誌, Vol. 7, No. 2, pp. 41–50 (2013)

[長谷川 07] 長谷川 幹根, 石川 佳治：T-Scroll: 時系列文書のクラスタリングに基づくトレンド可視化システム, 情報処理学会論文誌, Vol. 48, pp. 61–78 (2007)

[鳥海 14] 鳥海 不二夫, 榊 剛史, 岡崎 直観：「人工知能」の表紙に関する Tweet の分析 (小特集「人工知能」表紙問題における議論と論点の整理), 人工知能：人工知能学会誌：journal of the Japanese Society for Artificial Intelligence, Vol. 29, No. 2, pp. 172–181 (2014)

[風間 10] 風間 一洋, 今田 美幸, 柏木 啓一郎：Twitter の情報伝播ネットワークの分析, 第 24 回人工知能学会全国大会 (2010)