時系列ニュースの話題性の可視化
Visualization of the Topicality of News Streams

中山　豪
Go Nakayama
法政大学情報科学部コンピュータ科学科
E-mail: go.nakayama.5v@stu.hosei.ac.jp

Abstract

We are able to read much news online with mobile phones and personal computers. However, it takes much time to read all new articles, and it is difficult to find articles that we really want to read. To solve the problem, this paper proposes a method for visualizing the topicality of news streams. The method calculates the topicality of news streams by using the dynamic topic model and the burst detection algorithm, and it visualizes the result by using ThemeRiver and dynamic queries. Although ThemeRiver visualizes time-series data, it is not necessarily appropriate for the comparison of separately located topics. The characteristic of the proposed method is that it allows users to adopt dynamic queries to compare news topics by examining dynamic visualization results. For this analysis, the method obtains news from the RSS feeds of five Japanese major newspaper publishers and three overseas news agencies. This paper presents several examples of the experimental use of the method. Also, it shows that users can get to know news with high topicality and can reduce time to choose the articles that they want to read.

1. はじめに

近年，新聞記事や雑誌の電子化がますます進む状況において，新聞社などが Web 上で配信する数多くの記事を PC やスマートフォンで閲覧できるようになった．無料で読めるニュースが多いことは，ニュースを読む人にとってメリットである．一方でこの現状は情報過多であると言え，ニュースを選ぶ手間や時間がかかり，読者の負担が増えているととらえることもできる．本研究は，この負担を減らすことを目的としている．

先行研究で開発されたシステムの 1 つとして T-Scroll [1]が挙げられる．詳細については関連研究の節で説明するが，有用ではあるが使いやすさや分かりやすさに欠けるインタフェースであるという実験結果が出ている．したがって本研究では，ユーザにとって分かりやすい可視化結果が得られることに重点を置く．

本研究の提案手法は，動的トピックモデル [2]とバースト解析 [3]を組み合わせてニュースを解析し，ThemeRiver [4]と動的検索(dynamic query) [5]を用いて可視化するというものである．ThemeRiver は要素（本研究ではトピック）の時間軸に沿った変化を示すのに適しているが，離れた要素の比較は難しい．動的検索を用いて可視化結果を動的に変化させることでこの問題を解決する．ニュースの解析において，この可視化の方法を提案することが本研究における新規性である．また，ユーザが必要とする情報を与えることができるという有用性や，ニュース読者の負担を減らすという目的の達成に繋がる．

2. 関連研究

2.1. T-Scroll

ニュースの話題性の可視化システム T-Scroll [1]では，トピックを代表するキーワードをノードとし，キーワード同士の関係性の深さをエッジとして表現している．なお本研究とは違い，クラスタリングによってトピックを分類している．ノードの大きさはトピックの規模，色はクラスタの質を表している．トピックは時系列に並べられており，左から右へ見ていくことで，ユーザは時系列文書集合のトレンドやトピックの詳細を把握することができる．評価実験の結果，インタフェースの使いやすさや分かりやすさの改善が必要とされた．

2.2. ThemeRiver

ThemeRiver [4]は，分類された Theme（本研究ではトピックに該当）に関する情報の時間的推移を可視化するシステムである．図 1 に示されるように，データの時間経過による増減が一目で分かり，時系列データの全体像を掴むことに長けている．本研究では，この上下対称な川の流れを表すような可視化手法を取り入れたが，本来の ThemeRiver にある Theme を代表するキーワードは表示していない．また ThemeRiver の欠点である，離れた Theme 同士の比較が難しいことに対する解決策として，スライダによる動的検索を導入した．

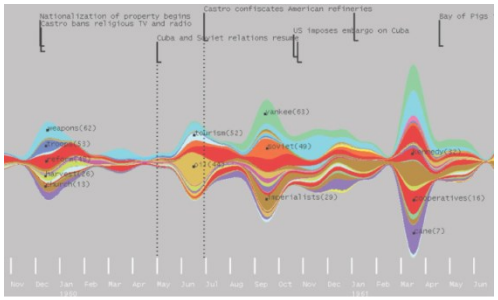

図 1 ThemeRiver [4]の概観．

Supervisor: Prof. Hiroshi Hosobe

3. 準備

3.1. 動的トピックモデル

トピックモデルは，ビッグデータから有用な情報を抽出する手法として注目を集めている．本研究では，Blei らが考案したトピックモデルの拡張の 1 つである動的トピックモデル(dynamic topic model) [2]を用いる．動的トピックモデルは文書データの時系列情報を扱うことができる．本研究でおこなうような新聞記事の解析をはじめとして，論文やブログの解析等に適している．

本研究における動的トピックモデルの役割は，取得したニュースを任意の数のトピックに分類し，トピックごとに関連性の高いキーワードを算出することである．具体的には，それぞれのニュースのタイトルに含まれるキーワードの集合をw，出力させたいトピックの数をKとすると，それぞれのトピック $z_i\,(i = 1, \dots, K)$のwの確率分布は$p(w \mid z_i)$と推定できる．また，文書dにおけるトピックz_iの確率分布は$p(z_i \mid d)$と推定できる．これら$p(w \mid z_i)$と$p(z_i \mid d)$の推定には Blei らが作成したライブラリを使用した．

3.2. バースト解析

バースト解析 [3]とは，Kleinberg によって考案されたバースト検知アルゴリズムを用いて，時系列データにおいて特定の情報の急激な増加を検知することができる解析手法である．

ニュースを解析する場合を例に説明すると，社会的に重大な事件 A が発生した後に，事件 A に関するニュースは一定期間多くなる．この期間を含むニュース記事の集合をバースト解析にかけると，事件 A に関するトピックのバースト度が高いと算出される．このトピックの状態をバースト状態と呼ぶ．逆に，あるトピックに関する報道が普段通りの数しかなされていない場合，そのトピックは非バースト状態である．本研究ではバースト状態のトピックを話題性の高いトピックとする．

本研究では，ニュースのデータを 1 時間に 1 回という離散時間で集めるため，enumerating バーストという手法を用いる．便宜上，引き続きニュースの解析と事件 A の例を使って説明する．まずn時間で集められるニュースデータの集合の内，ある 1 時間分の集合を$B_i\,(i = 1, \dots, n)$とし，B_iにおける記事の総数をd_i，事件 A に関係する記事の総数をr_iとする．非バースト状態を表す式は，ニュース収集期間における期待値である

$$p_0 = \frac{\sum_{i=1}^{n} r_i}{\sum_{i=1}^{n} d_i}$$

とする．またバースト状態を表す式は，p_0に任意のパラメータsをかけた

$$p_1 = p_0 s$$

とする．なお，このパラメータsは$s > 1$かつ$p_1 \leq 1$となる値でなければならない．パラメータsは，バースト状態と判定されやすいかどうかを決めるものである．sが 1 に近ければ近いほど，事件 A に関する記事の割合が少なくてもバースト状態であると判定されやすいという特性を持つ．本来 Kleinberg のバースト検知アルゴリズムは，トピックではなく文書内のキーワードのバースト度$\mathrm{bw}(i, w)$を算出するもので，

$$\mathrm{bw}(i, w) = -\log\left[\binom{d_i}{r_i} {p_0}^{r_i}(1 - p_0)^{d_i - r_i}\right] - \left(-\log\left[\binom{d_i}{r_i} {p_1}^{r_i}(1 - p_1)^{d_i - r_i}\right]\right)$$

で求められる．これをトピックのバースト度$\mathrm{bz}(t, z_i)$の算出に応用すると，

$$\mathrm{bz}(t, z_i) = \sum_{w} \mathrm{bw}(i, w) \cdot p(w \mid z_i)$$

とすることができる．

4. 提案手法

4.1. 概要

本研究で提案するシステムは，ニュースの解析を動的トピックモデルとバースト解析の組み合わせでおこない，解析結果の可視化には ThemeRiver と動的検索を用いる．

動的トピックモデルを用いることで，全てのニュース記事を任意の数のトピックとしてまとめ上げることができる．これらのトピックごとのバースト度を算出し，値が高いものを話題性の高いトピックとする．

解析結果の可視化を ThemeRiver のような積み上げグラフでおこなうことで，トピックの話題性の増減が一目で分かる．しかし，グラフ上で離れたトピック同士を比較することには向いていない．したがって本システムでは，スライダを用いた動的検索を導入してこの問題を解決する．

4.2. 単純な 2 トピック比較

図 2 は，2 トピック比較のための動的検索の例を示している．動的検索のスライダを動かしていくと，総当り的に 2 つのトピックの色が濃くなり，2 つのトピックの単純な比較を支援する．可視化結果のグラフは，流れの太さが話題性の高さを表しており，左から右へ見ていくことでトピックの話題性の時間的推移を確認することができる．図 2 では，12 月 11 日から 14 日にかけて高い話題性を示したトピックがあることを確認することができる．

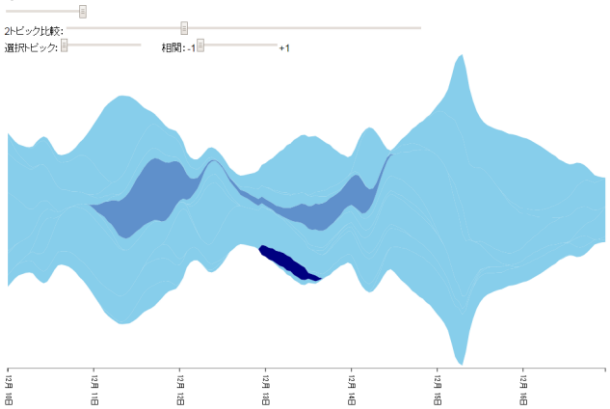

図 2 単純な 2 トピック比較の例．

4.3. 相関を用いたトピック比較

図 3 と図 4 は，トピック同士の相関係数を計算した結果をもとに，特定のトピックに対して任意の相関係数を持つトピックの動的検索をおこなう例を示している．まずユーザは，選択トピックのスライダで比較対象としたいトピックを選択する．このトピックはグラフにおいて最も濃い色で表示される．次に相関スライダで数値を選ぶ．このスライダは 20 ステップで，一番左の状態だと–1.0 から–0.9 を，1 つ右にスライドさせると–0.9 から–0.8 を選択したことになる．比較対象となったトピックと相関スライダで選択した範囲の相関係数を持つトピックが次に濃い色で表示され，それ以外のトピックは最も薄い色で表示される．

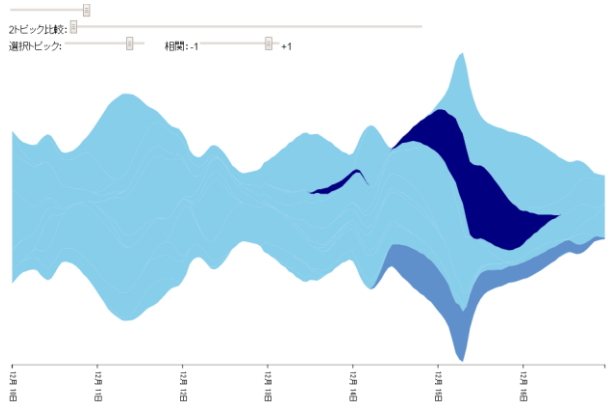

図 3 選択トピックと正の相関を持つトピックの表示．

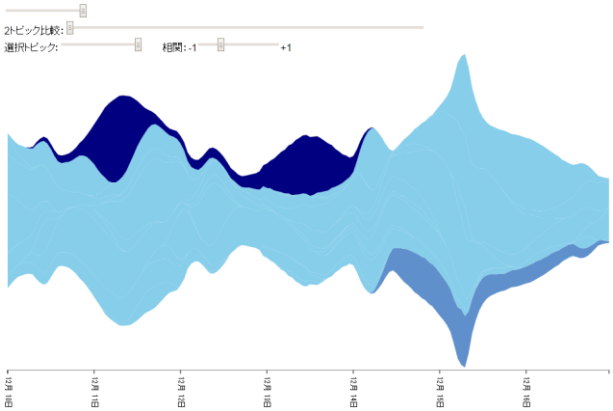

図 4 選択トピックと負の相関を持つトピックの表示．

相関係数の計算は以下のようにおこなった．期間nのある時点iにおける 2 つのトピック A と B の話題性をそれぞれA_i, B_iとし，A と B の話題性の平均を$\bar{A}, \bar{B}$とすると，A と B の相関係数ρ_{AB}は

$$\rho_{AB} = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\{\sum_{i=1}^{n}(A_i - \bar{A})^2\}\{\sum_{i=1}^{n}(B_i - \bar{B})^2\}}}$$

となる．相関係数が+1 に近ければ 2 つのトピックは高い正の相関をもち，同じような増減をする（図 3）．逆に–1 に近ければ，反比例したような増減をする（図 4）．

以上のように，スライダを使った可視化結果の動的な変化を示すことによって，ユーザは話題性の高いトピックやトピック同士の関連性を知ることができる．

5. 実装

本研究では，ニュースを収集するシステム，収集したニュースを解析して話題性を算出するシステム，結果を可視化するシステムを実装した．

ニュース収集システムは，日本の主要 5 紙（朝日・産経・日本経済・毎日・読売新聞）と海外の通信社 3 社（AFP・CNN・ロイター）の RSS 配信を 1 時間に 1 度取得する．使用したプログラミング言語は Java である．

次に，取得したニュースのタイトルを Java の日本語形態素解析器 Kuromoji を用いて形態素解析し，キーワードとなり得る名詞を抽出する．このとき，記号などが紛れ込まないようにするなどのクリーニングも同時におこなう．そしてキーワードごとに出現した回数，含まれるニュースの URL，共起するキーワードを記録する．この記録をもとに，動的トピックモデルを用いて複数のトピックに分類し任意の期間におけるバースト度を算出する．この 2 つ部分に関しては公開されているパッケージを用いた．

算出した話題性は，JavaScript のライブラリ D3.js を用いることで ThemeRiver のように上下対象の積み上げグラフとして可視化する．ユーザはこの可視化結果を Web ブラウザ上で閲覧することができる．

6. 実験

実験に使うデータは 2015 年 8 月 1 日から 12 月 31 日に取得したニュースである．また，解析の対象は 12 月 10 日 0 時台から 12 月 16 日 23 時台に取得したニュースである．本実験では，一般的にニュースを読む目的である「話題性の高い最新トピックについて知ること」を目指した，本研究において開発した話題性可視化システムの使い方を提案するとともに，本研究の有用性を示す．

まず，上記のデータをニュース解析のシステムにかけた結果を入力とし，可視化システムを起動する．起動した画面では図 2 から図 4 で示したように，ThemeRiver を用いた可視化結果が表示される．この時点で，ThemeRiver の効果として流れが太い要素，つまり話題性の高いトピックがある程度分かる．ここで筆者が提案するのは，2 トピック比較のスライダを用いて，流れが太い 2 つの要素を探す作業をすることである．なお，話題性が最大の 2 要素を厳密に探す必要はない．スライダを動かしているうちに話題性の高い要素がおおよそ見つかるのが，動的検索を用いた利点である．

この作業をおこなって見つけ出した 2 つの要素 A と要素 B を図 5 に示す．12 月 10 日頃話題性が高く，12 月 15 日頃から再び話題性が高くなったトピックを表す要素を A とした．また，12 月 11 日から 12 月 14 日にかけて話題性が高いトピックを表す要素を B とした．それぞれデータを参照すると，A のトピックが含むキーワードはオリンピック関係のものであり，B はバドミントンとフィギュアスケートに関するトピックであった．

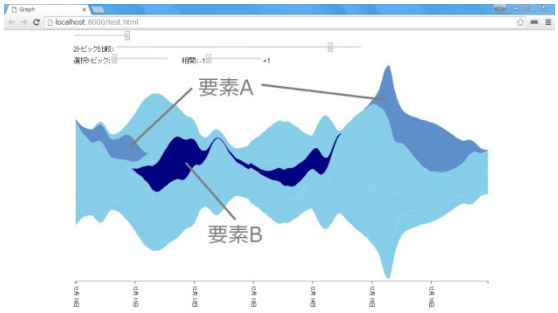

図 5 見つけ出した 2 要素.

ニュースを読むことにあてられる時間は場合によって様々である．A と B のトピックに関するニュースを読んで，まだ時間が次の作業に進むことを提案する．次は 4.3 節で説明したように，相関スライダを用いて別のトピックについても見ていく．例えば要素 A に関するトピックに興味があれば，A と高い正の相関をもつ要素をスライダ操作によって見つける．これが A と関連をもつことは保証できないが，動的トピックモデルによって統合できなかった，似ているトピックを表す要素の可能がある．逆に，A に関するトピックのニュースは十分だと感じたら，A と負の相関をもつ要素をスライダ操作によって見つける．この例を図 6 に示す．なお，この作業をある程度繰り返すことで，解析期間におけるトレンドやトピック同士の関連性を掴むことができる．

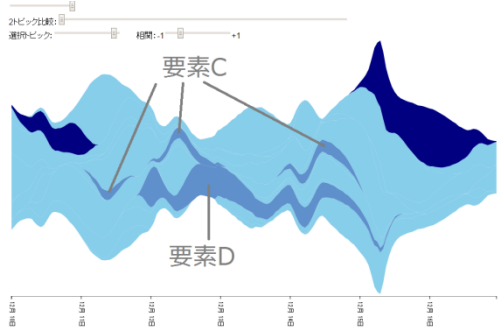

図 6 要素 A と負の相関をもつ 2 つの要素.

要素 C よりも長期間にわたって高い話題性を示した要素 D について詳しく見る．要素 D を表すトピックが含むキーワードはパリ協定に関するものであった．

以上の結果を，Google の急上昇ワードのランキングと比較していく．要素 A について，オリンピック関係の急上昇ワードは 12 月 10 日には見られなかったが，15 日には「新国立競技場」が 4 位である．要素 B については，12 月 11 日は「羽生結弦」「グランプリファイナル（この期間中におこなわれていた大会）」がそれぞれ 1 位と 6 位，12 日は「浅田真央」が 1 位，14 日は「安藤美姫」が 2 位と，フィギュアスケートに関するキーワードが急上昇している．このことから，本システムで話題性の高いトピックとされたものは，これを見る限りでは社会的にも話題性の高いものであることが分かった．

7. 議論

前節において，要素 A も B もスポーツに関するトピックであったが，意図的な選択はおこなっていない．そもそもバースト解析の特性として急激な情報の増加を検知するため，日々ある程度報道されている政治や経済に関するトピックよりも，スポーツなどのトピックがバースト状態になりやすい．

要素 B のトピックで違う競技のキーワードが混ざった原因として，先述のフィギュアスケートの大会と同時期に「バドミントン SS ファイナル」という大会も開催されていたからだと分かった．これを含む問題を解決するためには，より長期間のニュースを取得し，より多くのトピックに分割して解析する必要がある．

要素 C について実験の節では言及しなかったが，要素 C を表すトピックには政治に関するキーワードが含まれていた．要素 C のような細切れのトピックを可視化して得られる情報は少なく，これに対する改善が必要である．

8. おわりに

本論文では，時系列ニュースの話題性を可視化するシステムとその使い方を提案した．ニュースサイトで読むニュースを選んだり，多くの記事を読んだりする負担をかけずに，ユーザは話題性の高いトピックやトレンドを知ることができることを示した．

追加の実験として，どれほどニュース読者の負担が減るかなどを，実際にニュースサイトでニュースを読む場合と比較などが考えられる．今後の課題として，ニュースの収集・解析・可視化のプログラムが 1 つのシステムとしてまとまっておらず，収集から可視化まで自動でできるようにすることが必要である．また，可視化結果からニュース記事へのリンクを表示して，すぐにトピックに対応するニュースを読めるようにする必要がある．

JavaScript でプログラムした本可視化システムの Web での公開に対するハードルは低い．現状の課題を解決した後に，一般ユーザに利用してもらうことで新たな改善点の発見が見込め，本研究の発展が期待できる．

文　献

[1] 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：データベース, vol. 48, no. SIG 20 (TOD 36), pp. 61-78, 2007.

[2] D. Blei and J. Lafferty, "Dynamic Topic Models," *Proc. ICML,* pp. 113-120, 2006.

[3] J. Kleinberg, "Bursty and Hierarchical Structure in Streams," *Proc. KDD,* pp. 91-101, 2002.

[4] S. Havre, B. Hetzler and L. Nowell, "ThemeRiver: Vizualizing Theme Changes over Time," *Proc. InfoVis,* pp. 115-123, 2000.

[5] B. Shneiderman, "Dynamic Queries for Visual Information Seeking," *IEEE Software,* vol. 11, no. 6, pp. 70-77, 1994.