時系列中国語ニュース・ブログにおける トピックモデルの推定と比較対照分析*

鄭 立儀† 胡 碩† 小池 大地† 宇津呂 武仁‡ 吉岡 真治§ 神門 典子¶
筑波大学大学院 システム情報工学研究科† 筑波大学システム情報系‡
北海道大学大学院 情報科学研究科§ 国立情報学研究所¶

1 はじめに

ウェブ上の世界を始めとして，膨大な情報が溢れ，いわゆる情報爆発が起こっている．ウェブ上のニュースやブログの上に限っても，同様に多くの情報が流れている．これらのことを背景にして，時系列に沿って，情報を集約したり俯瞰的に把握するための技術が注目されている．例えば，本研究で利用した統計的トピックモデルのように，文書集合における主要なトピックを推定する技術が確立されてきた．

トピックモデルにおいては，文書が生成される背景において，潜在的に複数のトピックが寄与していることを想定し，文書の生成尤度を高めるようにモデルのパラメータを訓練する．トピックモデルの一種である潜在的ディリクレ配分法 (LDA, Latent Dirichlet Allocation) [1] は，与えられた文書集合から，文書ごとのトピックの確率分布と，トピックごとの語の確率分布を学習する．

以上の研究の成果をふまえて，本論文では，中国語の時系列ニュースおよびブログを対象として，教師なしの (時系列) トピックモデルを適用し，多様なニュースサイトにおいてバーストするトピックを中心として，中国語ブログの収集を行い，中国語における各新聞社の間，および，ニュースとブログの間での関心事項の違いや意見の有無について分析する．

具体的には，中国語において，報道姿勢，報道内容，政治的立場の異なる多様なニュースサイト，および，ブログを対象として，関心事項の違いや意見の有無について分析する．図 1 の流れに示すように，まず，各新聞社ごとに時系列のニュース記事を収集し，LDA を適用してトピックの分布を推定する．次に，ニュースにおいてバーストするトピックを中心として，中国語ブログの収集を行った後，ニュース記事およびブログ記事を全て混合した記事集合に対して再度 LDA を適用し，新聞社・ブログといった情報源間のトピック分布の異なりを分析する．分析結果の事例を図 2 に示す．

2 トピックのモデル化

2.1 トピックモデル

本研究では，トピックモデルとして，潜在的ディリクレ配分法 (LDA, Latent Dirichlet Allocation) を用いる．このトピックモデルにおいては，語 w の列によって表現される時間情報を含んだ文書の集合と，トピック数 K を入力とし，各単位時間について，各トピック $z_n(n = 1, \ldots, K)$ における語 w の確率分布 $p(w|z_n)(w \in V)$，及び，各文書 b におけるトピック z_n の確率分布 $p(z_n|b)(n = 1, \ldots, K)$ を推定する．ここで，V は文書中に出現する語の集合である．本論文では，$p(w|z_n)(w \in V)$，及び，$p(z_n|b)(n = 1, \ldots, K)$ の推定においては，LDA によりトピックモデルを推定するツール GibbsLDA++[1] を用いた．ハイパーパラメータ α，β と，トピック数 K は，$\alpha = 2.5$，$\beta = 0.1$，$K = 20$ とした．

2.2 文書とトピックの対応付け

本研究では，一つのニュース記事，あるいは，ブログ記事に対して，トピックを一意に割り当てることで，トピックごとの記事集合の要素数を測ることとした．

ある日における文書集合を D，トピック数を K，一つの文書を $d(d \in D)$ とすると，トピック $z_n(n =$

*Estimation and Comparative Analysis of Topic Models in Time Series Chinese News and Blogs
†Liyi Zheng, Shuo Hu, Daichi Koike, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Takehito Utsuro, Faculty of Engineering, Information and Systems, University of Tsukuba
§Masaharu Yoshioka, Graduate School of Information Scienceand Technology, Hokkaido University
¶Noriko Kando, National Institute of Informatics

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

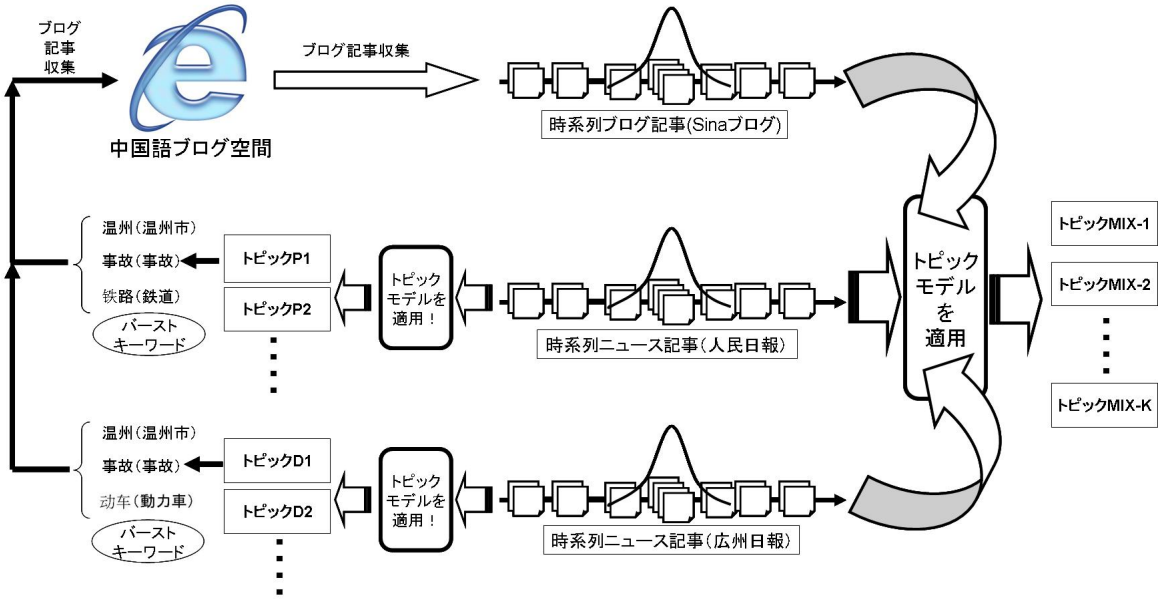


図 1: トピックモデルを利用した時系列の中国語ニュース (人民日報・広州日報)・ブログの比較対照分析の流れ

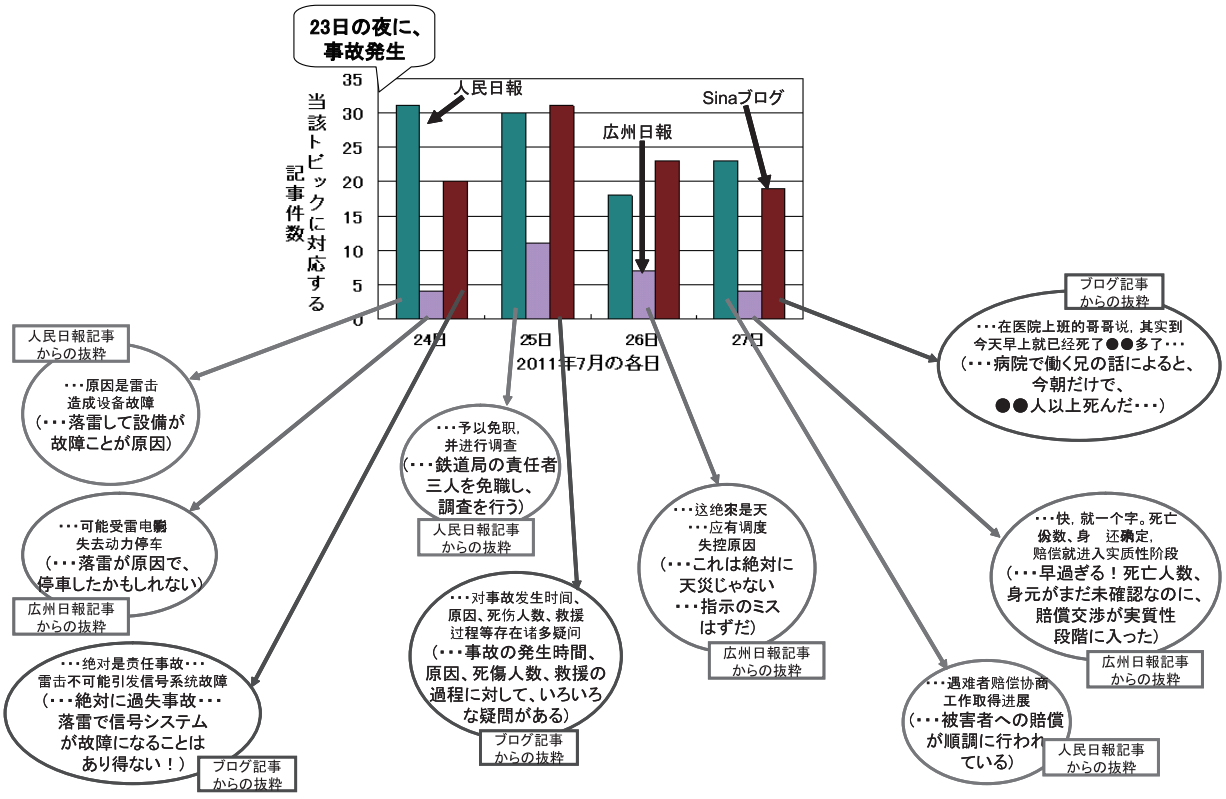


図 2: 中国語ニュース (人民日報・広州日報)・ブログにおける報道内容・関心・意見の比較対照分析結果の例

$1, \ldots, K$) の記事集合 $D(z_n)$(ニュース記事・ブログ記事の和集合) は以下の式で表される.

$$D(z_n) \quad = \quad \left\{ d \in D \middle| z_n = \underset{z_u (u=1,\ldots,K)}{\operatorname{argmax}} \; p(z_u|d) \right\}$$

これはつまり，文書 d におけるトピックの分布において, 確率が最大のトピックに，文書 d を割り当てていることになる.

人民日報（2011.7.18～2011.7.31） → 5937記事	
バーストするトピック	特徴的なキーワード（26個）
Topic②： ノルウェーテロ事件	爆炸(爆発)、挪威(ノルウェー)、枪击(銃撃)、袭击(襲撃)、死亡
Topic⑦： 消費生活関連	……
Topic⑩： 社会犯罪全般	……
Topic⑫： 中国列車事故	乘客(旅客)、列车(列車)、铁路(鉄道)、动车(動力車)、温州(温州市)、事故
Topic⑭： アメリカ軍関連	……
Topic⑰： 英国電話盗聴事件	英国(イギリス)、丑闻(不祥事)、默多克(ルパート・マードック)

広州日報（2011.7.18～2011.7.31） → 1446記事	
バーストするトピック	特徴的なキーワード（21個）
Topic①： スポーツ全般	游泳(水泳)、体育(スポーツ)、选手(選手)、金牌(金メダル)
Topic⑧： ノルウェーテロ事件	……
Topic⑩： 消費生活関連	……
Topic⑬： 社会犯罪全般	犯罪嫌疑人(犯罪容疑者)、犯罪、涉嫌(嫌疑)、法律
Topic⑲： 中国列車事故	乘客(旅客)、列车(列車)、铁路(鉄道)、死亡、温州(温州市)、事故

図 3: ニュース記事集合から選定された分析対象トピックおよび特徴的なキーワード

3 トピックモデルを利用した中国語ニュース (人民日報・広州日報) とブログの比較対照分析

3.1 分析手順の概要

本節では，図 1 の手順に沿って，中国語において，報道姿勢，報道内容，政治的立場の異なる多様なニュースサイト，および，ブログを対象として，関心事項の違いや意見の有無について分析する．具体的には，ニュースサイトとしては，中国政府の立場に最も近い人民日報[2] ，および，中国政府の立場から比較的離れた位置にある広州日報[3] をとりあげる．一方，中国語ブログホストとしては，Sina ブログホスト[4]を対象とする．

まず，2011 年 7 月 18 日から 31 日の二週間に渡って，人民日報 5,937 記事，広州日報 1,446 記事を収集した．次に，各新聞社の記事集合に対して独立に LDA を適用した．その結果，それぞれ 20 個のトピックのうち，ある程度記事集合の内容のまとまりが確認できたトピックとして，人民日報からは，

> 中国列車事故，英国電話盗聴事件，ノルウェーテロ事件，アメリカ軍関連，社会犯罪全般，消費生活関連

の 6 トピックが，一方，広州日報からは，

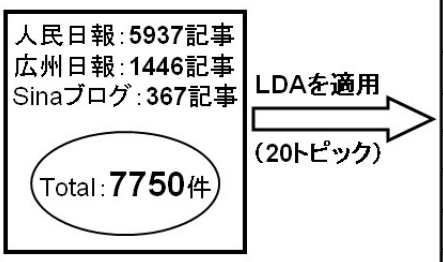
> 中国列車事故，ノルウェーテロ事件，社会犯罪全般，消費生活関連，スポーツ全般

の 5 トピックが，それぞれ得られた (図 3)．次に，これらのトピックに特徴的なキーワードを手動で選定し，人民日報からは 26 キーワードを，広州日報からは 21 キーワードを，それぞれ，ブログ記事の収集の際のクエリとした．これらのクエリを含む中国語ブログ記事の収集においては，Yahoo! Search BOSS API[5] を利用し，Sina ブログホストのドメインを対象としてブログ記事の収集を行った．その結果，収集されたブログ記事のうち，2011 年 7 月 18 日から 31 日の二週間の間に投稿されたものは 367 記事となった．この 367 ブログ記事，人民日報 5,937 記事，広州日報 1,446 記事を混合した記事集合に対して LDA を適用したところ，20 個のトピックのうち，ある程度記事集合の内容のまとまりが確認できたトピックは

> 中国列車事故，英国電話盗聴事件，ノルウェーテロ事件，アメリカ軍関連，社会犯罪全般，消費生活関連

の 6 トピックとなった (図 4)．このうち，「中国列車事故」に関する記事数は，人民日報 212 記事，広州日報 56 記事，Sina ブログ 159 記事であり，他のトピックと比較して，十分な数のブログ記事が収集できたため，次節ではこのトピックを対象として，人民日報，広州日報，Sina ブログの比較対照分析を行う．

[2] http://www.people.com.cn/
[3] http://www.dayoo.com/
[4] http://blog.sina.com.cn/
[5] http://developer.yahoo.com/search/boss/



2011.7.18~2011.7.31	人民日報	広州日報	Sinaブログ
中国列車事故	212記事	56記事	159記事
英国電話盗聴事件	348記事	33記事	30記事
ノルウェーテロ事件	405記事	44記事	7記事
アメリカ軍関連	540記事	23記事	14記事
社会犯罪全般	259記事	61記事	9記事
消費生活関連	235記事	75記事	11記事

図 4: 人民日報・広州日報・Sina ブログの混合記事集合に対して LDA を適用した結果

3.2 分析結果

当該の列車事故は 7 月 23 日に発生し，主として，24 日から 28 日の 5 日間の記事数がバースト傾向であった．そこで，記事内容の変遷と情報源の間の対比が象徴的であった 24 日から 27 日の 4 日間について，当該トピックに対応付けられた記事数の推移，および，主要な記事の内容の抜粋を図 2 に示す．この図から分かるように，人民日報の記事の内容が相対的に保守的な内容で占められるのに対して，広州日報の記事においては，当局に対する懐疑的な論調の内容が散見された．さらに，Sina ブログにおいては，事故発生直後から，当局の発表に対する疑念の声が観測され，公共のメディアでは報道されない口コミ情報も散見された．

なお，今回の分析においては，ブログ記事収集の際に用いたキーワードを手動で選定したが，今後は，このキーワード選定の過程を自動化する方式について取り組む予定である．

4 関連研究

本研究に関連して，関連ニュース記事の閲覧という観点においては，*Newsblaster*，*NewsInEssence*[6]，および，*Google News*[7] をはじめとするシステムやサービスがよく知られている．

ニュース記事に対して関連するブログ記事を対応付ける方式に関する関連研究は，大別すると，ニュース記事およびブログ記事のテキスト情報の間の関連性に基づく手法 [3, 6]，および，ブログ記事からニュース記事へのリンクによる引用情報を用いる手法 [2] に分けられる．このうち，本論文の手法は，文献 [3, 6] と同様に，ニュース記事およびブログ記事のテキスト情報の間の関連性に基づく手法に相当する．

文献 [4] においては，本研究の手法と同様に，震災に関するニュース記事・ブログ記事を収集し混合した文書集合に対してトピックモデルを適用し，ニュース・ブログの間での話題の相関，および，時系列での話題の変遷の様子を分析している．さらに，文献 [7] においては，進化的階層的ディリクレ過程を利用して，ニュース・ブログ・掲示板の間での話題の相関，および，それぞれの時系列での話題の変遷の様子を分析している．

[6] http://www.newsinessence.com/nie.cgi
[7] http://news.google.com/

5 おわりに

本論文では, 中国語において，多様なニュースサイトにおいてバーストするトピックを中心として，中国語ブログの収集を行い，中国語におけるニュースとブログの間での関心事項の違いや意見の有無について分析した．今後は，文献 [5] の成果等をふまえて，日中両国のブログにおいて共通に関心を持たれている事項について，意見の違いの分析を行う方式を確立する．

参考文献

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[2] M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. C. Konig. Blews: Using blogs to provide context for news articles. In *Proc. ICWSM*, pp. 60–67, 2008.

[3] 池田大介, 藤木稔明, 奥村学. blog とニュース記事の自動対応付け. 言語処理学会第 11 回年次大会論文集, pp. 1030–1033, 2005.

[4] 小池大地, 横本大輔, 牧田健作, 鈴木浩子, 宇津呂武仁, 河田容英, 吉岡真治, 神門典子, 福原知宏, 中川裕志, 清田陽司, 関洋平. ニュース・ブログにおける話題の相関と変遷の分析 — 震災に関する話題を例題として —. 第 4 回 DEIM フォーラム論文集, 2012.

[5] 中崎寛之, 川場真理子, 横本大輔, 宇津呂武仁, 福原知宏. 多言語 Wikipedia エントリを知識源とする特定トピックの日英ブログサイト検索と日英対照ブログ分析. 人工知能学会論文誌, Vol. 25, No. 5, pp. 613–622, 2010.

[6] 佐藤由紀, 横本大輔, 牧田健作, 宇津呂武仁, 福原知宏. ニュース記事中の話題に関連するブログ記事の収集手法. 第 3 回 DEIM フォーラム論文集, 2011.

[7] J. Zhang, Y. Song, C. Zhang, and S. Liu. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora. In *Proc. 16th SIGKDD*, pp. 1079–10881, 2010.