

政治ブログ群における潜在オピニオンのバースト性の分析

澤井 裕介

小野 景子

熊野 雅仁

木村 昌弘

龍谷大学理工学部電子情報学科

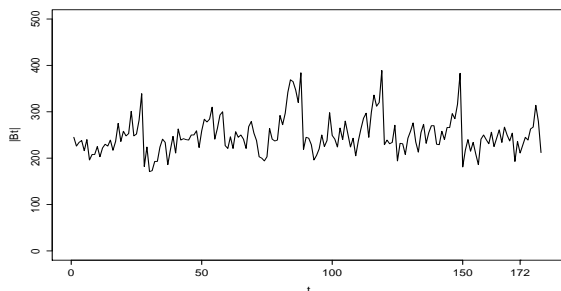
1 はじめに

近年、Web上で閲覧できる個人的な日誌の一種であるブログが発達し、その量は膨大になっている。また、2013年7月21日に行われた第23回参議院議員通常選挙では、公職選挙法の改正によりインターネット上での選挙活動が可能となり、インターネットを使用した選挙活動が活発に行われた。そのことにより、インターネット上により多くの政治的な内容の記事が投稿され、政治ブログへの関心が高まってきている。したがって、政治ブログ群の分析は重要な課題となりつつある。

本研究では、Web空間での政治問題に関する潜在オピニオンのバースト現象を分析するために、政治ブログ群における潜在トピックを抽出し、各潜在トピックに対しその盛り上がり度の変動を分析する手法を提案する。

2 分析データ

ブログランキングサイトに注目し、政治ブログの収集を行う。本研究では、ブログランキングサイトである「日本ブログ村」(<http://www.blogmura.com/>)を用い、それらの2013年2月1日から2013年7月31日までの政治カテゴリーに属するブログを収集した。そして、収集したブログエントリーを分析し、そのポストされた日付とテキスト情報を抽出した。テキスト情報はBoW表現に変換した。BoW表現における語彙の総数は6,573となった。 $B_t(t=1, \dots, T)$ を第 t 日目にポストされたブログ全体の集合とする。また、 t に対する $|B_t|$ の変動を図1に示す。収集した全ブログエントリー数は45,067となった。参議院選挙が行われた日は $t=172$ であることに注意する。

図1: t に対する B_t の分布

Burst analysis of latent opinions in political blogs

Yusuke SAWAI, Keiko ONO, Masahito KUMANO and Masahiro KIMURA

Department of Electronics and Informatics, Ryukoku University

3 分析法

本研究では、ブログ集合

$$B = \bigcup_{t=1}^{182} B_t \quad (1)$$

から潜在トピック $\{z\}$ を抽出する。そして各潜在トピック z に対してアノテーション、及びバースト分析を行う。

3.1 潜在トピックの抽出

各ブログ $b \in B$ は階層 Dirichlet 過程 (HDP) に基づく Latent Allocation (LDA) [1] モデルに従って生成されると仮定して、潜在トピック $\{z\}$ を抽出する。

3.2 潜在トピックのアノテーション

抽出した各潜在トピック z において、条件付き確率 $P(z|b)$ の値によりブログ $b \in B$ のランキングを行う。そして、その上位のブログ群に含まれる、特徴的な単語をフィッシャー検定を用いて抽出することにより z のアノテーションとする。

3.3 潜在トピックのバースト性の分析

潜在トピック z の第 t 日目の強度を

$$f_z(t) = \sum_{b \in B_t} P(z|b) \quad (2)$$

で定義する。 $f_z(t)(t=1, \dots, T)$ に関して期間 $[1, T]$ をミンシフト法 [2] を用いてクラスタリングすることにより z に関する $[1, T]$ のシーズン分解

$$[1, T] = \bigcup_{j=1}^{J_z} I_{z,j} \quad (3)$$

を構築する。 z の第 j シーズン $I_{z,j}$ のバースト度を、

$$g_{z,j} = \frac{1}{|I_{z,j}|} \sum_{t \in I_{z,j}} f_z(t) \quad (4)$$

と定義し、 z のバースト性を可視化する。

4 分析結果

抽出された潜在トピック z の総数は 237 となった。

$$\sum_{t=1}^T f_z(t) = \sum_{b \in B} P(z|b) \quad (5)$$

にしたがって潜在トピック z をランキングすることにより主要な潜在トピックを抽出した。上位5位は、 $z=54$ 、 $z=62$ 、 $z=36$ 、 $z=71$ 、 $z=108$ であった。それらの潜在トピック z に対して、表1に $P(z|b)$ ランキングに関する5位までのブログタイトルを示し、表2にアノ

テーション結果を示す。これらの結果及び、ブログエントリーを詳細に調べるにより、 $z = 54$ は参議院選挙トピック、 $z = 62$ は金融問題トピック、 $z = 36$ は地方議会トピック、 $z = 71$ は日常生活トピック、 $z = 108$ 原発問題トピックと考えられた。

表 1: $P(z|b)$ ランキング上位 5 位のブログタイトル

	ブログタイトル
54	自民の票を回して—京都・民主—共産党落としへ 共産党大躍進 8 氏が当選—非改選含め提案権獲得— 「泥舟から逃げる」民主党、崩壊第 2 幕 2013 年参議院議員選挙 制度の問題か、支持率か。—都議会議員選挙について— 資金供給量が過去最高に日銀、資 5 月末に 159 兆円超 編集委員・田村秀男 円安 = 株高持続の条件 金融量的緩和で脱デフレ効果。 消費税増税中止で設備投資に火を点ける！ 全面安の展開となった。
62	議会運営委員会 長野市長選出県議会議員と長野市との懇談会 藤原市長、3 期目がスタート 「長野県議会災害対策連絡本部設置要綱」を決定 平成 25 年みどり市議会本議会開会
71	何故が続く...中部圏への旅 歴史を感じて... 歓迎会 かめま春祭りに参加しました。 ヒロシ 酒豪の彼女
108	新基準「5 年猶予」-原子力規制委員が議論せず提示 柏崎刈羽の運転再開申請へ...7 月の新基準施行受け 原発安全審査始まる 福島第一 汚染水 破綻明かさず 福島原発から汚染水が流出していたのに、

表 2: 潜在オピニオン z に対するアノテーション

54	62	36	71	108
選挙	日銀	議会	時間	原発
候補	金融	委員	投稿	原子
議席	デフレ	質問	写真	汚染
民主	市場	市議	開催	小出
自民	株価	市議会	散歩	規制

潜在トピック $z = 54$ に関して $f_z(t)$ の変動とその密度推定の結果を図 2 に示す。また、図 3、図 4 にそれぞれ潜在トピック $z = 54$ 、 $z = 62$ のバースト可視化の結果を示す。潜在トピック $z = 62$ では、目立ったバースト性は観測されなかった。しかし、潜在トピック $z = 54$ では、参議院選挙が行われた日 ($t = 172$) の前後でバースト度が高くなっていることが観測される。

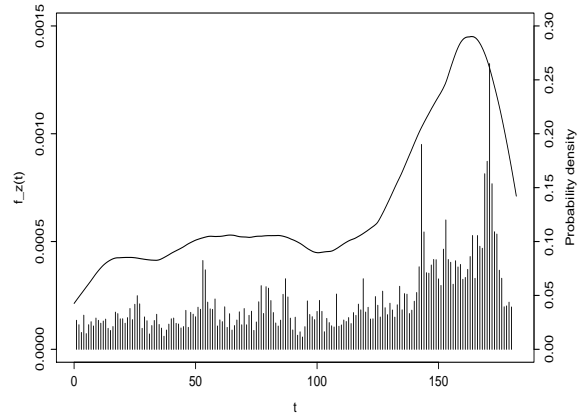


図 2: $z = 54$ の t に対する $f_z(t)$ と Probability density

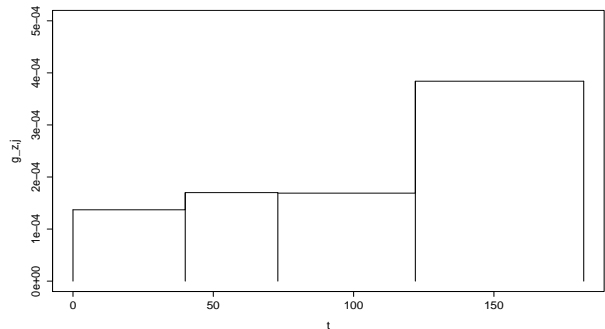


図 3: $z = 54$ のバースト性の可視化

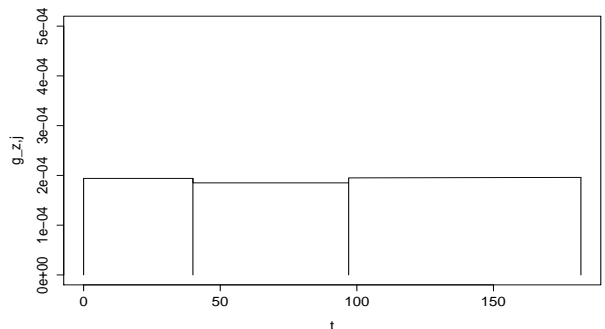


図 4: $z = 62$ のバースト性の可視化

5 まとめ

Web 空間での政治問題に関する潜在オピニオンのバースト現象を分析するために、政治ブログ群における潜在トピックを抽出し、各潜在トピックに対しその盛り上がり度の変動を分析する手法を提案した。第 23 回参議院議員選前後の政治ブログ群を収集し、本手法の有効性を検証した。

参考文献

- [1] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, Vol. 101, No. 476, pp. 1566–1581, 2006.
- [2] D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. In *ICCV*, pp. 438–445, 2001.