3E1-R-6-9

時系列ニュース・ブログにおける話題同定に関する分析 —震災を例題として—

Analysis on Identifying Topics in Times Series News and Blogs: A Case Study of Earthquake Disaster

小池 大地[*1] Daichi Koike　牧田 健作[*1] Kensaku Makita　宇津呂 武仁[*2] Takehito Utsuro　河田 容英[*3] Yasuhide Kawada　吉岡 真治[*4] Masaharu Yoshioka　福原 知宏[*5] Tomohiro Hukuhara

[*1]筑波大学大学院システム情報工学研究科 Grad. Sch. Sys. & Inf. Eng, Univ of Tsukuba

[*2]筑波大学システム情報系 Fclty. Eng, Inf. & Sys, Univ of Tsukuba

[*3](株) ナビックス Navix Co., Ltd.

[*4]北海道大学大学院情報科学研究科 Grad. Sch of Inf. Sci. & Tech, Hokkaido Univ.

[*5]独立行政法人 産業技術総合研究所 サービス工学研究センター Center for Service Research, National Institute of Advanced Industrial Science and Technology

This paper presents results of analyzing correlation of topics in time series newspaper articles and blogs after the recent earthquake disaster until the end of the year of 2011, as well as changes in sub-topics as time passed after the disaster. In this analysis, we first collect blog posts that are closely related to the earthquake disaster and then apply a topic model to the union of the set of newspaper articles and that of blog posts. We observed that most topics are closely related to both news articles and blog posts, while some topics are only specific to news articles or to blog posts. We show that the proposed analysis method is quite effective in efficiently analyzing correlation of news and blogs as well as changes in sub-topics as time passed.

1. はじめに

現代の情報社会においては，情報の氾濫の問題が顕著であり，このことは，いわゆる情報爆発の問題を引き起こしている．そして，そのように爆発的に増大する情報の集約や，俯瞰をするための技術の開発が強く望まれている．中でも，情報爆発が最も顕著に現れているのはウェブである．ウェブ上には，様々なメディア上で情報が氾濫しているが，その中でも，ニュースやブログなどは，実世界において注目すべき出来事が起るとその事実をニュースが報道し，一方，その出来事に対して，一般個人のレベルでの反応や感想，意見がブログに書かれる，というサイクルで情報が行き交うことになる．

このように，ウェブ上で情報が氾濫する状況をふまえて，我々は，ウェブ上の情報の中でも，特に，ブログ空間における多種多様な記事内容を俯瞰的に閲覧する方式の研究を行ってきた [横本 11]．具体的には，Wikipedia を知識源として話題ラベル[*1] の体系を構築し，この Wikipedia の体系を元に，ブロガーのブログ記事集合に対して話題を対応付ける方式を提案した．また，そのようなウェブ上のニュースとブログを関連付けることにより情報の集約を行う，という方向の研究も行われている．それらの研究の基盤となる技術は，ニュース記事とブログ記事の間で話題の対応をとる技術であるが，それらの技術は，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 [池田 05, 佐藤 11]，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 [Gamon08] に分けられる．

以上の研究の成果をふまえて，本論文では，特に，一定期間におけるニュース・ブログの話題の相関と変遷の分析を行った結果を示す．特に，題材として，2011 年 3~12 月の期間において，「東日本大震災」に関連する話題のニュース記事，および，ブログ記事を収集し，ニュース・ブログの間の話題の相関と変遷の分析を行った結果を報告する．

2. 分析手順の概要

本論文で用いた手法の外観図を図 1 に示す．この手法においては，まず，2011 年 3~12 月の期間のニュース記事，および，ブログ記事を収集したものを一つの文書集合とみなして，トピックモデル (本論文においては，潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [Blei03] を用いた[*2]) を適用し，ニュース記事，および，ブログ記事を混合した文書集合におけるトピックを推定する．次に，各ニュース記事 d，あるいは，ブログ記事 d に対して，確率値 $P(z_n|d)$ が最大となるトピック z_n を割り当てる．これにより，各トピックに，どの程度の数のニュース記事，あるいは，ブログ記事が対応しているのかの分析を行う．また，各トピックにおいて，中心的な内容が時系列にどのように変遷するかの分析を行う．さらに，ニュース特有の内容，および，ブログ特有の内容について分析を行う．

これらの分析においては，まず，各トピック z_n において，確率値 $P(w \mid z_n)$ が上位の語 (実際には，Wikipedia エントリタイトルを利用) を参照して，全期間に渡ってトピック z_n に特有の特徴を表すとする．その一方で，クエリ尤度モデルの枠組み [Ponte98] に基づき，Wikipedia エントリタイトルを話題ラベルとみなして，個々のニュース記事，および，ブログ記事に付与し [小池 12]，各文書の特徴付けを行ったうえで分析

連絡先: 小池 大地，筑波大学大学院システム情報工学研究科，〒 305-8573　茨城県つくば市天王台 1-1-1, 029-853-5427

*1 本論文では，トピックモデルにおける用語 “トピック” に対して，2. 節の手法によって各トピックに割り当てられた記事中の詳細な内容を表す用語として “話題” を用いる．

*2 ツールとして，GibbsLDA++(http://gibbslda.sourceforge.net/) を用いた．本論文では，トピック数を 50，および，100 としてトピック推定を行い，得られたトピックを人手で見比べ，トピックの推定結果の性能がより高くなったトピック数 50 を採用した．

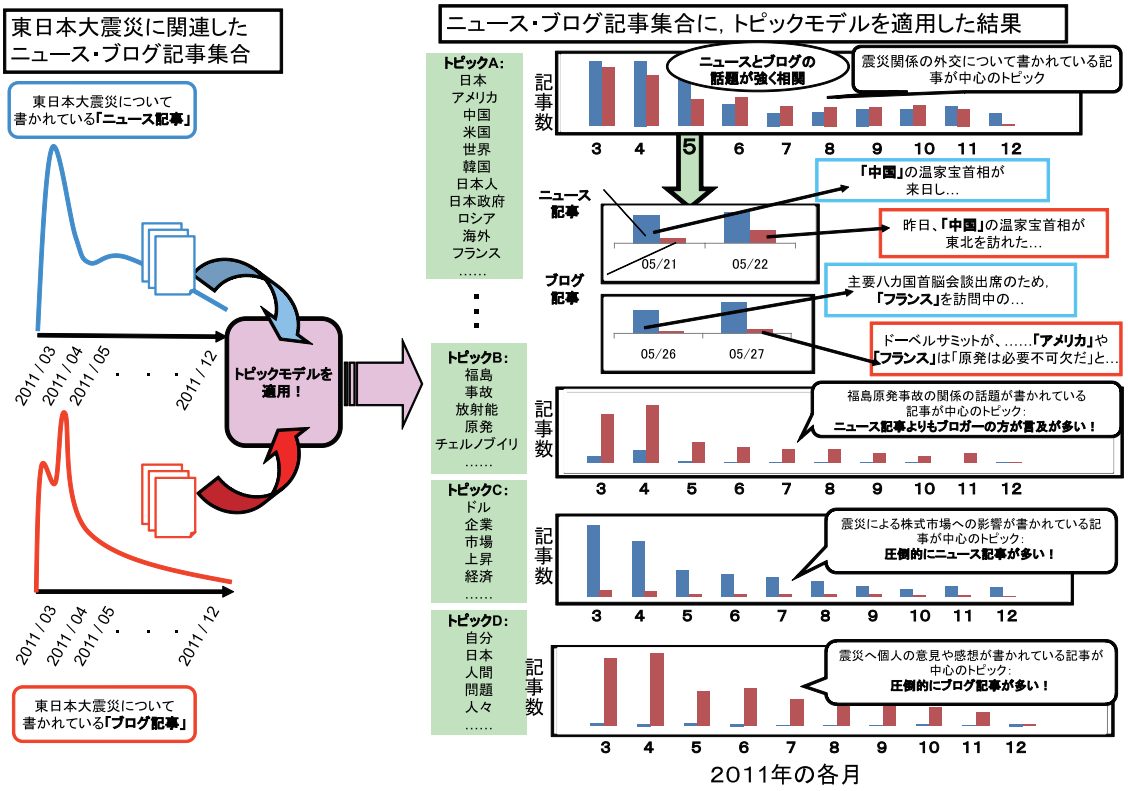


図 1: ニュース・ブログにおける話題の相関・変遷の分析の流れ

を行っている．

3. 分析

3.1 分析対象

ニュース記事としては，2011 年 3 月 11 日から 12 月 29 日までの日付のものを，日経新聞[*3]，朝日新聞[*4]，読売新聞[*5]の各新聞社のサイトから収集した 70,005 記事，23,237 記事，および，50,286 記事の合計 143,528 記事を用いた．その後，震災関係の 7 語[*6] およびそのリダイレクトを Wikipedia から収集し，それらのうちの少なくとも一つがニュース記事中に出現するものだけを分析対象とした．その結果，各新聞社の記事数は，日経新聞が 11,006 記事，朝日新聞が 4,988 記事，読売新聞が 8,368 記事，合計 24,458 記事となった．

一方，ブログ記事としては，震災関係の 7 語 (上述) の一つ一つを初期クエリ t_0 として，関連するブログ記事集合を収集した結果を用いた．初期クエリ t_0 を含む日本語ブログの収集においては，Yahoo! Search BOSS API[*7] を利用し，日本語ブログホスト大手 6 社[*8]のドメインを対象として，2011 年 11 月下旬から 12 月下旬に，2011 年 3 月 11 日以降の日付の記事を対象として，ブログ記事の収集を行った [小池 12]．その後，上述の震災関係の 7 語およびそのリダイレクトを Wikipedia から収集し，それらのうちの少なくとも一つがブログ記事中に出現するものだけを分析対象とした．その結果，分析対象のブログ記事は，34,826 記事となった．

3.2 分析結果

図 2 に，50 個のトピックのうちの主要なものについて，$P(w|z_n)$ が上位の語，および，ニュース記事，および，ブログ記事の典型例をそれぞれ示す．ニュース記事，および，ブログ記事中の赤字の語は，「$P(w|z_n)$ が上位の語」の欄に示した語である．これらの 7 個のトピックは，いずれも，震災関係において，典型的に観測されるトピックであり，これらのトピックにおいては，ニュース記事における報道内容とブログにおける関心事項が高い相関を示す場合が多い．

次に，ニュース・ブログ間の相関および時系列の遷移に関する分析を行った [小池 12]．その典型例を，図 1 の右半分，および図 3 に示す．

ニュース・ブログ間の相関に関して，図 1 のトピック A については，全体の期間を通して記事件数の間に強い相関が示されていることが分かる．一方，トピック B，および，トピック D については，全体の期間を通して，圧倒的にブログ記事数の方がニュース記事数よりも多いという傾向がある．一方，トピック C については，全体の期間を通して，圧倒的にニュース記事数の方がブログ記事数よりも多いという傾向がある．

図 1 のトピック B に関して，その詳細を図 3 に示す．実際にその記事内容を見てみると，「福島原子事故の放射能汚染」に関する話題が中心となるトピックであるが，ブログにおいては，一貫して，チェルノブイリ事故と比較しての福島原発事故の放射能汚染の影響について書かれたブログ記事が多数観測された．その中で，深刻度に関する国際評価が「レベル 7」に引き上げられたという 4 月 12 日の報道の直後のみ，関連ニュー

*3 http://www.nikkei.com/
*4 http://www.asahi.com/
*5 http://www.yomiuri.co.jp/
*6 福島県，放射能，津波，東京電力，原子力発電所，放射線，原子力発電．
*7 http://developer.yahoo.com/search/boss/
*8 fc2.com，yahoo.co.jp，ameblo.jp，goo.ne.jp，livedoor.jp，hatena.ne.jp

トピック ID	各トピックにおいて $P(w\|z_n)$が上位の語	各トピックのニュース記事		各トピックの各ブログ記事	
		数	例	数	例
0 ⋮	セシウム、放射性物質、ヨウ素、ベクレル、土壌、濃度、基準、結果、放射能、半減期、水道水、・・・	396	金町浄水場の水道水から、基準を上回る放射性ヨウ素が検出された。(2011/03/23)	1201	土壌の放射性物質はどれくらい稲作に影響があるのだろう。(2011/08/16)
2 ⋮	処理、放射性物質、汚染、施設、焼却、廃棄物、汚泥、がれき、処分、環境、地下、流出、装置、環境省、・・・	803	環境省は31日、ごみ焼却施設で発生する放射性物質を含む汚染焼却灰の処分方法を発表した。(2011/08/31)	636	今日、環境局一般廃棄物対策課で、「東京都が受け入れる震災がれき」の話があり参加してきた。(2011/11/15)
8	宮城、津波、岩手、被害、東日本、石巻氏、大震災、遺体、死亡、不明、避難所、行方不明者、・・・	1347	2日目を迎えた12日岩手、宮城、福島の東北3県の沿岸部を中心に被害実態が明らかになってきた。(2011/03/12)	734	東北地方地震。昨日から凄い被害状況ですが亡くなられた方や行方不明者も多いんですね。(2011/03/12)
9 ⋮	津波、メートル、被害、高さ、大津波、対策、防災、地図、大震災、海岸、逃げ、防潮堤、堤防、・・・	440	仙台新港など太平洋岸の各地に高さ10メートルクラスの津波が到来。(2011/03/11)	1267	津波による被害は想像を絶するものです。防災対策や避難対策が機能したことを祈るばかりです。(2011/03/12)
14 ⋮	原子炉、爆発、冷却、容器、東電、燃料、可能、水素、プール、電源、漏れ、海水、炉心・・・	1497	建屋内に原子炉から漏れた水素がたまり爆発するおそれがあると発表した(2011/03/13)	1356	緊急炉心冷却装置が作動しなかった福島原発の原子炉が気になります。(2011/03/12)
36 ⋮	計画、午後、停電、時間、午前、グループ、東京電力、発表、対象、地域、一部、時間帯、・・・	576	東京電力は17日、計画停電の規模が、14日に初実施して以降、最大規模になる見込みだと発表した。(2011/03/17)	161	不足と言われた関東の電気を一部地域の2時間の停電だけですませた東京電力はすごい。(2011/03/13)
39 ⋮	自衛隊、派遣、活動、放水、米軍、東日本、ロボット、被災地、ヘリ、救助、輸送、車両、対応、・・・	896	米軍と自衛隊による被災地への支援活動が16日に本格化した。(2011/03/16)	436	ヘリコプターによる放水作業という危険な活動を続ける自衛隊に感謝。(2011/03/20)

図 2: トピックの抜粋およびニュース記事・ブログ記事の典型例

ス記事がやや増加した.

3.3 トピック・日付ごとの記事内容のまとまりの評価

各トピックの各日付について合計 389 箇所の評価を行った. 評価をするに当たり，ニュース，ブログ記事の集まっている記事件数によって日付を，バーストしている期間，バースト期間以外で一定数以上記事が集まっている期間，その他の無作為に選んだ期間に分けて評価を行った. その結果を図 4 に示す.

図 4 (a) においては，合計 389 箇所中，ニュース，ブログ記事がそれぞれ 3 記事以上集まっている期間について，確率値 $P(z_n|d)$ の上位最大 5 記事の内，半数以上が確率値 $P(z_n|d)$ 一位の記事と同一の話題であると判断された割合を示す. この結果から，一定数以上記事が集まっている期間の方が，同一の話題について書かれた記事が集まっている割合が高いことが分かる.

図 4 (b) においては，ニュース，ブログ記事について，どちらとも半数以上が同一の話題と判断された期間中，そのニュース，ブログ記事の話題を比較した時に、ニュースとブログが同一の話題と判断された期間の割合を示す. この結果からも，一定数以上記事が集まっている期間の方が，ニュース，ブログ記事で同一の話題について書かれた記事が集まっている割合が高いことが分かる.

4. おわりに

本論文では，題材として，2011 年 3~12 月の期間において，「東日本大震災」に関連する話題のニュース記事，および，ブログ記事を収集し，ニュース・ブログの間の話題の相関と変遷の分析を行った. なお，本研究に関連して，ニュース，ブログといった複数の相互に関連しあっている時系列の情報源を対象としてトピックモデルを適用し，各トピックの時系列の特徴をとらえる方式 [Zhang10] がある. また，[風間 11] は，東日本大震災における Twitter のトピックを分析するために，名詞の共起を調査するとともに，名詞群の出現頻度の時間的変化とトピックとの関係を分析している.

参考文献

[Blei03] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).

[Gamon08] Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M. and Konig, A. C.: BLEWS: Using Blogs to Provide Context for News Articles, *Proc. ICWSM*, pp. 60–67 (2008).

[池田 05] 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).

[風間 11] 風間一洋, 鳥海不二夫, 篠田孝祐, 榊剛史, , 栗原聡, 野田五十樹：名詞出現頻度の時間的変化に着目した東日本大震災時の Twitter のトピックの分析, WebDB Forum 2011 論文集 (2011).

[小池 12] 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).

[Ponte98] Ponte, J. M. and Croft, W. B.: A Language Modeling Approach to Information Retrieval, *Proc. 21st SIGIR*, pp. 275–281 (1998).

[佐藤 11] 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).

[横本 11] 横本大輔, 林東權, 牧田健作, 宇津呂武仁, 河田容英, 福原知宏, 神門典子, 吉岡真治, 中川裕志, 清田陽司：特定トピックに関するブログ記事集合の観点分類における Wikipedia の利用, 第 3 回 DEIM フォーラム論文集 (2011).

[Zhang10] Zhang, J., Song, Y., Zhang, C. and Liu, S.: Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora, *Proc. 16th SIGKDD*, pp. 1079–10881 (2010).

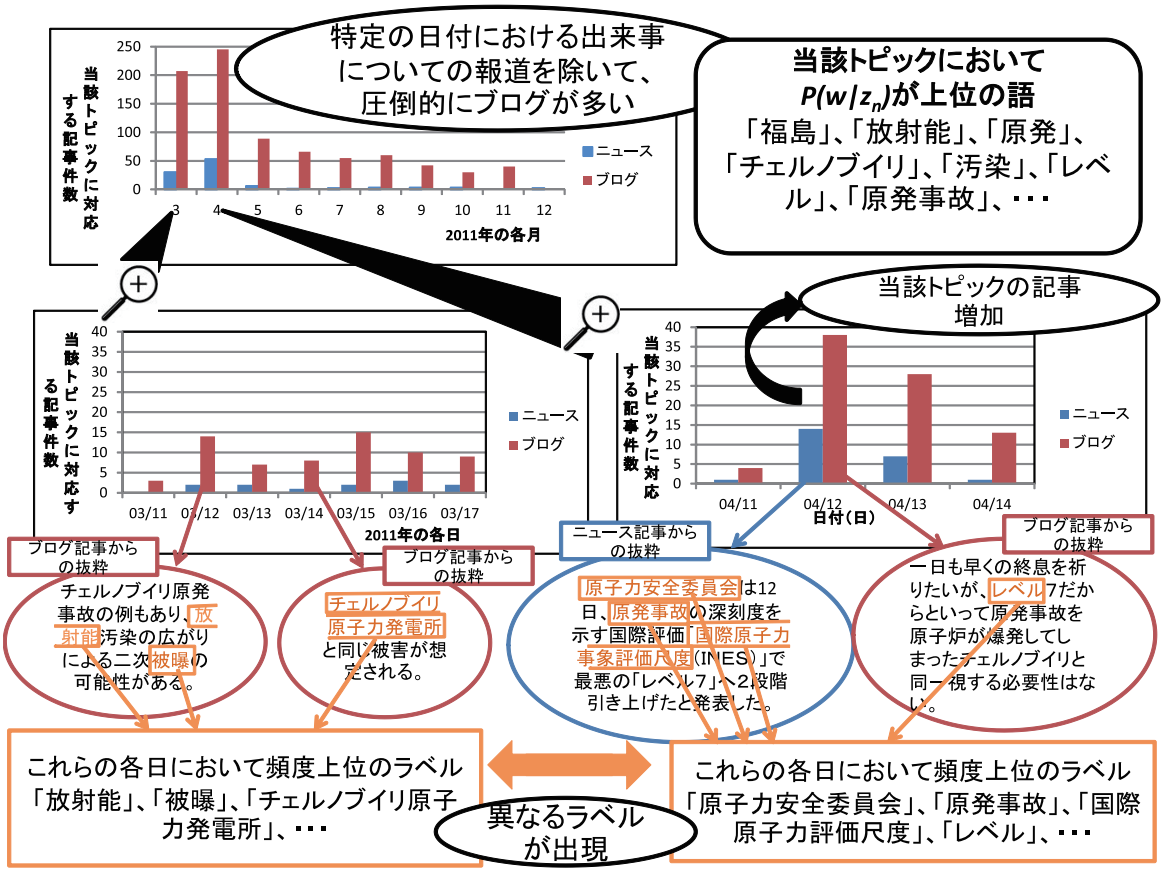
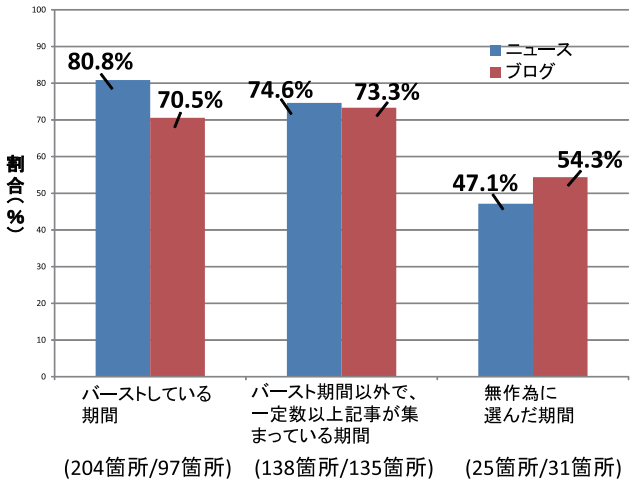
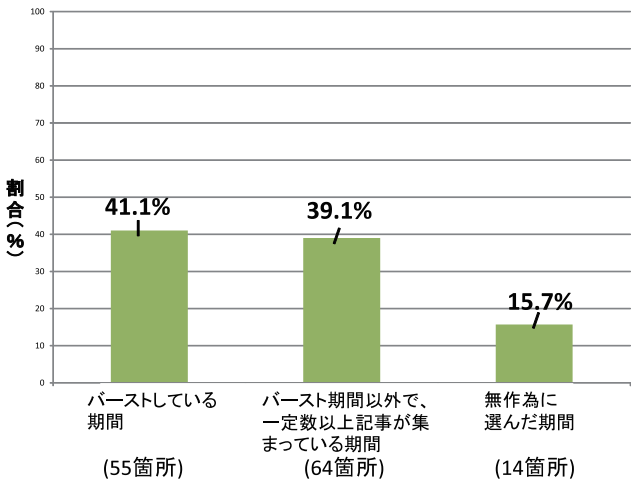


図 3: ニュースにおける報道よりも，ブログにおける関心の方が高い例：「福島原発事故の放射能汚染」関係 (ニュース: 103 記事, ブログ: 835 記事)



(a) 確率値 $P(z_n|d)$ の上位最大 5 記事の内，半数以上が確率値 $P(z_n|d)$ 一位の記事と，同一の話題である割合



(b) (a) の内，ニュースとブログが同一の話題について書かれている割合

図 4: トピック・日付ごとの記事内容のまとまりの評価