DEIM Forum 2012 E2-4

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 —

小池　大地[†1]　横本　大輔[†2]　牧田　健作[†2]　鈴木　浩子[†2]　宇津呂武仁[†3]
河田　容英[†4]　吉岡　真治[†5]　神門　典子[†6]　福原　知宏[†7]　中川　裕志[†]
清田　陽司[†]　関　洋平[††]

†1 筑波大学理工学群工学システム学類　〒 305-8573 茨城県つくば市天王台 1-1-1
†2 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
†3 筑波大学 システム情報系 知能機能工学域　〒 305-8573 茨城県つくば市天王台 1-1-1
†4 (株) ナビックス　〒 141-0031 東京都品川区西五反田 8-3-6
†5 北海道大学大学院 情報科学研究科　〒 060-0808 北海道札幌市北区北 8 条西 5 丁目
†6 国立情報学研究所　〒 101-8430 東京都千代田区一ツ橋 2-1-2
†7 産業技術総合研究所　〒 135-0064. 東京都江東区青梅 2-3-26
† 東京大学 情報基盤センター　〒 113-0033 東京都文京区本郷 7-3-1
†† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2

あらまし　本論文では，震災に関する話題についての時系列のニュース記事集合，および，ブログ記事集合を対象として，トピックモデルを用いたトピック同定を行う．そして，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する．分析の結果，ニュース・ブログ間の相関が高いトピック，ニュース記事特有のトピック，ブログ記事特有のトピックなどの違いを容易に発見することができた．

キーワード　ニュース，ブログ，話題，時系列分析

Analyzing Correlation of Topics in News and Blogs and their Changes: A Case Study of Topics on Earthquake Disaster

Daichi KOIKE[†1], Daisuke YOKOMOTO[†2], Kensaku MAKITA[†2], Hiroko SUZUKI[†2], Takehito UTSURO[†3], Yasuhide KAWADA[†4], Masaharu YOSHIOKA[†5], Noriko KANDO[†6], Tomohiro FUKUHARA[†7], Hiroshi NAKAGAWA[†], Yoji KIYOTA[†], and Yohei SEKI[††]

†1 College of Eng. Sys., School of Science and Engineering, University of Tsukuba,　Tsukuba 305-8573 Japan
†2 Grad. Sch. of Systems and Information Engineering, University of Tsukuba,　Tsukuba 305-8573 Japan
†3 Faculty of Engineering, Information and Systems, University of Tsukuba,　Tsukuba 305-8573 Japan
†4 Navix Co., Ltd.　Tokyo 141-0031, Japan
†5 Graduate School of Information Science and Technology, Hokkaido University, Sapporo, 060-0808, Japan
†6 National Institute of Informatics, Tokyo 101-8430, Japan
†7 National Institute of Advanced Industrial Science and Technology,　Tokyo 135-0064 Japan
† Information Technology Center, University of Tokyo, Tokyo 113-0033, Japan
†† Faculty of Library, Information and Media Science, University of Tsukuba,　Tsukuba 305-8550 Japan

Key words　news, blog, topic, time series analysis

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話題を俯瞰的に閲覧する方式の研究を行ってきた [8], [9], [12]．具体的には，基本的な方式 [12] として，Wikipedia を知識源として話題の体系を構築し，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，そのほか，複数の言語間で話題の分布を比較分析する方式 [8]，あるいは，時系列方向の話題の分布を分析する方式 [9] 等を提案した．

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 [3], [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 [2], [4], [6] 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．

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，LDA (Latent Dirichlet Allocation) [1] を用いた) を適用し，トピックを推定する (2.1 節)．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 (2.2 節)．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，中心的な話題が時系列にどのように変遷するのかについての分析を行う．さらに，ニュース特有の話題，および，ブログ特有の話題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 [10] 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 [13](3. 節)，各文書の特徴付けを行ったうえで分析を行っている．

分析の結果，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関連を示す場合が多いトピック，ニュース記事特有のトピック，ブログ記事特有のトピックなどの違いを容易に発見することができた．また，各日に特徴的な話題ラベルを同定することにより，同一のトピックにおいても，時系列に沿って話題がめまぐるしく変遷する様子を容易に観測することができた．

また，以上の方式と同一の枠組みにより，ニュース (新聞記事)，ブログに加えて，2011 年 3 月 11 日から 12 月 31 日の期間の NHK 放送字幕テキスト(注1) を混合した文書集合を対象として，ニュース・ブログとテレビ放送との間の話題の相関と変遷の分析を行った結果についても報告する．

2. トピックモデルを用いた話題分布の分析

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) [1] を用いる．LDA を用いたトピックモデルの推定においては，語 w の列によって表現された文書の集合と，トピック数 K を入力として，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．これらを推定するためのツールとしては，GibbsLDA++(注2) を用いた．LDA のハイパーパラメータである α，β には，GibbsLDA++の基本設定値である $\alpha = 50/K$，$\beta = 0.1$ を用いた．LDA ではトピック数 K を人手で与える必要があるが，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．なお，このツールは推定の際に Gibbs サンプリングを用いているが，その反復回数は 2,000 とした．

2.2 文書に対するトピックの割り当て

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．文書集合を D，トピック数を K，1 つの文書を d $(d \in D)$ とすると，トピック z_n $(n = 1, \ldots, K)$ 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．

$$D(z_n) = \left\{ d \in D \mid z_n = \underset{z_u \ (u=1,\ldots,K)}{\operatorname{argmax}} P(z_u|d) \right\}$$

(注1)：本字幕テキストデータは，国立情報学研究所平成 23 年度共同研究『NII 研究用テレビジョン放送アーカイブを用いた東日本大震災の社会的影響の学術的分析』(戦略研究公募型) No.74 「テレビジョン放送アーカイブと新聞・ブログ・マイクロブログの特性を考慮した東日本大震災の社会的影響の学術的分析」の一環として利用しているものである．

(注2)：http://gibbslda.sourceforge.net/

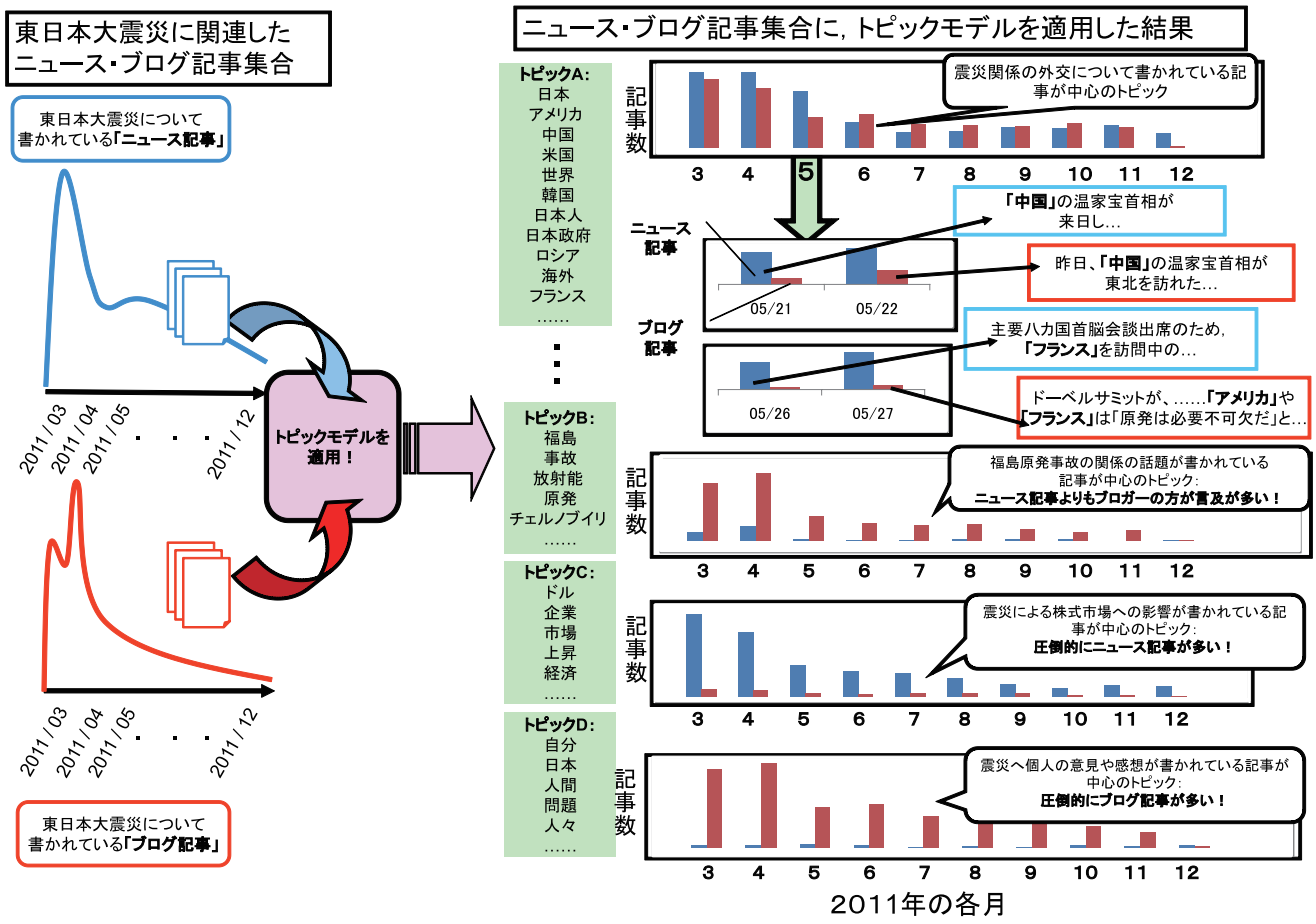

図 1　ニュース・ブログにおける話題の相関・変遷の分析の流れ

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．

3. Wikipedia を知識源とする文書への話題ラベルの付与

3.1 クエリ尤度モデル

本節では，情報検索の手法 1 つであり，本研究で用いているクエリ尤度モデル [10] について説明する．

クエリと文書集合が与えられ，クエリに適合する度合いに従って文書をランキングしたい場合，文書 d ($\in$ 文書コレクション C) がクエリ q に適合する確率 $P(d \mid q)$ を求めることができれば，これに従ってランキングを行うことができる．このとき，ベイズの定理を用いることにより，次式を得ることができる．

$$P(d \mid q) = \frac{P(q \mid d)P(d)}{P(q)} \tag{1}$$

$P(q)$ は文書 d に依存しないので定数とみなせる．また 文書 d に関しての何らかの事前知識がない限り，$P(d)$ は一様であるとみなす．このとき，式 (1) を次のように簡略化できる．

$$P(d \mid q) = \frac{P(q \mid d)P(d)}{P(q)} \propto P(q \mid d)P(d) \propto P(q \mid d)$$

上式に基づく情報検索手法をクエリ尤度モデルと呼ぶ．

次に，文書中の単語は独立に生起すると仮定して，文書をユニグラム言語モデルとして表現する．このとき，文書 d におけるユニグラム言語モデル θ_d からクエリ q が生成される尤度，すなわちクエリ尤度 $P(q|\theta_d)$ は，語彙 $w \in V = \{w_1, \ldots, w_{|V|}\}$ のクエリ q における出現頻度 $\mathrm{c}(w_i, q)$ を用いて次式のように定義される．

$$P(q \mid \theta_d) = \prod_{w_i \in V} P(w \mid \theta_d)^{\mathrm{c}(w_i, q)} \tag{2}$$

上式を求めるためには，文書モデルのパラメータ θ_d を推定しなければならない。そのための推定方法の 1 つとしては，文書 d における最尤推定値 $P_{ML}(w_i \mid \theta_d)$ と，文書コレクション C における最尤推定値 $P_{ML}(w_i \mid \theta_C)$ を線形補間するという手法をもちいる．まず，文書 d における最尤推定値 $P_{ML}(w_i \mid \theta_d)$ は，相対頻度を用いて以下のように定義される．

$$P_{ML}(w_i \mid \theta_d) = \frac{\mathrm{c}(w_i, d)}{|d|} \tag{3}$$

ここで，$\mathrm{c}(w_i, d)$ は文書 d において語 w_i が出現する頻度，$|d|$ は文書 d の文書長すなわち総語数を示す．一方，文書コレクション C における最尤推定値 $P_{ML}(w_i \mid \theta_C)$ は次のようにして推定できる．

$$P_{ML}(w \mid \theta_C) = \frac{\sum_{d \in C} \mathrm{c}(w_i, d)}{\sum_{d \in C} |d|} \tag{4}$$

ここで，文書コレクション C は，検索対象の文書集合である．

式 (3)，および，式 (4) と，補間の度合いを表現する定数

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

$\lambda \in [0,1]$ を用いて，線形補間法による推定値は次のように定義される．

$$P(w_i \mid \theta_d) \;=\; \lambda P_{ML}(w_i \mid \theta_d) \;+\; (1-\lambda) P_{ML}(w_i \mid \theta_C) \tag{5}$$

上式を用いることで，式 (2) のクエリ尤度 $P(q|\theta_d)$ を求めることができる．

3.2　文書への話題ラベルの付与

本節では，前節で述べたクエリ尤度モデルの考え方を用いて，対象文書集合の個々の文書に対して話題ラベルを付与する手法について説明する．なお本節以降では，文書を表す記号として A を用いる．

本研究では，文書中に出現する Wikipedia エントリタイトルから，文書の話題ラベルとして相応しいものを自動選定する．そのために，文書をクエリとみなして，文書中にエントリタイトルが出現した Wikipedia エントリ集合のランキングを行う．はじめに，対象文書集合 $\mathbb{A}$ の個々の文書 $A \in \mathbb{A}$ は，文書中に出現した Wikipedia エントリタイトルの集合として表現される．

$$A = \left\{ t(E_1), \ldots, t(E_n) \right\}$$

まず，対象文書集合 $\mathbb{A}$ において，エントリタイトル $t(E)$ が 10 個以上の文書に出現した Wikipedia エントリを集めて，Wikipedia エントリ集合 $\mathbb{E}(\mathbb{A})$ を作成する．

$$\mathbb{E}(\mathbb{A}) \;=\; \left\{ E \;\middle|\; \mathrm{df}(\,\mathbb{A},\; t(E)\,) \;\geq 10 \right\}$$

次に，文書 A にエントリタイトル $t(E)$ が出現し，かつ，Wikipedia エントリ集合 $\mathbb{E}(\mathbb{A})$ に含まれる Wikipedia エントリ E を抽出し，文書 A に対する話題ラベルの候補集合に対応する Wikipedia エントリ集合 $\mathbb{E}(A)$ を作成する．

$$\mathbb{E}(A) \;=\; \left\{ E \in \mathbb{E}(\mathbb{A}) \;\middle|\; t(E) \in A \right\}$$

そして，クエリ尤度モデルに基づいて，文書 A をクエリとみなして，Wikipedia エントリ集合 $\mathbb{E}(A)$ のランキングを行う．具体的には，$q = A$，$d = E \in \mathbb{E}(A)$，$C = \mathbb{E}(\mathbb{A})$ として，式 (2) の $P(A \mid \theta_E)$ を求める．

以上のように推定した $P(A \mid \theta_E)$ を用いることで，文書 A に付与する話題ラベル集合 $L(A)$ を以下のように決定する．

$t(E) \in L(A)$ の選定手順

- $E \in \mathbb{E}(A)$
- $P(A \mid \theta_E) \geq \left(\alpha \times \max_{E' \in \mathbb{E}(A)} P(A \mid \theta_{E'}) \right)$
- $P(A \mid E)$ の大きいものから順に 10 個まで選ぶ

具体的には，Wikipedia エントリ集合 $\mathbb{E}(A)$ における $P(A \mid \theta_E)$ の最大値に対して，その α 倍以上の $P(A \mid \theta_E)$ を持つ Wikipedia エントリのタイトル $t(E)$ を，文書 A の話題ラベルとして抽出する．なお，本論文では $\alpha = 0.6$ とした．また，1 文書に付与する話題ラベルの数は最大 10 個とし，$P(A \mid \theta_E)$ の大きいものから順に上位 10 個までのエントリタイトルを $L(A)$ とした．

4.　ニュース・ブログ間の話題に関する分析

4.1　分析対象

4.1.1　ニュース記事

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(注3)，朝日新聞(注4)，読売新聞(注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

福島県，放射能，津波，東京電力，原子力発電所，放

(注3)：http://www.nikkei.com/
(注4)：http://www.asahi.com/
(注5)：http://www.yomiuri.co.jp/

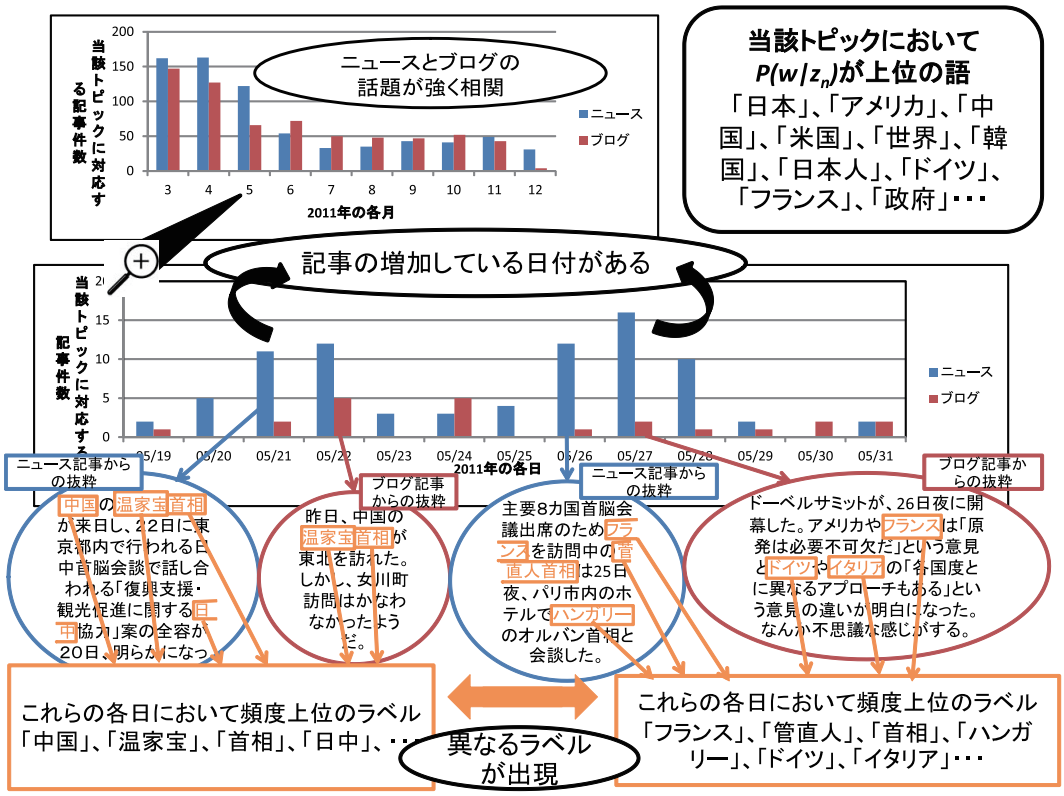

図 3　ニュース・ブログ間の時系列の相関と各日における記事の参照関係の例：「海外とのやり取り」 (ニュース: 733 記事，ブログ: 656 記事)

射線，原子力発電

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．

4.1.2　ブログ記事

前節で述べた，震災関係の 7 語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(注6) を利用し，日本語ブログホスト大手 6 社(注7)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．検索の際には，複数のドメインを一度に指定して検索し，1,000 件の記事を取得する．次に，ブログ記事検索後，検索結果の URL をブログサイト単位にまとめる．その結果，一つの検索クエリあたり約 200 前後のブログサイトが取得される．次に，各ブログサイトをドメイン指定し，初期クエリ t_0 を検索クエリとすることにより，各ブログサイト中において初期クエリ t_0 を含むブログ記事を収集し，ブログ記事集合を作成する．その後，上述の震災関係の七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．

4.2　分 析 結 果

前節で述べたニュース記事およびブログ記事，合計 59,284 記事を混合した文書集合を対象として，LDA を適用した(注8)．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関連を示す場合が多い．一方，図 3〜 図 6 には，特に，以下の特徴を強く持つトピックの例を示す．

- 3. 節の手法により，各ニュース記事，および，ブログ記事に対して，Wikipedia を知識源として，文書中の話題の特徴を的確に表す話題ラベルを付与し，この話題ラベルを日毎に集計をして頻度上位のラベルを ⟨ トピック，日 ⟩ の組に対して付与している．この話題ラベルと，各トピックに対して付与されている「$P(w|z_n)$ が上位の語」を比較し，日毎に特徴的な話題

(注6)：http://developer.yahoo.com/search/boss/

(注7)：fc2.com，yahoo.co.jp，ameblo.jp，goo.ne.jp，livedoor.jp，hatena.ne.jp

(注8)：ニュース記事集合，あるいは，ブログ記事集合に対して個別に LDA を適用してはいないため，それぞれの記事集合に対してトピック数を最適化するという手順はとっていない．本論文の手法を用いることにより，本節で述べるように，ニュース記事が中心となるトピック，あるいは逆に，ブログ記事が中心となるトピック，等も自然に観測することが可能となる．なお，本論文で分析対象としたトピックにおいて，文書のまとまりがどの程度とらえられているか，あるいは，同一トピックに対応するニュース記事とブログ記事の間で，どの程度厳密に話題の対応がとれているかについての評価結果の詳細は，文献 [7] において示す．

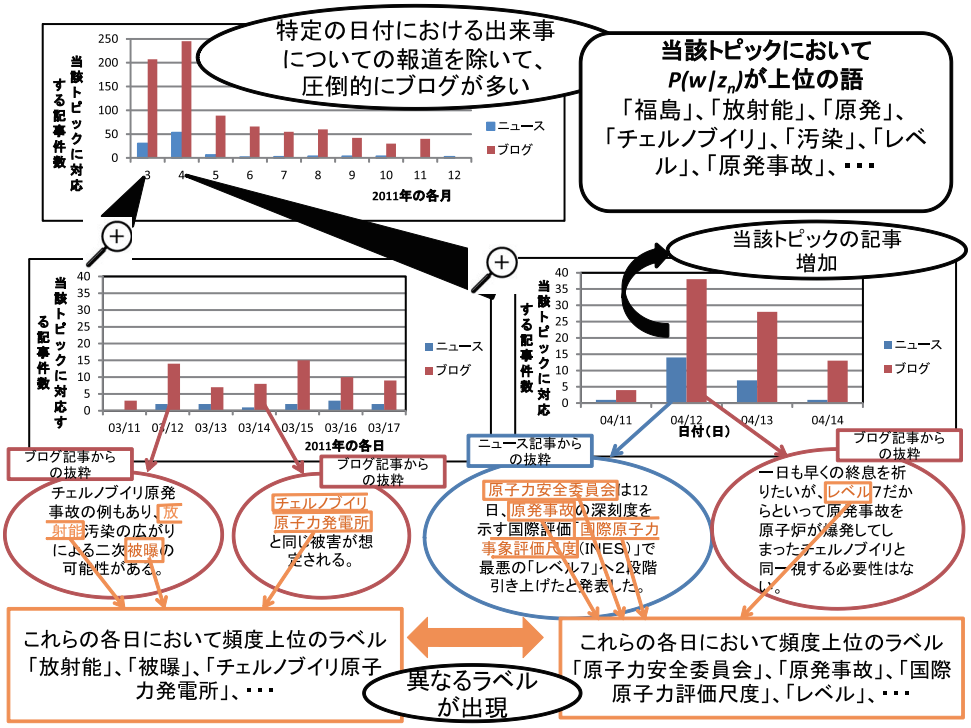

図 4　ニュースにおける報道よりも，ブログにおける関心の方が高い例：「福島原発事故の放射能汚染」関係 (ニュース: 103 記事，ブログ: 835 記事)

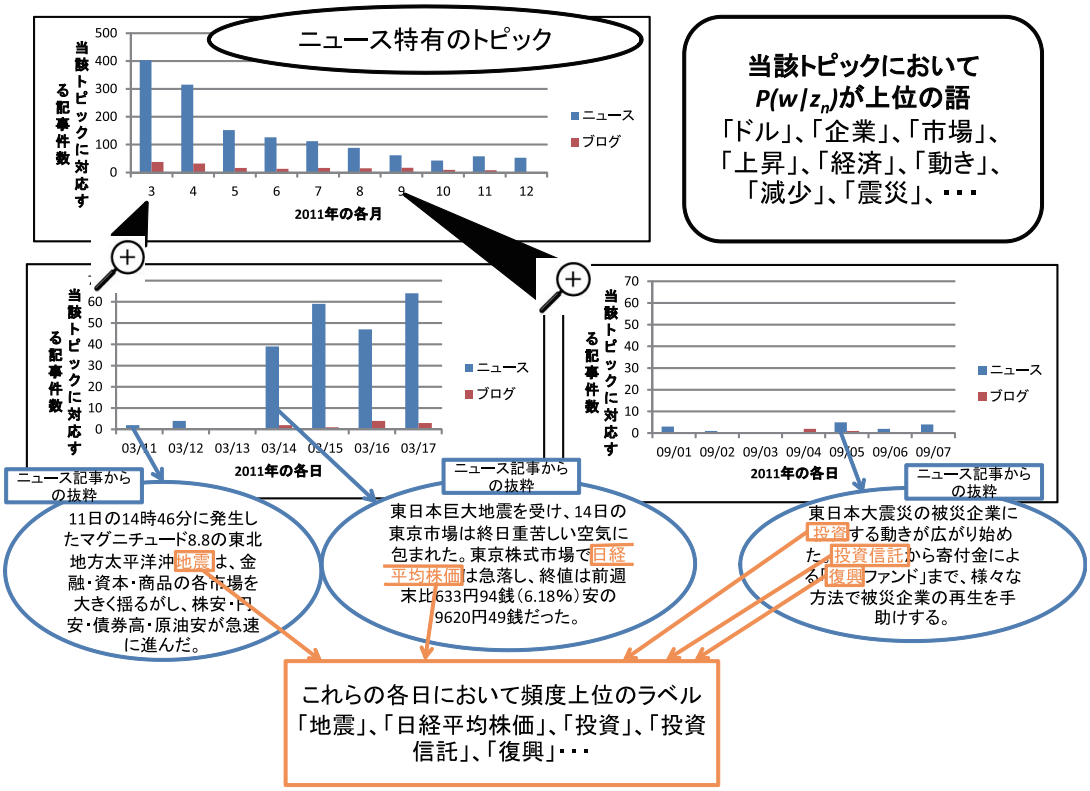

図 5　ニュース特有のトピックの例：「株式市場への影響」 (ニュース: 1,412 記事，ブログ: 165 記事)

ラベルが大きく異なる場合に，そのトピックを，特徴的な話題ラベルを持つ日のニュース記事，ブログ記事の例とともに提示 (図 3，図 4).

- ブログ記事の数よりもニュース記事の数の方が圧倒的に多く，ニュースに特有のトピック (図 5).
- ニュース記事の数よりもブログ記事の数の方が圧倒的に多く，ブログに特有のトピック (図 4，図 6).

図 3 の場合には，このトピック全体としては，「海外とのやりとり」に関連する多様なニュース記事，ブログ記事が対応しているが，日毎の出来事に応じて，頻度上位の話題ラベルが大き

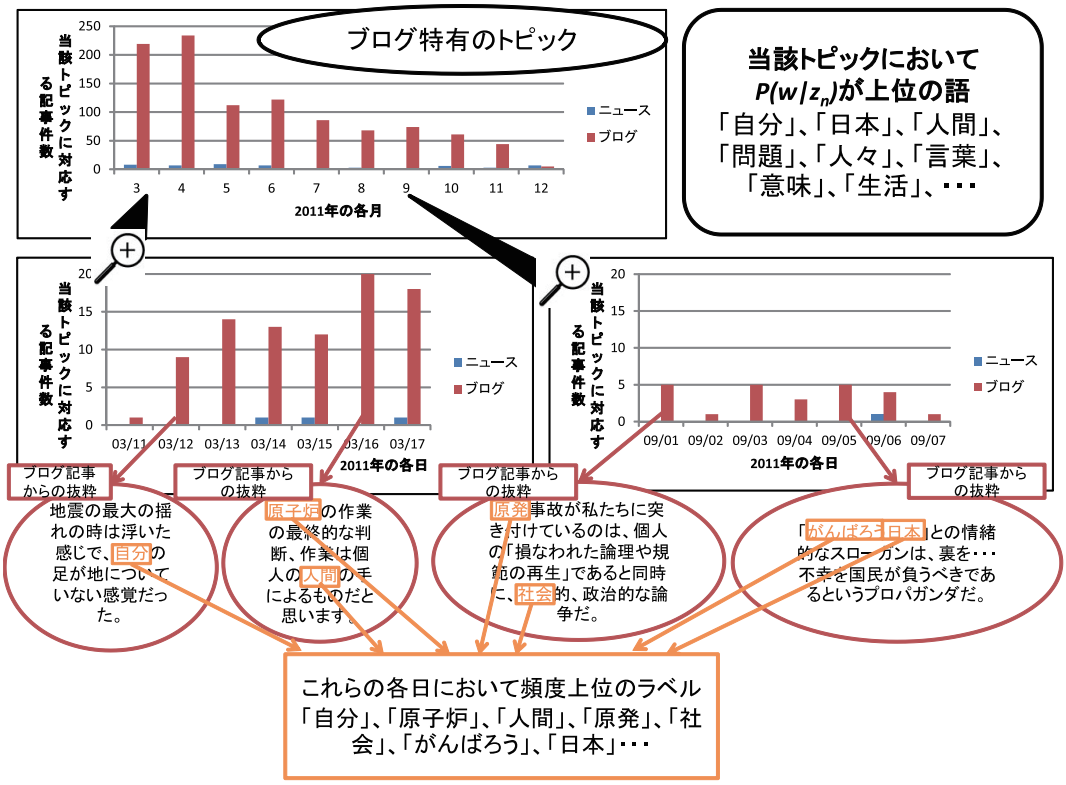

図 6　ブログ特有のトピックの例：「個人の意見や感想が中心」 (ニュース: 54 記事，ブログ: 1,025 記事)

く異なることが分かる．一方，図 4 は，特定の日付における出来事のついての報道を除いては，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．この場合は，「福島原発事故の放射能汚染」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を話題にする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，関連ニュース記事がやや増加する傾向がみられた．

図 6 は，震災発生以降，12 月に至るまで，多種多様な個人の意見や感想が集められた，極めてブログ特有のトピックとなっている．逆に，図 5 は，「震災による株式市場への影響」について報道するニュース記事が集められたトピックで，この話題については，ブログのおける関心があまり高くないことが分かる．

5. 新聞とテレビ放送との間の話題の分析

前節までの方式と同一の枠組みにより，ニュース (新聞記事)，ブログに加えて，2011 年 3 月 11 日から 12 月 31 日の期間の NHK 放送字幕テキストを混合した文書集合を対象として，ニュース (新聞)・ブログとテレビ放送との間の話題の相関と変遷の分析を行った．分析においては，ニュース (新聞) とテレビ放送を混合した文書集合にトピックモデルを適用したもの，および，ブログとテレビ放送を混合した文書集合にトピックモデルを適用したものの両方を対象として分析を行ったが，本論文では，前者の結果について簡単に述べる．

分析においては，まず，字幕テキストのうちの無音区間に挟まれたテキスト区間を一文書とし，前節と同様に，震災関係の七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文書中に出現するものだけを分析対象とした．その結果，分析対象の文書数は 32,847 文書となった．これらの文書および 4.1.1 節のニュース (新聞) 記事を混合した文書集合に対して，トピック数を 50 としてトピックモデルを適用した．

図 7 に示すように，トピック「子供を放射線から守る」について，特にテレビ放送に特有の現象として，NHK においては，「放射線」に焦点を当てた特集番組がいくつも放送されている．そのため，それらの特集番組の放送日においては，テレビ放送において，トピック「子供を放射線から守る」の文書数が多くなる要因の一つとなっていた．一方，ニュース (新聞)・ブログにおいては，これらの日，および，その後の数日において，これらの特集番組の影響によりトピック「子供を放射線から守る」の記事数が増加する，という現象は観測されなかった．

6. 関 連 研 究

文献 [14] においては，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 (EvoHDP; evolutionary hierarchical Dirichlet process) を適用し，各トピックの時系列の特徴をとらえる方式を提案している．この方式では，ニュース，ブログといった情報源ごとに、各月ごとのトピックを推定する．月ごとのトピックを推定する際には，隣接する月の間でトピックを関連付けてトピックの推定を行う．これに対して，本論文では，ニュース記事集合およびブログ記事集合の和集合に対して LDA を適用し，各トピックに対応するニュース記事およびブログ記事の分布を分析する

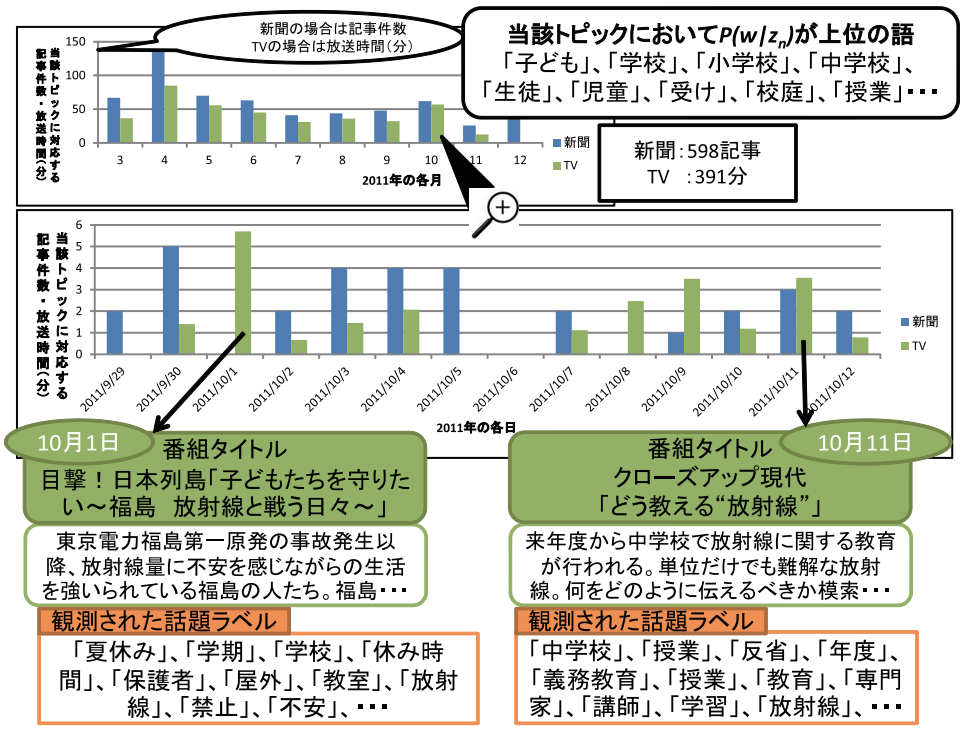

図 7　ニュース (新聞) とテレビ放送の比較におけるテレビ放送特有の内容の例：トピック「子供を放射線から守る」における特集番組

というより簡便な手法を用いている．また，各記事の内容を把握するためのラベル付けにおいては，Wikipedia を知識源として用いている．この方式を，震災関連のニュース記事，および，ブログ記事に適用することにより，多様な話題についてのニュース報道の動向，および，ブログ記事における関心の変遷をとらえることができることを示した．

文献 [5] 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．本論文の方式により，ニュース記事およびブログ記事といった他のメディアの記事とあわせて，Twitter の時系列データに対してトピックモデルを適用することにより，トピックとの関連や他のメディアとの現象上の違いの発見が容易になると考えられる．

7. おわりに

本論文では，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した．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．分析の結果，ニュース・ブログ間の相関が高いトピック，ニュース記事特有のトピック，ブログ記事特有のトピックを容易に発見することができた．

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] 石崎諒, 青野雅樹. Web ニュースに対するブログ意見の分析ツール. 電子情報通信学会技術研究報告, WI2-2008-52, pp. 11–12, 2008.

[5] 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.

[6] 小原恭介, 山田剛一, 絹川博之, 中川裕志. Blogger の嗜好を利用した協調フィルタリングによる Web 情報推薦システム. 第 19 回人工知能学会全国大会発表論文集, 2005.

[7] 小池大地, 牧田健作, 宇津呂武仁, 吉岡真治, 河田容英, 福原知宏. 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 —. 第 26 回人工知能学会全国大会論文集, 2012.

[8] 牧田健作, 横本大輔, 鈴木浩子, 宇津呂武仁, 河田容英, 福原知宏. Wikipedia を多言語知識源とするブログ集合の話題分析. 電子情報通信学会技術研究報告, NLC2011-18, pp. 95–100, 2011.

[9] 牧田健作, 横本大輔, 宇津呂武仁, 福原知宏. トピックに関する話題の時系列分布に着目したブログ分析. 第 3 回 DEIM フォーラム論文集, 2011.

[10] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. 21st SIGIR*, pp. 275–281, 1998.

[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.

[12] 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.

[13] 横本大輔, 鈴木浩子, 牧田健作, 宇津呂武仁, 河田容英, 福原知宏. 文書集合の話題俯瞰のためのクラスタリング手法. 第 4 回 DEIM フォーラム論文集, 2012.

[14] 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.