DEIM Forum 2012 F7-4

ブロガーの話題分布の俯瞰と分析

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

†1 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
†2 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
†3 (株) ナビックス　〒 141-0031 東京都品川区西五反田 8-3-6
†5 産業技術総合研究所　〒 135-0064. 東京都江東区青梅 2-3-26
†4 国立情報学研究所　〒 101-8430 東京都千代田区一ツ橋 2-1-2
†7 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 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

あらまし　本論文では，ブログ記事の書き手であるブロガーに注目し，特定のブロガーのブログ記事集合を対象として，話題のまとまりを同定し，一人のブロガーのブログ記事集合における話題分布の俯瞰，および，分析を行う方式を提案する．また，分析結果の一例として，幅広い話題にわたってブログ記事を投稿しているブロガーについて，話題の傾向・共通点を分析した結果を報告する．具体的な手法として，一人のブロガーのブログ記事を収集した文書集合に対して，トピックモデルを適用し，ブロガーのブログ記事集合におけるトピック分布を推定する．そして，各トピックに割り当てられたブログ記事の話題がどの程度まとまっているのかの評価を行う．さらに，各ブロガーのトピックを「医療」,「震災」,「国際」,「政治」といった分野に分類し，一人のブロガーのブログ記事の分野の多様性を分析する．
キーワード　ブログ, ブロガー, 話題分布, 集約

Analyzing Topic Distribution of a Blogger

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

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

Key words　blog, blogger, distribution of topics, aggregation

1. はじめに

現代の情報社会においては，情報の氾濫，すなわち，いわゆる情報爆発が起こっている．そして，そのように爆発する情報の集約や，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．ウェブ上の情報の一例として，近年，一般個人が自由に情報を発信するツールであるブログが世界中で普及し，各地域の人々がそれぞれインターネット上で個人の意見や評判を発信することが可能になった．それに伴い，様々な情報がブログに記載され，様々な人々の意見や評判が Web 上に氾濫するようになった．このような状況を鑑みて，本論文の前段として，我々は，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, 10–12, 16, 17]．具体的には，基本的な方

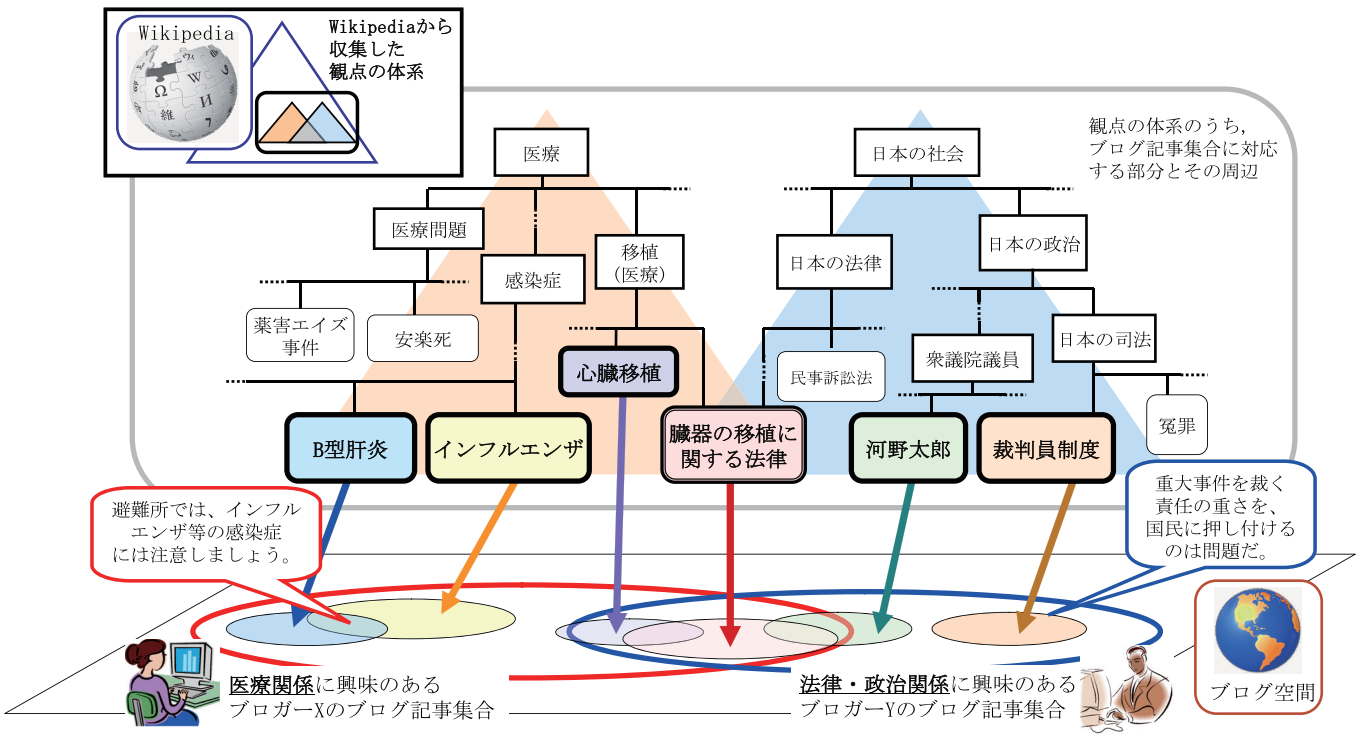

図 1　Wikipedia を知識源とするブロガーの話題分布の模式図

式 [8, 16, 17] 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 (ここで，話題として対応付けられた Wikipedia エントリを観点と呼ぶ) 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 [10, 12]，あるいは，時系列方向の話題の分布を分析する方式 [11] 等を提案した．

以上の研究の成果を背景として，本論文では，特に，ブログ記事の書き手であるブロガーに注目し，特定のブロガーのブログ記事集合を対象として，話題との対応付け，および，話題の分布の分析を行うことにより，ブロガーそのものの分析を行う方式を提案する．本論文の方式によるブロガーの話題分布の俯瞰の様子の模式図を図 1 に示す．この図では，「臓器の移植に関する法律」という共通の話題を持った二人のブロガーを対象として，ブログ記事の話題分布を示している．この図からは，ブロガー X が特に医療関係全般に渡って広く関心を持っているのに対して，ブロガー Y は，法律・政治全般に渡って広く関心を持っていることが容易に把握できる．このように，提案方式を用いることにより，ブロガーの話題の分布を効率よく俯瞰することが可能となる．

本論文では，具体的な手法として，一人のブロガーのブログ記事を収集した文書集合に対して，トピックモデル (本論文においては，LDA (Latent Dirichlet Allocation) [2] を用いた) (4.1 節) を適用し，ブロガーのブログ記事集合におけるトピック分布を推定する．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 (4.2 節)．これにより，各トピックに，どの程度の数のブログ記事が対応しているのかの分析を行う．また，各トピックに割り当てられたブログ記事の話題がどの程度まとまっているのかの評価を行う (5. 節)．さらに，各ブロガーのトピックを「医療」,「震災」,「国際」,「政治」といった分野に分類し，一人のブロガーのブログ記事の分野の多様性を分析する (6. 節)．

ここで，特定のブロガーのブログ記事集合における話題の分布を俯瞰するための既存の情報源としては，個々のブロガーによって整理されたサイドバーの情報が考えられる．しかし，それらはブロガーのブログ記事の内容を正確に知るためには不十分であることが多い．実際，社会問題に興味・関心の強いブロガーを中心とした 50 人程度のブロガーに対して調査を行ったところ，サイドバーのタグの粒度が粗いブロガーが全体の 74%を占めた．これに対して，本論文で提案する方式によれば，適切な粒度のもとでの話題の分布を容易に示すことができる点に大きな利点がある．

2. Wikipedia エントリとブログ記事の類似度

まず，ブログ記事集合中の各ブログ記事に対して観点を付与する手順について以下に述べる．

観点を付与する際には，観点とブログ記事の類似度を計算し，計算された類似度に基づいて付与する観点を決定する．類似度の計算においては，まず Wikipedia エントリ e の本文中に含まれる重要な語を関連語として抽出し，Wikipedia エントリ e の関連語の集合 $R(e)$ とする．そして，観点となる Wikipedia エントリ e の関連語 r ($\in R(e)$) がブログ記事 d の本文により多く出現しているほど類似度が高いとする．

具体的には，文献 [16] に基づいて次のように定義する．まず，

Wikipedia エントリ e の本文から収集した関連語 r ($\in R(e)$) の逆文書頻度 (idf, inverse document frequency)(注1) を重みとして関連語 idf ベクトル $\vec{I}$ を定義する．

$$\vec{I}(e) = \left(idf(r_1),\ \ldots,\ idf(r_n) \right)$$

一方，ブログ記事 d においても，Wikipedia エントリ e の関連語 r の出現頻度 $freq(d,r)$ を重みとして d のターム頻度ベクトル $\vec{G}(d,e)$ を次のように定義する．

$$\vec{G}(d,e) = \left(freq(d,r_1),\ \ldots,\ freq(d,r_n) \right)$$

そして，Wikipedia エントリ e とブログ記事 d の類似度 $Sim(e,d)$ は，2 つのベクトルの内積として次のように定義する．

$$Sim(e,d) \ = \ \vec{I}(e) \cdot \vec{G}(d,e) \ = \ \sum_{r \in R(e)} w(r) \times freq(d,r)$$

3. 分析対象のブロガー及びブログ記事の収集

3.1 特定キーワードについてのブログ記事の収集

本節では，評価の対象としたブログ記事集合の収集方法について述べる．本論文の評価においては，特定のキーワード t_0 を含むブログ記事を検索し，その結果得られたブログ記事集合を対象とした．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2) を利用し，日本語ブログホスト大手 6 社(注3)のドメイン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た．

3.2 分析対象ブロガーの選定

前節において，特定のキーワード t_0 を用いて収集したブログ記事のうち，ブロガー b によるブログ記事集合を $D(b,t_0)$ とする．そして， ブロガー b とトピック t_0 の Wikipedia エントリ $e(t_0)$ の類似度を

$$Sim_b(e(t_0),\, b) = \sum_{d \in D(b,t_0)} Sim(e(t_0),\, d)$$

として定義し，$Sim_b(e(t_0), b)$ の降順に 20 ブロガーを選定した．

3.3 ブロガーごとのブログ記事の収集

ブロガー b について，投稿日時が新しいものから順に最大約 500 件，ブログ記事を収集し，その集合を $D(b)$ と定義する．

(注1)：$idf(r) = \log\left(\frac{Wikipedia\ の総エントリ数}{関連語\ r\ が出現したエントリ数} \right)$ として定義する．

(注2)：http://developer.yahoo.com/search/boss/

(注3)：fc2.com, yahoo.co.jp, ameblo.jp, goo.ne.jp, livedoor.jp, hatena.ne.jp

4. トピックモデルを用いた話題分布の分析

4.1 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) [2] 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4)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，今回はトピック数を 10 から 50 まで 10 刻みでトピック推定した結果を人手で見比べ，5. 節の評価において最も精度のよいトピック推定結果が得られたトピック数を採用した．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．

4.2 文書に対するトピックの割り当て

本研究では，1 ブロガーごとに，ブロガーの書いた各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，ブログ記事を分類することとした．あるブロガーにおける文書集合を 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 $(n = 1, \ldots, K)$ のブログ記事集合 $D(z_n)$ は以下の式で表される．

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

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．

5. ブロガーのトピック分布の評価

本節では，各ブロガーのブログ記事集合に対して，トピックモデルにより推定されたトピックについて，各トピックに割り当てられたブログ記事の話題がどの程度まとまっているのかの評価を行う．

まず，評価対象のブロガーおよびブログ記事を収集するために，キーワードとして,「臓器移植」,「著作権侵害」,「電子書籍」を用い,「臓器移植」について 3 ブロガー,「著作権侵害」について 4 ブロガー,「電子書籍」について 5 ブロガーをそれぞれ選定し，ブログ記事集合を収集した．

次に，各ブロガーのトピック z_n について，ブログ記事集合 $D(z_n)$ 中において，$P(z_n|d)$ の降順で上位 5 位までのブログ記事について，同じ話題について書かれたブログ記事が $k(k = 3$ または 4) 記事以上含まれる割合について評価を行った．また，同じ話題について書かれたブログ記事が 3 記事以上含まれており，かつ，$P(z_n|b)$ 1 位のブログ記事の冒頭 500 文字の話題と 3 記事の話題が一致する割合について評価を行った．以上の結果を

(注4)：http://gibbslda.sourceforge.net/

表 1　各トピックにおけるブログ記事の話題のまとまりの評価

初期キーワード	ブロガー ID	トピックモデルのトピック数 K	評価対象トピック数	$P(z_n\|b)$ の上位 5 記事中に，同じ話題について書かれたブログ記事が k 記事以上含まれる割合 (%)		同じ話題について書かれたブログ記事が 3 記事以上含まれており，かつ，$P(z_n\|b)$ 1 位のブログ記事の冒頭 500 文字の話題と 3 記事の話題が一致する割合 (%)
				$k=4$	$k=3$	
「臓器移植」	1	30	12	75 (9 / 12)	100 (12 / 12)	92 (11 / 12)
	2	40	16	75 (12 / 16)	75 (12 / 16)	69 (11 / 16)
	3	40	7	57 (4 / 7)	57 (4 / 7)	57 (4 / 7)
	合計 / 平均 (ミクロ)	110	35	71 (25 / 35)	80 (28 / 35)	74 (26 / 35)
「著作権侵害」	4	40	16	88 (14 / 16)	94 (15 / 16)	94 (15 / 16)
	5	40	14	86 (12 / 14)	93 (13 / 14)	86 (12 / 14)
	6	40	29	69 (20 / 29)	93 (27 / 29)	93 (27 / 29)
	7	30	7	100 (7 / 7)	100 (7 / 7)	100 (7 / 7)
	合計 / 平均 (ミクロ)	150	66	80 (53 / 66)	94 (62 / 66)	92 (61 / 66)
「電子書籍」	8	40	12	92 (11 / 12)	100 (12 / 12)	83 (10 / 12)
	9	40	11	100 (11 / 11)	100 (11 / 11)	100 (11 / 11)
	10	40	12	83 (10 / 12)	100 (12 / 12)	92 (11 / 12)
	11	40	13	77 (10 / 13)	92 (12 / 13)	85 (11 / 13)
	12	40	11	73 (8 / 11)	91 (10 / 11)	82 (9 / 11)
	合計 / 平均 (ミクロ)	200	59	85 (50 / 59)	97 (57 / 59)	88 (52 / 59)
全キーワード合計 / 平均 (ミクロ)		460	160	80 (128 / 160)	92 (147 / 160)	87 (139 / 160)

表 1 に示す.

この結果から，$k=4$ および $k=3$ いずれの場合も，同じ話題のブログ記事が含まれる割合は，ブロガーごと，および，キーワードごとに異なるものの，平均して，80%，あるいは，90%程度の割合であることが分かる．また，$P(z_n|b)$ 1 位のブログ記事の冒頭 500 文字の内容から，トピック全体の話題が同定できる割合は 87%程度であり，比較的高いと言える．

6. ブロガーのトピックの分野分類

前節の結果から，各ブロガーについて，トピックモデルによって推定されたトピック z_n を参照して，$P(z_n|b)$ 1 位のブログ記事の冒頭 500 文字の内容を把握すれば，87%程度の精度で，トピック z_n の話題を把握できることが分かる．このことをふまえて，本節では，キーワード「臓器移植」を用いて収集した 3 ブロガーについて，トピックモデルによりブログ記事集合のトピック推定を行った結果について，全トピック (3 ブロガーについて，それぞれ，30 トピック，40 トピック，40 トピック) を分野に分類し，ブロガーのブログ記事の話題の多様性を分析した．この分類作業においては，作業の効率を重視して，本論文の著者の一人が，各トピック z_n について，$P(z_n|b)$ 1 位のブログ記事の冒頭 500 文字の内容のみを参照して，トピック z_n を分野に分類した.

図 2 に，3 ブロガーのトピック・ブログ記事を分野に分類した結果を示す．また，表 2 に，分野および各分野におけるトピック・ブログ記事の例を示す．これらの結果から，分析対象の 3 ブロガーについては,「医療」,「震災」,「国際」,「政治」といった多岐に渡る分野について，多様なトピックについてブログ記事を書いていることが分かる.

7. 関 連 研 究

本論文で焦点を当てた方式を一般化すると，文書集合に対して話題の分布を推定し，俯瞰する方式としてとらえることができる．以下では，そのような方式についての関連研究を概観する.

データに対して様々な観点のラベルを振り，データを検索する際に，ラベルを付与していくことで検索結果を絞り込んでいく，という考え方を，ファセット検索 [15] と呼ぶ．このファセット検索に関連する研究として，TREC-2009 におけるブログ検索タスク [9] においては，ファセット検索によるブログサイト検索タスクが導入され,「意見の有無」,「個人的情報・公的情報の別」,「トピックについて専門的あるいは詳細な情報を含むか否か」の 3 種類のファセットをブログサイトに付与するタスクが行われた．また，Web ページの検索結果に対して，観点を付与し，クラスタリングを行う手法の研究 [1, 4, 14] としては，Web ページの検索結果を分類し，各分類に対して適切な要約文を付与する手法 [4]，検索された個々の Web ページに対してラベルの付与を行い，付与されたラベルに基づいて分類を行う手法 [1, 14] 等が提案されている．ただし，これらの方式においては，分類対象の Web ページの情報のみを用いて分類を行っているため，分類対象のデータの規模が十分でなければ，分類が容易でなくなる，という問題がある.

また，その他には，トピック，ブロガー，リンク先，主観といったファセットを通じてブログを閲覧するもの [3] や，Wikipedia の記事を検索するインタフェースとして，エントリ集合から自

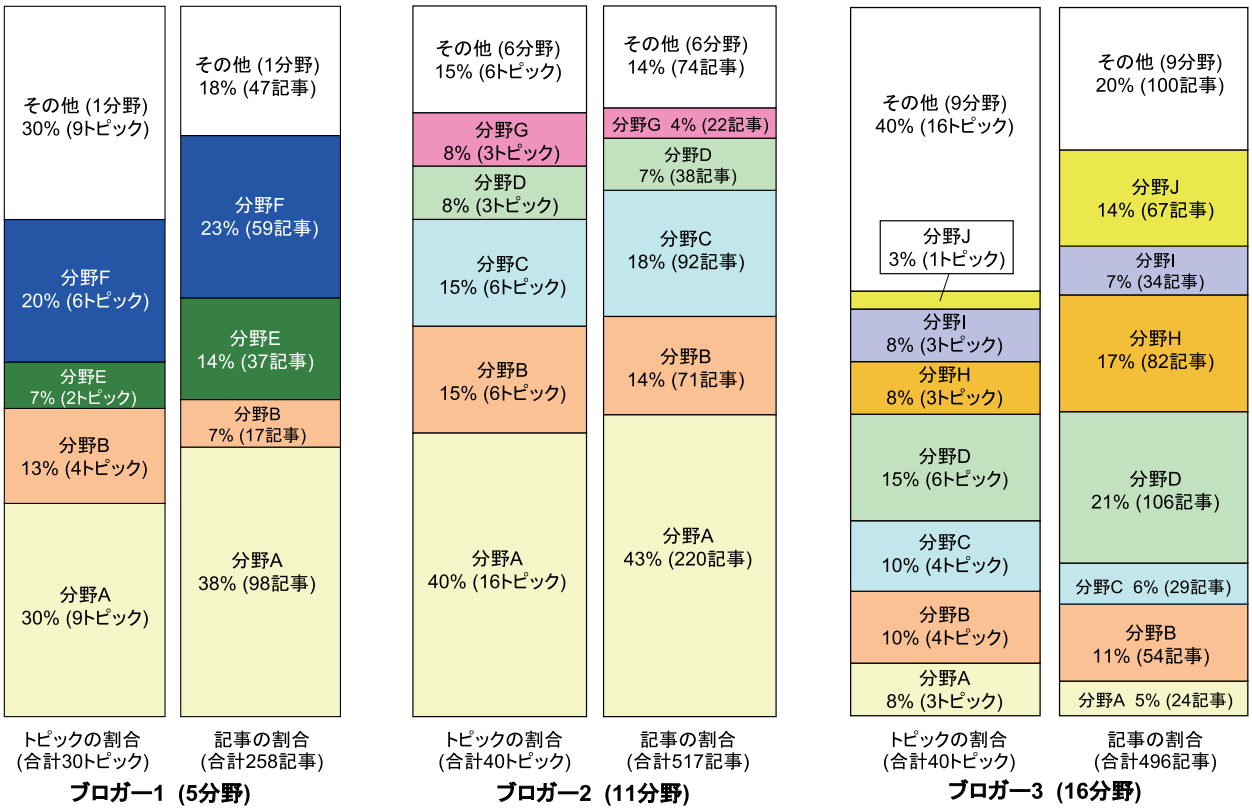

図 2　ブロガーのトピック・ブログ記事の分野分類の内訳 (キーワード:「臓器移植」)

表 2　分野・トピック・ブログ記事の例 (キーワード:「臓器移植」)

分野	ブロガー数	トピック数	トピックの具体例	ブログ記事の具体例	トピックの記事数	分野の記事数
医療	3	28	臓器移植推進パレード	臓器移植推進街頭パレードについて.	4	342
			臓器移植法改正	移植法改正後 1 年たった後の，総ドナー数について.	36	
			B 型肝炎訴訟	B 型肝炎訴訟の和解案についての意見.	29	
震災	3	14	計画停電について	計画停電が回避されたことについて.	11	142
			福島第一原発事故	福島原発のタービンがある建物からの水漏れについて.	23	
			被災地での医療活動	震災時の重病の患者や，老人介護についての意見.	29	
国際	2	10	反捕鯨団体	反捕鯨団体のメンバーの裁判について.	17	121
			海外・日本の原発について	世界で広がる反原発運動について.	11	
			パレスチナ問題	イスラエルとパレスチナの紛争についての意見.	13	
政治	2	9	介護保険制度改正	介護保険の改正案の紹介.	8	144
			発電施設建設	八ツ場ダムの建設についての意見.	9	
			与党寄りの意見	与党代表の辞任についての意見.	35	

動構築したファセット体系を利用するもの [7] などがある.

また，本論文に関連して，文献 [6] においては，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．特に，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ることを示している．文献 [13] においては，時系列ニュースに対してトピックモデルによりトピックを推定した後，トピック単位でのバーストを検出する手法を提案している．一方，文献 [5] においては，同様の手法により日本語・中国語二言語の時系列ニュースを対象として，トピック

モデルによりトピックを推定した後，二言語間でのトピックの対応を同定する手法を提案している．

8. おわりに

本論文では，ブログ記事の書き手であるブロガーに注目し，特定のブロガーのブログ記事集合を対象として，話題との対応付け，および，話題の分布の分析を行うことにより，ブロガーそのものの分析を行う方式を提案した．具体的な手法として，一人のブロガーのブログ記事を収集した文書集合に対して，トピックモデルを適用し，ブロガーのブログ記事集合におけるトピック分布を推定した．そして，各トピックに割り当てられたブログ記事の話題がどの程度まとまっているのかの評価を行った．さらに，各ブロガーのトピックを「医療」,「震災」,「国際」,「政治」といった分野に分類し，一人のブロガーのブログ記事の分野の多様性を分析した．

本論文の分析においては，各ブロガーのトピックを分野に分類する作業を人手で行ったが，今後の課題として，この過程を自動化する方式を実現する．具体的には，文書集合における話題の冗長性を削除して俯瞰性を高める方式 [18] を適用することにより，話題の粒度が比較的細かいトピックの単位を分野にまとめることを実現する．

また，本論文の冒頭では，図 1 において，複数のブロガーの間の話題の関連性を模式図で示したが，本論文の評価および分析の範囲においては，各ブロガーごとにトピックモデルを適用し，評価および分析も各ブロガーごとに行った．今後は，同一分野において共通の関心を持つ複数のブロガーのブログ記事集合を混合した文書集合を対象としてトピックモデルの適用を行い，話題の分布を分析することにより，複数のブロガー間の話題の関連性を自動的に同定する方式を確立する．

一方，5. 節の評価，および，6. 節の分野への分類においては，各トピックにおいて，$P(z_n|b)$ 1 位のブログ記事の冒頭 500 文字の内容によってそのトピックにおける話題の大勢が把握する方式を示し，その有効性を定量的に実証した．今後は，この実績に基づき，トピック集合および文書集合中を効率よくファセット検索するインタフェース，あるいは，そのようなインタフェースを用いたブロガーコミュニティ発見タスクといった，より上位の機構において，本論文の手法の有用性を評価する．

文　献

[1] 馬場康夫, 黒橋禎夫. キーワード蒸留型クラスタリングによる大規模ウェブ情報の俯瞰. 情報処理学会論文誌, Vol. 50, No. 4, pp. 1399–1409, 2009.

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

[3] 藤村考, 戸田浩之, 井上孝史, 廣嶋伸章, 片岡良治, 杉崎正之. マルチファセット型ブログ検索システム BLOGRANGER の開発. 電子情報通信学会技術研究報告, OIS2005-92, pp. 19–24, 2006.

[4] 原島純, 黒橋禎夫. PLSI を用いたウェブ検索結果の要約. 言語処理学会第 16 回年次大会論文集, pp. 118–121, 2010.

[5] 胡碩, 高橋佑介, 牧田健作, 横本大輔, 宇津呂武仁, 吉岡真治. 日中時系列ニュースにおけるトピックの推定と二言語間対応付け. 言語処理学会第 18 回年次大会論文集, pp. 179–182, 2012.

[6] 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.

[7] C. Li, N. Yan, S. B. Roy, L. Lisham, and G. Das. Facetedpedia: Dynamic generation of query-dependent faceted interfaces for Wikipedia. In *Proc. 19th WWW*, pp. 651–660, 2010.

[8] D. Lim, D. Yokomoto, K. Makita, T. Utsuro, and T. Fukuhara. Utilizing Wikipedia as a knowledge source in categorizing topic related Korean blogs into facets. 言語処理学会第 17 回年次大会論文集, pp. 876–879, 2011.

[9] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC-2009 blog track. In *Proc. TREC-2009*, 2009.

[10] 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.

[11] 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 回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, 2011.

[12] 鈴木浩子, 横本大輔, 牧田健作, 宇津呂武仁, 河田容英, 福原知宏. Wikipedia を知識源とする日英ブログ記事集合の観点分類と言語間対照分析. 情報処理学会研究報告, Vol. 2011, No. (2011–DBS–153), November 2011.

[13] 高橋佑介, 横本大輔, 宇津呂武仁, 吉岡真治, 河田容英, 神門典子, 福原知宏, 中川裕志, 清田陽司. 時系列トピックモデルにおけるバーストの同定. 第 4 回 DEIM フォーラム論文集, 2012.

[14] 戸田浩之, 中渡瀬秀一, 片岡良治. 特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案. 情報処理学会論文誌：データベース, Vol. 46, No. SIG 13(TOD 27), pp. 40–52, 2005.

[15] D. Tunkelang. *Faceted Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.

[16] D. Yokomoto, K. Makita, T. Utsuro, Y. Kawada, and T. Fukuhara. Utilizing Wikipedia in categorizing topic related blogs into facets. In *Proc. 12th PACLING,* #20, 2011.

[17] 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 回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, 2011.

[18] 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.