DEIM Forum 2013 C2-5

トピックモデルを用いたブロガー・コミュニティの収集と俯瞰

牧田　健作† 　鈴木　浩子† 　小池　大地† 　鄭　　立儀† 　宇津呂武仁††

河田　容英††† 神門　典子††††

† 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
†† 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
††† (株) ログワークス　〒 151-0053 東京都渋谷区代々木 1-30-15 天翔代々木ビル 6F
†††† 国立情報学研究所　〒 101-8430 東京都千代田区一ツ橋 2-1-2

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，数百人規模のブロガー集合のブログ記事を対象としてトピックモデルを適用することにより，ブロガー集合をコミュニティへと分類する．そして，トピックモデルによって推定されたトピックを利用することにより，ブロガー・コミュニティの俯瞰を行う方式を提案する．実際に,「にほんブログ村」に登録されているブロガーを収集し，コミュニティに分類・俯瞰した結果を報告する．さらに，推定されたトピックモデルの情報を利用して,「にほんブログ村」に未登録のブロガーの収集を行い，コミュニティを拡張した結果，約 3,300 ブロガーをコミュニティに追加できたことを報告する．

キーワード　ブロガー，ブログ，コミュニティ，話題分布，俯瞰，トピックモデル，トピック

Collecting and Overviewing Bloggers' Communities based on a Topic Model

Kensaku MAKITA†, Hiroko SUZUKI†, Daichi KOIKE†, Liyi ZHENG†, Takehito UTSURO††,

Yasuhide KAWADA†††, and Noriko KANDO††††

† Grad. Sch. of Systems and Information Engineering, University of Tsukuba, Tsukuba 305-8573 Japan
†† Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba 305-8573 Japan
††† Logworks Co., Ltd.　Tokyo 151-0053, Japan
†††† National Institute of Informatics, Tokyo 101-8430, Japan

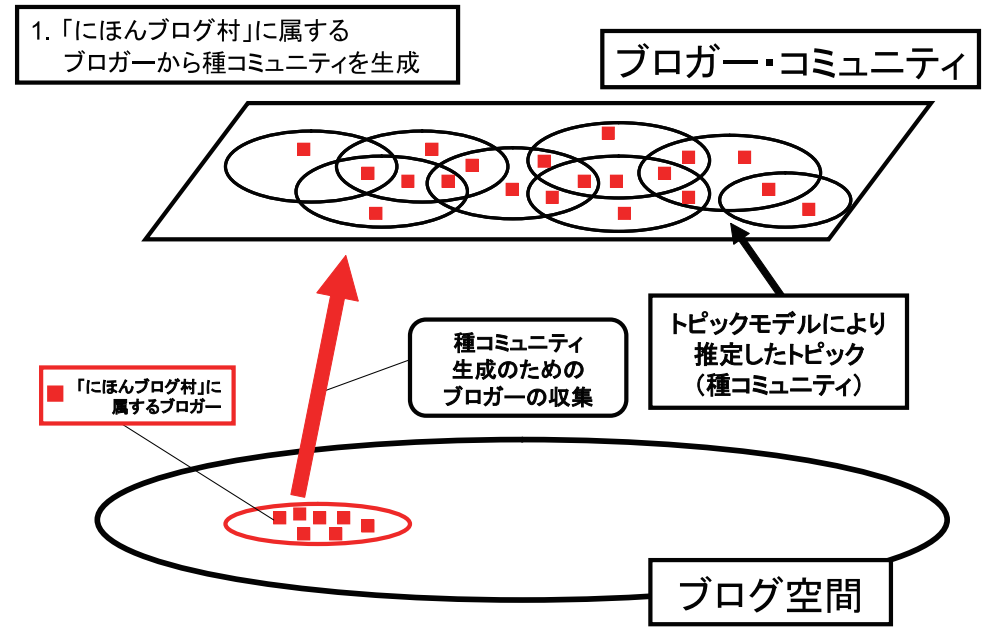
Abstract　This paper first studies how to apply a topic model to blog posts collected from a few hundred bloggers and then to classify bloggers into communities. Then, we study how to exploit the estimated topics in the task of overviewing the bloggers' communities. In the evaluation, we collect a few hundred bloggers from a well-known blogger community service "Nihon Blog Mura" in Japan, and then automatically generate 36 communities and overview them based on a topic model. Furthermore, we collect about 3,300 bloggers outside "Nihon Blog Mura" and then expand the 36 bloggers' communities by classifying the 3,300 bloggers into the 36 communities.

Key words　blogger, blog, community, distribution of topics, overview, topic model, topic

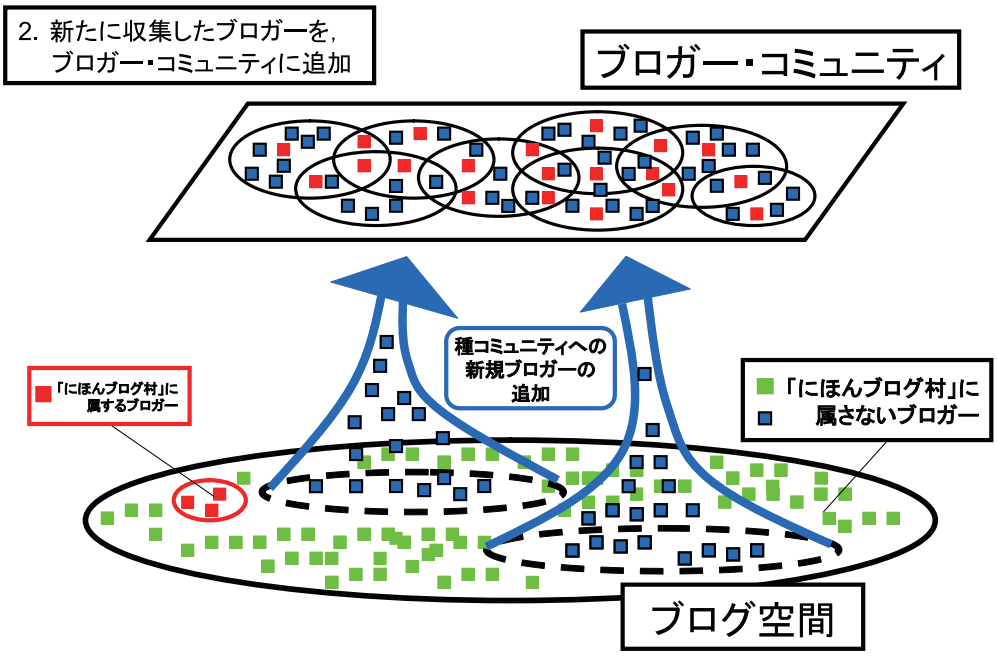
1. はじめに

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ウェブ上に氾濫するようになった．

本論文では，効率的な俯瞰を実現するために，ブログの書き手であるブロガーに注目し，同一の興味を持つブロガーのコミュニティを作成し，俯瞰を行う．ブログ記事を直接俯瞰するのではなく，ブロガーという情報発信者の単位で俯瞰を行うこ



(a) ブロガー・コミュニティの生成



(b) ブロガー・コミュニティの拡張

図 1　ブロガー・コミュニティの生成及び拡張

とで，人々の話題への関心の度合いを考慮した，効率的な俯瞰が実現できると考えられる．

ここで，人手で作成されたブロガー・コミュニティの代表的なものとして，「にほんブログ村」(注1) が挙げられる．しかし，「にほんブログ村」においては，登録を希望するブロガーが最大 3 個までのコミュニティへの登録を手動で行うだけにとどまっており，「にほんブログ村」外のブロガーを自動的に収集し，コミュニティに対して大規模に自動登録する機能は提供できていない．

そこで，本論文では，図 1 に示すように，「にほんブログ村」に所属するブロガーのブログ記事集合を用いて種コミュニティを自動生成し，種コミュニティと密接に関連するブロガーを「にほんブログ村」外から自動収集し，種コミュニティに追加することにより，コミュニティの自動拡張及び俯瞰を行う方式を提案する．

以下に，本論文の流れについて述べる．

(注1)：http://www.blogmura.com/

まず，2. 節において,「にほんブログ村」から，ブロガー・コミュニティ作成のためのブロガーおよびブログ記事を収集してくる手順について説明する．具体的には，2.1 節において,「にほんブログ村」についての基本情報について説明し，2.2 節において，具体的なブロガーおよびブログ記事収集手順について述べる．

次に，3. 節において，2. 節において収集したブロガーに対して，トピックモデル (本論文においては，LDA (Latent Dirichlet Allocation) [2] を用いた) の推定によるブロガー・コミュニティの生成手法について述べる．(注2) 具体的には，まず，3.1 節において，本論文におけるブロガー・コミュニティの定義について述べる．次に，3.2 節で LDA についての基本的な説明を述べ，3.3 節で LDA によって推定されたトピックに対するブログ記事の割り当て手法について述べる．そして，3.4 節において，トピックへのブログ記事の割り当ての精度評価と，トピックに割り当てられたブログ記事の話題のまとまりについて分析する．最後に， 節において，ブロガー・コミュニティの作成手順を述べ,「にほんブログ村」のカテゴリーに対して，より詳細なブロガー・コミュニティが生成されていることを示す．

最後に，4. 節において，生成したコミュニティに関わるブロガーを「にほんブログ村」外から収集し，ブロガー・コミュニティを拡張する手法について述べる．まず，4.1 節で，コミュニティ拡張のためのブロガーおよびブログ記事の収集手順について述べる．次に，4.2 節において，収集したブロガーをブロガー・コミュニティに追加する手順について述べる．最後に，4.3 節で，ブロガー・コミュニティの拡張性能の評価を行い，本手法によって，ブロガー・コミュニティの拡張が容易に達成できることを示す．

2. 分析対象のブロガー及びブログ記事の収集

2.1 「にほんブログ村」

「にほんブログ村」とは，日本最大級のブロガー・コミュニティであり，表 1 に示すように，多数のブロガーが多様なカテゴリーに登録されている．

「にほんブログ村」は，まず，ブロガーが属する一番大きなまとまりとして,「カテゴリー」があり，さらに，ひとつのカテゴリー内で「サブカテゴリー」として細かく分類されている．特に,「企業」「ベンチャー」「経営」「経済」カテゴリーにおけるサブカテゴリーを表 2 に示す．そして，カテゴリーおよびサブカテゴリーにおいて，属するブロガーへのアクセス数によって，人気ランキングが提示されている．

「にほんブログ村」においては，このように「カテゴリー」と「サブカテゴリー」の階層構造としてブロガーを分類することで，利用者が関心のある話題について言及しているブロガーへの効率の良いアクセスを実現している．

(注2)：先行研究 [6] においては，1 ブロガーのブログ記事集合に対して個別にトピックモデルの推定を行うことで，複数ブロガーの話題分布の分析を行なっていたが，本研究では，約 8,500 ブロガーを対象に，同一のトピックモデルを用いて，話題分布の推定を行なっている．

表 1 「にほんブログ村」のカテゴリー数およびブロガー数

カテゴリー数	サブカテゴリー数	ブロガー数
121	約 5,500	681,041

表 3 分析対象ブロガー数およびブログ記事数

	ブロガー数	ブログ記事数
「にほんブログ村」から収集したブロガー	240	7,708
「にほんブログ村」外から収集したブロガー	8,552	145,015

2.2 「にほんブログ村」に属するブロガーおよびブログ記事の収集

本研究では前節で述べた「にほんブログ村」から，ブロガーを収集する手順について述べる．本研究では,「にほんブログ村」のカテゴリーの中から,「企業」,「ベンチャー」,「経営」,「経済」の 4 カテゴリーに着目し，それぞれのカテゴリーに属するブロガーを収集した．まず，各 4 カテゴリーにおけるブロガーの人気ランキングの上位ブロガーから，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，各カテゴリーにつき 100 ブロガー，合計で 400 ブロガーを選定した．その後，各ブロガーごとに，最新の記事から記事が書かれた日付の降順に最大 50 記事収集し，その結果，記事が正しく収集され，分析対象となったブロガーは表 3 の「「にほんブログ村」から収集したブロガー」の欄に示すように，240 ブロガーとなり，収集されたブログ記事数は 7,708 記事となった．この 240 ブロガーの集合を B，収集されたブログ記事集合を D と置く．

3. トピックモデルを用いたブロガー・コミュニティの俯瞰

3.1 概　　要

本節では本研究における「ブロガー・コミュニティ」の定義について述べる．

本研究においては，前節において収集したブロガーが書いたブログ記事集合に対して，トピックモデルを推定することで，コミュニティを作成する．具体的には，ブロガーのブログ記事に対して，(本論文においては，LDA (Latent Dirichlet Allocation) を適用することによってトピックを推定し，トピックに対してブログ記事を分類する．このとき，各トピックに対して，同一トピックへのブログ記事のまとまりから，トピックの話題を人手で同定し，トピックに対する話題のラベル付けを行う．そして，あるブロガーの記事が一定数以上，同一トピックに張り付いたときに，ブロガーをそのトピックに貼り付ける．

以上の手順によって生成されるトピックごとのブロガー集合を,「ブロガー・コミュニティ」と定義する．このとき，トピックにおける話題ラベルが同様に，ブロガー・コミュニティの話題を示す話題ラベルとなる．

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

表 2　ブロガー・コミュニティの収集において用いた「にほんブログ村」のカテゴリー

カテゴリー	サブカテゴリー
企業	酪農・畜産業，林業・漁業，農業，鉱業，建設業，製造業 (食料品・飲料)，製造業 (繊維・衣服)，製造業 (木材・紙・印刷)，製造業 (化学・金属)，製造業 (機械・電気通信機械)，製造業 (電子部品・デバイス)，製造業 (その他) 電気・ガス・水道・熱供給，情報通信業，運輸業，卸売・小売業，金融・保険業，不動産業，飲食店・宿泊業，医療・福祉業，教育・学習支援業，その他サービス業，ベンチャー企業
ベンチャー	海外起業・海外独立，海外起業・海外独立支援，農林水産業，建設業，製造業，情報通信業，卸・小売業，飲食店・宿泊業，サービス業，その他業種，起業・独立支援，合同会社・合同会社設立，ベンチャー社長，社会起業家， シニア起業家，女性起業家，学生起業家，ベンチャーキャピタル，エンジェル投資家，ベンチャーの基礎知識，ベンチャー情報
経営	会長，オーナー社長，零細企業社長，中小企業社長，女性社長，若手社長，学生社長，老舗社長，二代目社長，三代目社長，後継者，IT 社長，経営者，役員・経営幹部，秘書・アシスタント，マネジメント，経営企画，組織・人材，人事労務・総務，財務・経理，法務・知財，生産・在庫，広報・IR，M & A，コーポレートガバナンス，企業 節税・税金対策， 海外進出支援・海外支援，コンサルタント，カウンセラー，コーチ，ビジネスボイストレーニング， 会話・コミュニケーション，ビジネスマナー，モチベーション，仕事術，営業，広告・マーケティング， フランチャイズ，オフィス・事務所，自営業・個人事業主，ファイナンシャルプランナー， 慈善事業・社会貢献，経営哲学・経営理念，経営学，MBA・MBA 取得，ビジネススクール・セミナー，異業種交流会， ロータリークラブ・ライオンズクラブ，倫理法人会，法人会，青年会議所・商工会議所，経営情報
経済	世界経済，アメリカ経済，ヨーロッパ経済，アジア経済・中国経済，日本経済， 地域経済，実体経済，金融経済，消費経済，経済学

3.2　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，今回はもっともトピックにおける記事のまとまりが良かった 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．

3.3　文書に対するトピックの割り当て

本研究では，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．ブログ記事集合を 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．

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

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．

3.4　トピックにおける話題の分析

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．

評価に際しては，各 $D(z_n)$ について，確率の降順に 5 記事選定し，人手でブログ記事の内容を判断し，5 記事中，3 記事以上同一の話題があったトピックを，正解とした．その結果，推定された 50 トピックのうち，約 70%の 36 トピックが正解と判定された．ここで，人手によって同定された話題が，トピックにおけるブロガー・ミュニティにおける話題を示すラベルとなる．同定されたブロガー・コミュニティの話題ラベルを，表 4 [注5] における「ブロガー・コミュニティの話題」の欄に示す．なお，これらの 36 個のコミュニティを，「にほんブログ村」中のカテゴリー・サブカテゴリーと比較したところ，すべてのコミュニティに対して，内容的に対応するカテゴリーもしくはサブカテゴリーが存在した．このことから，本論文で作成したコミュニティの範囲においては，「にほんブログ村」中のカテゴリーもしくはサブカテゴリーは十分網羅性があることが確認できた．

3.5　ブロガー・コミュニティへのブロガーの割り当て

本節では，前節で話題がまとまっていると判定されたされた 36 トピックに対して，ブロガーを割り当てることで，ブロガー・コミュニティを作成する手法について述べる．

まず，評価対象のブロガー集合 B について，B 中のブロガーを $b\,(b \in B)$ とする．そして，トピック z_n におけるブロガー b の記事集合を $D(z_n, b)$ とする．ここで，$D(z_n, b)$ が 5 記事以上となるような，トピック z_n に対して，ブロガー b を割り当てる．そして，B 中のすべての b について，トピックへブロガーの割り当てを行い，トピックごとのブロガー集合 $B(z_n)$ を，以下のように定義する．

$$B(z_n) = \left\{ b \in B \;\middle|\; |D(z_n, b)| \geqq 5 \right\}$$

各トピックに対する $B(z_n)$ におけるブロガー数の一覧を，表 4 の「「にほんブログ村」のみのブロガー数」の欄に示す．また，各コミュニティに対する，「にほんブログ村」の各カテゴリーの

(注4)：http://gibbslda.sourceforge.net/

(注5)：4.3 節の評価実験においては，ID1〜ID7 のトピックを対象とした予備実験の結果を報告する．そのため，表の表記として，ID1〜ID7 を強調している．

表 4　ブロガー・コミュニティの一覧

ID	ブロガー・コミュニティの話題	「にほんブログ村」のみのブロガー数 $\|B(z_n)\|$	ブロガー・コミュニティ拡張のためのクエリ	ブロガー・コミュニティ拡張後のブロガー数 $\|B(z_n) \cup B_{out}(z_n)\|$
1	**政治**	10	政治	336
2	**東日本大震災**	7	東電	219
3	**親子・家族**	5	子供	126
4	**経済**	14	経済	137
5	**保険・収入・家計**	12	収入	90
6	**企業運営**	8	企業	22
7	**起業支援**	2	起業	19
8	IT	7	インターネット	636
9	メディア・娯楽	3	映画	320
10	国際・外交	8	海外	272
11	人生論・ライフハック	20	人生	238
12	レストラン・グルメ	11	料理	221
13	法律・制度	10	課税	211
14	事件・時事・社会問題	6	事件	186
15	スポーツ	7	選手	181
16	社会学・思想	4	研究	134
17	建築・住宅	8	工事	128
18	就職・転職	5	採用	113
19	株式市場	9	為替	102
20	ネット通販	5	ネットショップ	90
21	工業	9	板金	85
22	勉強・スキルアップ	2	試験	84
23	電気・通信	1	発電	81
24	起業家向けの勉強会	7	セミナー	80
25	農業	15	収穫	70
26	貿易	7	貿易	70
27	インテリア	3	インテリア	51
28	仕事論・人生論	10	仕事	44
29	出版	4	原稿	41
30	経営戦略	3	コンサルティング	30
31	財政	2	税金	28
32	セミナー・勉強会	3	コーチング	26
33	融資・金融・経営	3	経営	22
34	物流業界	4	物流	11
35	ボディージュエリー	1	エアーブラシ	15
36	接客業	1	サービス	5
	合計 (延べ数)	236	—	4,524
	合計 (異なり数)	158	—	3,507

ブロガーの分布を，表 5 の「企業」,「ベンチャー」,「経営」,「経済」の欄に示す．

この結果から,「企業」「ベンチャー」「経営」「経済」の 4 カテゴリーのブロガーに対して，36 のブロガー・コミュニティが作成されており，より細かい粒度でのコミュニティ生成ができていることがわかる．これは,「企業」などのカテゴリーに関する話題が主であるブロガーであっても,「東日本大震災」や「レストラン・グルメ」など，日常的な関心事項をブログ記事に書くことも多いため，それらのブロガーの主題とは異なる話題を含む，幅広いコミュニティが生成できたのだと考えられる．これらのコミュニティを種コミュニティとして，次節でブロガー・コミュニティの拡張を行う．

なお，ここで，本節の手法によってブロガーをコミュニティに割り当てた結果を,「にほんブログ村」において各ブロガーが手動で登録を行ったカテゴリー・サブカテゴリーと比較して，その重複度合いの評価を行った．まず，36 個のコミュニティのうち少なくとも一つのコミュニティへ割当てられたブロガーを無作為に 50 ブロガー選定した．そして，それらのブロガーが「にほんブログ村」において所属している合計 105 のカテゴリー・サブカテゴリーを，本節の手法で割り当てを行った 66 個のコミュニティと比較したところ，内容的に対応するコミュニティの数は 16 個となった．この結果から,「にほんブログ村」のカ

表 5　ブロガー・コミュニティにおける「にほんブログ村」のブロガー数の分布

ID	ブロガー・コミュニティの話題	「にほんブログ村」におけるカテゴリー			
		企業	ベンチャー	経営	経済
1	**政治**	0	0	1	9
2	**東日本大震災**	1	0	0	6
3	**親子・家族**	0	4	1	0
4	**経済**	0	0	1	13
5	**保険・収入・家計**	4	1	2	5
6	**企業運営**	1	1	4	2
7	**起業支援**	0	2	0	0
8	IT	0	3	1	3
9	メディア・娯楽	0	0	2	1
10	国際・外交	1	3	0	4
11	人生論・ライフハック	0	6	11	3
12	レストラン・グルメ	7	1	2	1
13	法律・制度	3	3	2	2
14	事件・時事・社会問題	2	0	0	4
15	スポーツ	1	1	4	1
16	社会学・思想	0	0	0	4
17	建築・住宅	5	1	2	0
18	就職・転職	1	3	1	0
19	株式市場	0	0	0	9
20	ネット通販	0	3	1	1
21	工業	7	0	1	1
22	勉強・スキルアップ	1	1	0	0
23	電気・通信	0	0	0	1
24	起業家向けの勉強会	0	4	3	0
25	農業	14	0	1	0
26	貿易	2	0	0	5
27	インテリア	0	2	1	0
28	仕事論・人生論	0	2	7	1
29	出版	1	2	1	0
30	経営戦略	0	0	3	0
31	財政	0	0	0	2
32	セミナー・勉強会	0	2	1	0
33	融資・金融・経営	1	1	1	0
34	物流業界	3	0	1	0
35	ボディージュエリー	0	0	1	0
36	接客業	0	0	1	0
	合計 (延べ数)	55	46	57	78
	合計 (異なり数)	41	31	39	47

テゴリー・サブカテゴリーに対する再現率は，15.2% (16/105) となった．一方，本節の手法によって,「にほんブログ村」においては登録されていなかったブロガー・コミュニティ間の新規の所属関係を同定した割合は 75.8%(50/66) となった．この結果から，提案手法によって，ブロガー自身によるカテゴリー登録結果とは異質なコミュニティ所属関係が同定できることが分かった．

4. ブロガー・コミュニティの拡張

4.1 「にほんブログ村」外からのブロガーおよびブログ記事の収集

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．

まず，前節において生成した各トピック z_n における語 w の出現確率 $P(w|z_n)$ の上位 50 語から，そのトピックをできるだけうまく表すようなキーワード w_0 をクエリとして人手で選定する．各トピックから選定したクエリを，表 4 の「ブロガー・コミュニティ拡張のためのクエリ」の欄に示す．

次に，クエリ w_0 を含むブログ記事を収集する．クエリ w_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

(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

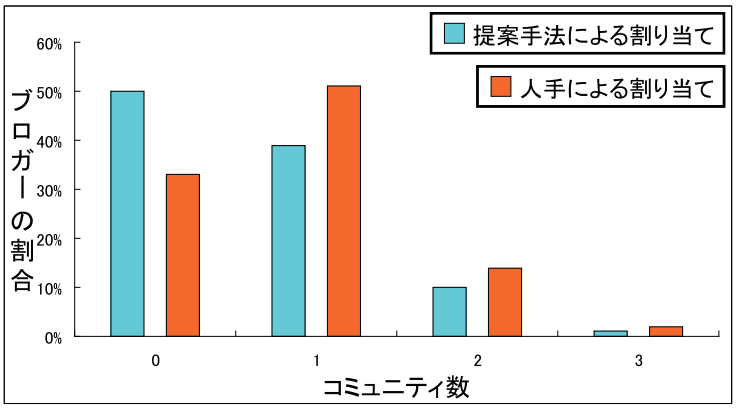


図 2　1 ブロガーに割り当てられたトピック数の分布

ンを対象としてブログ記事の収集を行った．収集の際には，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．そして，収集したブログ記事を書いているブロガーをブロガー・コミュニティ拡張のためのブロガーとする．

上記の手順をすべてのクエリに対して行い，その後，収集されたブロガーをごとに，最新の記事から記事が書かれた日付の降順に最大 20 記事収集する．

以上の手順によって，すべてのキーワードに対してブロガーの収集を行った結果，正しく記事が収集されたブロガー数およびブログ記事数を表 3 の「「にほんブログ村」外から収集したブロガー」の欄に示す．表から分かるように，収集したブロガー数，ブログ記事数はそれぞれ 8,552 ブロガー，145,015 記事となり，もともとの「にほんブログ村」から収集した 240 ブロガー，7,708 記事に対して，約 40 倍のブロガーと，約 20 倍のブログ記事を収集することができた．

ここで，新たに収集したブロガー集合を B_{out}, ブロガーが書いたブログ記事集合を D_{out} と定義する．

4.2　ブロガー・コミュニティへのブロガーの追加

本節では，前節で収集したブロガーを，ブロガー・コミュニティに追加する手順について述べる．

まず，ブロガーの書いた各ブログ記事 d ($d \in D_{out}$) に対して，3. 節によって推定したトピックモデルを用いて，ブログ記事のトピックを推定する．このとき，ブログ記事 d に対して，各トピックに対する出現確率 $P(z_n|d)$ $(n = 1, \ldots, K)$ が推定される．(ただし，3. により，$K = 50$)．

次に，3.3 節と同様に，以下の定義に従って，各ブログ記事 d に対してトピックを一意に割り当てることで，トピック z_n におけるブログ記事集合を定義する．

$$D_{out}(z_n) = \left\{ d \in D_{out} \;\middle|\; z_n = \underset{z_u\ (u=1,\ldots,50)}{\operatorname{argmax}} P(z_u|d) \right\}$$

そして，3.5 節と同様に，トピック z_n におけるブロガー $b(\in B_{out})$ のブログ記事集合を $D_{out}(z_n, b)$ とし，トピックごとのブロガー集合 $B_{out}(z_n)$ を以下のように定義する．

$$B_{out}(z_n) = \left\{ b \in B_{out} \;\middle|\; |D_{out}(z_n, b)| \geqq 5 \right\}$$

この $B_{add}(z_n)$ を，ブロガー・コミュニティに新たに追加された ブロガー集合とする．以上の手順で拡張された，各ブロガー・コミュニティの総ブロガー数（$B_{add}(z_n)$ と，「にほんブログ村」から収集したブロガーにおける，トピックごとのブロガー集合 $B(z_n)$ の和集合の要素数）を，表 4 の「ブロガー・コミュニティ拡張後のブロガー数」の欄に示す．表から分かる通り，「にほんブログ村」の 158 ブロガーで生成した種コミュニティに対して，約 22 倍の 3,507 人のブロガー・コミュニティを生成することができている．

表 6　ブロガー・コミュニティの拡張における追加先コミュニティの同定性能 (%) (評価対象 100 ブロガー)

適合率	再現率	F 値
88.7 (55 / 62)	64.7 (55 / 85)	74.8

4.3　評価及び分析

本節では，ブロガー・コミュニティ拡張の性能について評価を行う．

具体的には，前節で定義された，$B_{out}(z_n)$ について，ブロガー集合中のブロガー $b(z_n)(\in B_{out}(z_n))$ が，z_n に属するブロガーとして適切であるか否かの判定を行う．

評価対象として，表 4 に示す，ID1〜ID7 のトピック(注8)から選定したキーワードをクエリとして収集したブロガーの中から，100 ブロガー，1945 記事を選定(注9)した．

そして，100 ブロガーそれぞれの記事について，記事が属するトピックが適切であるかどうかを，手動で判定し，また，誤ったトピックが記事に張り付いていた時には，その記事に対して正解のトピックを人手で付与した．ここで，ブロガー $b(\in B_{out})$ における記事集合を $D_{out}(b)$ とし，$B^r_{out}(z_n)$ を，以下のように定義する．

$$B^r_{out}(z_n) = \left\{ b \in B_{out} \;\middle|\; D_{out}(b) \text{ 中, 5 記事以上} \right.$$

(注8)：3.4 節で述べたように，ここでは，ID1〜ID7 のトピックを対象とした予備実験の結果を報告する．現在，他トピックを対象とした本実験による評価を進めている．

(注9)：コミュニティに所属しているブログ記事のトピックへの出現確率の和の降順で 100 ブロガーを選定

$$\text{が } z_n \text{に属すると判定} \Big\}$$

そして，以下のように適合率・再現率を定義し，評価を行った．

$$\text{適合率} = \frac{\sum_{z_n} |B_{out}(z_n) \cap B^r_{out}(z_n)|}{\sum_{z_n} |B_{out}(z_n)|}$$

$$\text{再現率} = \frac{\sum_{z_n} |B_{out}(z_n) \cap B^r_{out}(z_n)|}{\sum_{z_n} |B^r_{out}(z_n)|}$$

その結果を，表 6 に示す．結果から，F 値が約 75%の性能で，ブロガーの拡張ができていることが分かる．

また，各ブロガーにおける，所属コミュニティ数の割合を図 2 に示す．このグラフから，自動判定ではどこのコミュニティにも属さないと判定されたブロガーが半数にのぼるが，手動での判定では 3 割程度となり，再現率の低さが最も顕著に現れていた．また，1 ブロガーが複数のコミュニティに属している割合が 16%程度であり，ブロガーの所属コミュニティ数が少ない傾向にあった．

5. 関 連 研 究

ウェブ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 [1,3,8] としては，ウェブ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 [3]，検索された個々のウェブ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 [1,8] 等が提案されている．ただし，これらの方式においては，分類対象のウェブ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，本論文に関連して，文献 [5] 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 [7] 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 [4] 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．

6. お わ り に

本論文では，多数の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ことで，同一の話題を持つブロガー・コミュニティを作成する方式を提案した．特に，人手で作成されたブロガー・コミュニティである「にほんブログ村」に対して，コミュニティへの大規模なブロガーの追加による，ブロガー・コミュニティの自動拡張を実現した．具体的には，まず,「にほんブログ村」から，特定のコミュニティに属するブロガーおよびブログ記事を収集し，収集したブロガーのブログ記事集合に対し，トピックモデルを推定し，ブロガー・コミュニティを作成した．そして，生成したコミュニティと密接に関連するブロガーを「にほんブログ村」外から収集し，ブロガー・コミュニティに追加した．以上の方式の有効性の評価を行ったところ,「にほんブログ村」の 240 ブロガーのブログ記事集合に対して，36 のコミュニティを自動生成することができ，さらに，これらのコミュニティと密接に関連する「にほんブログ村」外の約 3,300 ブロガーを 36 のいずれかのコミュニティに追加し，ブロガー・コミュニティの拡張及び俯瞰が行えることを実証した．

なお，本論文の評価実験の範囲においては，作成した 36 個のコミュニティは，いずれも,「にほんブログ村」のいずれかのカテゴリー・サブカテゴリーに対応することが確認できたが，今後の課題として,「にほんブログ村」内外のブロガー収集結果から自動的にコミュニティ生成を行い,「にほんブログ村」には登録されていないコミュニティを発見することが挙げられる．特に，その過程においては，新規に収集したブロガーに対して,「にほんブログ村」の全カテゴリー・サブカテゴリーとの所属関係を同定し，いずれのカテゴリー・サブカテゴリーにも該当しないことを自動で検出する方式の実現が不可欠であり，この方式の確立について重点的に取り組む予定である．

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] 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.

[4] 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.

[5] 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.

[6] 牧田健作, 横本大輔, 鈴木浩子, 宇津呂武仁, 河田容英, 神門典子, 福原知宏, 中川裕志, 吉岡真治, 清田陽司. ブロガーの話題分布の俯瞰と分析. 第 4 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, 2012.

[7] 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.

[8] 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.