トピックモデルのラベリングによる潜在的コミュニティの分析

白井　匡人† 　島田　諭†† 　三浦　孝夫†

† 法政大学 工学研究科　〒 184–8584 東京都小金井市梶野町 3-7-2
† 法政大学 マイクロ・ナノテクノロジー研究センター　〒 184–0003 東京都小金井市緑町 3-11-15
E-mail: †masato.shirai.5f@stu.hosei.ac.jp, ††{satoshi.shimada.57,miurat}@hosei.ac.jp

あらまし　近年，トピックモデルを用いた研究が多方面で提案されている．例えば，ソーシャルネットワークのユーザ分析やコミュニティ抽出に用いる．しかしながら，ここで言うトピックとは，確率的基準による一種のクラスタであり，ユーザやコミュニティの意味を明示的に捉えることは自明ではない．本研究では，予め与えたラベルを用いて潜在的トピックをラベル付けし，明示的な意味を付与する手法を提案する．

キーワード　テキストマイニング, ソーシャルネットワーク，トピックモデル, ラベリング

Masato SHIRAI†, Satoshi SHIMADA††, and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
† Research Center for Micro-Nano Technology, HOSEI University　3-7-2, MidoriMachi, Koganei, Tokyo, 184–0003 Japan
E-mail: †masato.shirai.5f@stu.hosei.ac.jp, ††{satoshi.shimada.57,miurat}@hosei.ac.jp

Abstract

Key words　Text Mining, Social Network,topic model,labeling

1. 前　書　き

近年，トピックモデルは文書集合の分析に非常に有効なモデルとして注目されており，ジャンル分類や文書要約など幅広く用いられてる [8]．また，twitter や facebook に代表されるソーシャルネットワークは極めて重要な情報源として注目されており，トピックモデルによるソーシャルネットワークのユーザ分類やコミュニティ抽出が行われている [7] [9].

トピックモデルとは Blei らによって提案された潜在的ディリクレ配分法に代表される確率的生成モデルであり，対象の文書や抽出したい知識に合わせてパラメータを追加することで，柔軟にモデルを構築することができる．例えばソーシャルネットワークを対象としたモデル化では文書の著者，文書の受信者，コミュニティ等のパラメータを加えることで，文書の内容とリンク構造を加味して潜在的コミュニティを抽出できることが示されている [2] [3]．これらのモデルでは著者とコミュニティがトピックによって特徴付けられているが，トピックとは確率的基準によって集められた一種のクラスタであり，潜在的トピックの混合で表されるユーザやコミュニティの意味は自明ではない。また，トピックには中身に一貫性のあるトピックと一貫性が定かでないトピックが混在する．この一貫性が無いトピックはストップワードのような多数の文書で頻度が高くなる語が集まっていることが多く，出現回数が多いことからトピック分布の中でも確率が高くなる傾向にあり，トピック分布の解釈が困難となる．このためトピック分布の混合で表される潜在的コミュニティに自動的にラベル付けする手法が望まれる．

トピックのラベル付けを行う手法はいくつか提案されており，Mei らにトピックのラベル付け [4] では, 実験データ中から候補ラベルを生成し，複数のスコア値を用いることで，トピックのラベル付けが行えることを示している．また，Ramage らによる Labeled LDA [5] では教師有り学習を行い，ラベルにトピックを 1 対 1 で対応付けて学習することにより文書をラベルとトピックの混合としてモデル化することができ，Twitter 上のユーザやツイートの特徴付けにも用いられている [6].

これらの手法はトピックにラベル付けする手法としては有用であるが．ラベルとトピックの関係しか考慮していないため，ユーザ分布とトピック分布から抽出される潜在的コミュニティのラベル付けに用いるには不十分である．例えばオリンピックや原発事故のような話題が大きく偏るイベントが発生した際に，コミュニティごとに特徴があるにも関わらず，どのコミュニティでも特定のトピックの確率が高くなり、一部のトピックの影響により各コミュニティで同じラベルが付けられる可能性がある．このためトピック分布のみによる潜在的コミュニティのラベル付けは困難であり，ユーザ分布やリンク構造など他の情報も加味する必要がある．

本研究では、実験データから抽出された候補ラベルを用いて潜在的トピックをラベル付けし、トピック分布とユーザ分布を加味してコミュニティにラベル付けすることで，潜在的コミュ

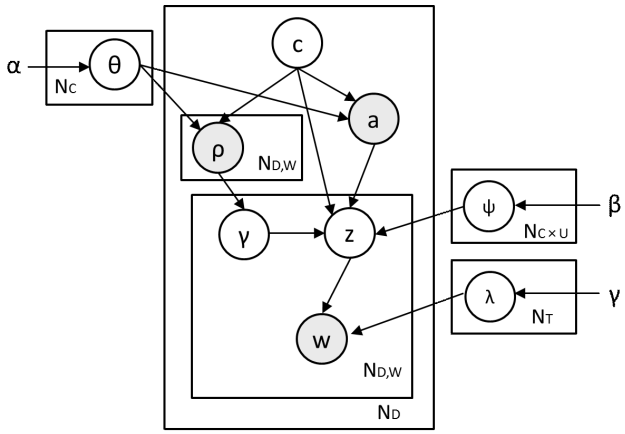

図 1　CART モデル

ニティに明示的な意味を付与する手法を提案する．

第 2 章では関連研究について述べ，第 3 章ではコミュニティ抽出手法について述べる．第 4 章ではラベリング手法について述べ．第 5 章では実験により有効性を示す．第 6 章で結論とする．

2. 関 連 研 究

2.1 トピックモデル

トピックモデルとは, 1 つの文書が複数のトピックの混合として表現されるという仮定である．1 つの文書が 1 つのトピックで表される混合多項分布に比べ, トピックモデルは文書が複数のトピックの混合分布として, 各トピックが単語の分布として表現され, 高い精度で文書をモデル化する可能性がある．

ソーシャルネットワークを対象としたモデルとしては Pathak らによる CART(Community-Author-Recipient-Topic) モデルが挙げられる．CART モデルではネットワークのリンク構造と文書の内容の両方を考慮してコミュニティを抽出することができる．トピックがコミュニティ，著者，受信者によって決まるとし，各パラメータを加えることでモデル化を行っている．ここで CART モデルのグラフィカルモデルを図 1 に示す．パラメータとして α,β,γ はディリクレ事前分布のパラメータ，θ はコミュニティ空間の多項分布,ψ はトピック空間の多項分布，λ は単語空間の多項分布である．ρ は受信者のセットであり，文書には複数の受信者が含まれる可能性を考慮している．この受信者のセットから 1 単語ごとランダムに受信者が選ばれ，コミュニティ c, 著者 a, 受信者 r の関係からトピックが決まる．これによりコミュニティごとに異なる著者の分布とトピックの分布が得られる．このモデルはソーシャルネットワークからコミュニティ抽出するモデルであるが，比較的小規模な会社内での E メールネットワークである Enron email コーパスを対象としており，大規模なソーシャルネットワークからコミュニティ抽出は行っていない．

3. コミュニティ抽出

3.1 Twitter からのコミュニティ抽出

本章では、トピックモデルを用いて Twitter の特徴を考慮した潜在的なコミュニティ抽出手法を提案する．大規模なソーシャルネットワークである Twitter に対応するために，CART モデルを変形しコミュニティ抽出に用いる．

Twitter 上では多数のユーザが文章を投稿しており，これらは特定のユーザへの投稿であるリプライや，他のユーザの投稿を再投稿するリツイートによってユーザ同士が繋がっている．Twitter からのコミュニティ抽出をモデル化するにあたり，ツイートの送り手と受け手のようなネットワーク構造を考慮することは重要である．しかしながら，トピックがコミュニティ, 著者，受信者の関係から求まるとした場合，必要なパラメータ数はコミュニティ数，著者数，受信者数の組み合わせであり，非常に多数のユーザが存在する Twitter では受信者をパラメータ化することは困難である．また、1 つのツイートは 140 文字以内の文字列であり，多くのツイートには数個の単語しか使われていない．このような Twitter の特性に対応するため，本手法では受信者のパラメータをモデルに組み込まない代わりに文書整形を行う．一定時間内ではユーザは似たトピックに関してツイートすると仮定し，同一時間内で同じユーザのツイートを受信者ごとに 1 つにまとめる．あるユーザが時間 T で複数回のツイートを行った場合，時間 T 内でのリプライの付いていない文書のまとまり 1 つとリプライしたユーザの種類数の文書を持つ．これにより文書の単語数を増やすことができ，さらに受信者の違いによる話題の違いを考慮することが可能となる．

ツイートから単語を抽出するには文字列を分割する必要があるが，ツイートでは新聞や小説などの文章より口語に近いくだけた表現が多く使われているため，形態素解析の精度が著しく悪化することが知られている．このため、本研究では文章から 2 文字以上連続する漢字、カタカナのみを抽出して単語として用いる．また，連続する文字数が多くなると複数の単語がまとまって 1 つの語として抽出されていることが考えられるので，抽出された語に対して形態素解析を行い単語を分割する．例えば、”昨日、首相官邸前抗議行動に行ってきた”という文章からは”昨日”,”首相官邸抗議行動”という 2 つの単語が抽出される。さらに各語に形態素解析を行うことで”昨夜”,”首相”,”官邸”,”前”,”抗議”,”行動”に分解され,2 文字以上の単語のみを用いることから，最終的に”昨日”,”首相”,”官邸”,”抗議”,”行動”の 5 つがこの文章の単語として抽出される．

3.2 コミュニティ抽出手法

潜在的コミュニティの抽出には文書が属するコミュニティと著者の関係をモデル化した CAT(Community-Author-Topic) モデルを用いる．グラフィカルモデルを図 2 に示す．パラメータとして α,β,γ はディリクレ事前分布のパラメータ，θ はコミュニティ空間の多項分布,ψ はトピック空間の多項分布，λ は単語空間の多項分布である．この生成モデルの生成過程は以下の通りである．

（1） 文書 d を生成するために，コミュニティ c_d が一様に選

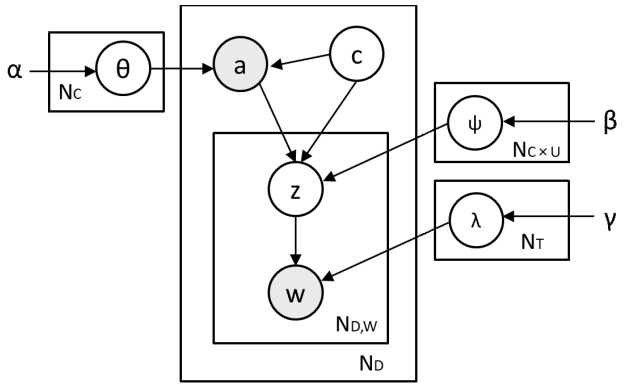

図 2　CAT モデル

ばれる．

（2） コミュニティ c_d を基に, 著者 a_d が選ばれる．

（3） コミュニティ c_d と著者 a_d を基に，トピック $z_{(d,i)}$ が選ばれる．

（4） トピック $z_{(d,i)}$ を基に，単語 $w_{(d,i)}$ が生成される．

このモデルでは著者はコミュニティに依存しており，トピックはコミュニティと著者に依存して選ばれる．これによりコミュニティに所属している著者の分布とコミュニティから生成されたトピックの分布が得られる．これらの結合確率は以下の式で表される．

$$p(c_d, a_d, \mathbf{z}_d, \mathbf{w}_d) = p(c_d)pr(a_d|c_d)\prod_{i=1}^{N_d} p(w_{(d,i)|z_{d,i}})p(z_{(d,i)}|c_d, a_d) \quad (1)$$

潜在変数の推定する手法としては，EM アルゴリズムや変分ベイズ法が挙げられるが，本研究では十分な反復回数の基で高い精度で潜在変数を推定できるギブスサンプリングを用いる．コミュニティに関する更新式は以下の式で求まる．

$$p(c_d = c|\mathbf{c}_{-d}, \mathbf{a}, \mathbf{z}, \mathbf{w}) \propto \frac{(n^{CU}_{-d,cu_i} + \alpha)}{\sum_{u=1}^{U}(n^{CU}_{-d,cu_i}) + U\alpha} \times \frac{\prod_{z=1}^{T}\Gamma(e_{d,z} + n^{(CU)T}_{-d,(c_d a_d)z}) + \beta)}{\Gamma(\sum_{z=1}^{T}(e_{d,z} + n^{(CU)T}_{-d,(c_d a_d)z})) + T\beta} \quad (2)$$

ここで n^{CU}_{-d,cu_i} は文書 d を除いてコミュニティ c からユーザ u_i が生成された回数,$e_{d,z}$ は文書 d でトピック z が生成された回数,$n^{(CU)T}_{-d,(c_d a_d)z}$ は文書 d を除いてコミュニティ c_d, 著者 a_d の組み合わせでトピック z が生成された回数,U はユーザの数,T はトピックの数である．次に各単語に対して割り振られるトピックに関する更新式は以下の式で求まる．

$$p(z_{d,i} = z|\mathbf{c}_{-d}, \mathbf{a}, \mathbf{z}_{-(d,i)}, \mathbf{w}_{-(d,i)}, c_d, a_d, w_{d,i} = w) \propto \frac{n^{TW}_{-(d,i),zw} + \gamma}{\sum_{v=1}^{W} n^{TW}_{-(d,i),zv} + W\gamma} \times \frac{n^{(CU)T}_{-d,(c_d a_d)z} + \beta}{\sum_{z=1}^{T} n^{(CU)T}_{-d,(c_d a_d)z} + T\beta} \quad (3)$$

ここで $n^{TW}_{-(d,i),zw}$ は文書 d の i 番目の単語以外で単語 w にトピック z が割り振られた回数，V は総単語数である．

4. ラベリング手法

4.1 候補ラベルの生成

本研究では，実験データ中から抽出された候補ラベルを用いてコミュニティのラベル付けを行う．コミュニティの内容を表す語の単位として 1 単語では不十分であるため，候補ラベルの語の単位を 2-gram とし，実験データの各文書から 2-gram を抽出して出現頻度の上位 1000 語を候補ラベルとする．2-gram は繋がりあった語を抽出するのに有用であるが，ここで抽出される候補ラベルは単語頻度に影響を受けるため，各単語の頻度が高い語のみとなる．例えば，頻度の高い”今日”や”明日”などに続く語は繋がりが弱い語であっても多くの語が候補ラベルとなる．逆に”太陽光”と”発電”という語の間には強い繋がりがあると考えられるが，各単語自体の頻度が高くないために候補ラベルとならない．このため高頻度語以外の候補ラベルを抽出するために，語の繋がり捉える指標としてコロケーション抽出などに用いられる相関ルールの確信度を使用する．共起する語 w_1,w_2 の出現回数を n_1,n_2, 共起回数を n_{12} としたとき，w_1 を条件として w_2 が出現する確信度 $w_1 \Rightarrow w_2$ は以下の式で求まる．

$$C(w_1 \Rightarrow w_2) = \frac{n_{12}}{n_1} \quad (4)$$

ここで求まる確信度の上位 1000 語を候補ラベルに加える．

4.2 ラベルの推定

まず潜在的コミュニティのラベル付けをするために潜在的トピックのラベル付けを行う．トピックはそれぞれ単語分布を持っていることから，トピック z の基でラベル l が出現する確率を求め，尤度の最も高い候補ラベルをトピックのラベルとする．トピック z のラベル l は以下の式で求まる．

$$Ans(l) = \arg\max_{l} p(l|z) \quad (5)$$

次に潜在的コミュニティのラベル付けを行う．潜在的コミュニティはトピックの確率分布 T とユーザの確率分布 U で表されており, この 2 つの確率分布を掛け合わせることで候補ラベルを決定する．これにより，例えトピック分布が同じであってもユーザ分布の違いにより異なるラベルが付けられる．コミュニティのラベルとなる候補ラベル l は以下の式より求まる．

$$Ans(l) = \arg\max_{l} p(l|z)p(z|u, c) \quad (6)$$

5. 実　　験

本章では 2 つの実験を行う．第 1 に Twitter から潜在的コミュニティの抽出を行う．第 2 に抽出された潜在的コミュニティに対してラベル付けを行う．

ユーザ数	文書数	異語数	総単語数
8870	169979	63778	3166073

表 1　実験データ

5.1 実験準備

実験に用いる Twitter のコーパスは 2012 年 7 月 1 日の 1 日分であり，この中でツイート数が上位 10000 となるユーザを使用する．ツイート数が非常に多いユーザに関しては”bot”(注1)である可能性が高くなるので，1 日で 200 回以上のツイートを行っているユーザを除外する．3 章で述べた手法によりツイートから語を抽出し，抽出された語の形態素解析には mecab を用いる．2 時間ごとにツイート情報をまとめる．ここで文書の単語数が 2 以下である文書では正確なトピックを推定することが困難なため取り除く．また，実験データ中でノイズとなる出現頻度が 1 の単語と高頻度語となりトピックの語分布の特徴付けを妨げる出現頻度が 4000 以上の単語を取り除く．取り除かれる単語の異語数はそれぞれ 66026 語，87 語である．最終的に使用する実験データのユーザ数，文書数，異語数，総単語数を 1 に示す．相関ルールの最小支持度は 300/総単語数 とし，トピックモデルの各パラメータは $\alpha = 0.01$,$\beta = 0.01$,$\gamma = 0.01$, トピック数 40, コミュニティ数 6, ギブスサンプリングの繰り返し回数 200 とする．

5.2 比較方法

抽出された確率分布の比較には JS ダイバージェンスを用いる．JS ダイバージェンスとは左右対称な確率分布の差の尺度であり確率分布 P と確率分布 Q が与えられたとき，JS ダイバージェンスは以下の式で求まる．

$$JS(P//Q) = \frac{1}{2}(\sum_x P(x) log \frac{P(x)}{R_(x)} + \sum_x Q(x) log \frac{Q(x)}{R(x)}) \quad (7)$$

ここで R は確率分布 P と Q の平均であり，$R = \frac{P+Q}{2}$ となる．確率分布の差が大きいほど JS ダイバージェンスの値も大きくなり P と Q が等しいとき JS ダイバージェンスは 0 となる．

5.3 実験結果

第 1 の実験より，抽出された潜在的コミュニティのトピックとユーザのトップ 10 を表 2 に示す．各コミュニティで最も出現する確率の高いトピックはコミュニティ1 から順に 16,28,33,16,16,28 であり，コミュニティ1,4,6 とコミュニティ2,6 で同じトピックが最上位に出現している．各コミュニティで最上位となるユーザは 956,847,4145,695,4045,2296 と全て異なっている．次に各コミュニティでトップになったユーザのトピック分布を 4 に示し，トピックの単語分布を表 5 示す．4 はコミュニティで最上位になったユーザであるが，ユーザ 695,2296 以外では最上位となるトピックが異なっている．各コミュニティのユーザ分布とトピック分布を JS ダイバージェンスにより比較した結果を表 3 に示す．次に第 2 の実験より，トピックのラベル付けの結果を表 6 に示し，潜在的コミュニティのラベル付け結果を表 7 に示す．各トピックに付けられるラベルはすべて異なっており，同じトピックが最上位になるコミュニティであっても異なるラベルが付けられている．

5.4 考　　察

表 2 より，各コミュニティのトピック分布は多くの同じトピックが上位に出現しているが，ユーザ分布ではトップ 5 に含まれる著者がすべて異なっている．これは表 3 の JS ダイバージェンスによる比較が示すようにコミュニティ同士のトピック分布の JS ダイバージェンスの平均が 4.19×10^{-4} であるようにトピック分布の差が小さくなっているためであると考えられる．コミュニティ1,4,5 でトップになったトピック 16 は，コミュニティ2,3,6 でも上位に出現しいる．トピック 16 のラベルでは上位語同士の組み合わせのラベルが出現しておらず，3 番目のラベルでは”チケット”,”発売”の両方がトップ 10 に含まれていない．これはトピック内の上位語に対する関連した語が集まっていないためであり，様々な文書で現れた語の集まった一貫性の無いトピックであると考えられる．表 5 と表 6 よりトピックのラベルとなる語はトピックの単語分布でトップ 10 に入る語の組み合わせだけでなく，トップ 10 に入ってない語との共起語が選ばれている．これは相関ルールの確信度による候補ラベルを加えたためだと考えられる．またトピック 2 のラベルでは実験データの日時である 7 月 1 日に公開する映画に関するラベル，トピック 4 では”プール 冷却”,”冷却 装置”など実験データ中に起こったイベントに関するラベルが付けられている．コミュニティのラベルでは，コミュニティでトップになったトピックとは違うラベルが付けられている．表 4 で示した各コミュニティでトップになったユーザのトピックとコミュニティで上位に出現するトピックが異なっているためであり，コミュニティのラベル付けにユーザの分布も条件に与えたためであると考えられる．

6. 結　　論

本論文では，潜在的コミュニティのラベリング手法を提案した．トピックモデルにより Twitter から潜在的コミュニティを抽出することができ，ラベル付けの際にユーザ分布を条件として与えることでコミュニティのトピック分布の差が少なくとも異なるラベル付けが行えることを示した．

(注1)：Twitter 上には大量の文章を自動的に投稿するプログラムである bot が多数存在する

文　　献

[1] Blei, D. M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, Journal of Machine Learning Research 3, pp. 993-1022, 2003.

[2] Ding Zhou, Eren Manavoglu, Jia Li, C.LeeGiles, Hongyuan Zha“Probabilistic Models for Discovering E-Communities”,Proceedings of the 15th international conference on World Wide Web(WWW’06),2006

[3] Nishith Pathak, Colin DeLong, Kendrick Erickson,and Arindam Banerjee “Social Topic Models for Community Extraction”, The 2nd SNA-KDD Workshop’08(SNA-KDD’08), 2008.

[4] Qiaozhu Mei, Xuehua Shen, and Chengxiang Zhai “Automatic Labeling of Multinomial Topic Models”,Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining(KDD’07),pp.490-

C1		C2		C3		C4		C5		C6	
ユーザ番号	トピック	ユーザ	トピック	ユーザ番号	トピック	ユーザ番号	トピック	ユーザ番号	トピック	ユーザ番号	トピック
956	16	847	28	4145	33	695	16	4045	16	2296	28
471	34	1521	30	4256	16	6732	30	6512	28	480	16
6381	20	2738	15	67	28	910	28	2037	18	2640	11
7884	33	744	20	844	11	1970	18	4059	34	2705	36
69	28	925	33	2536	20	3191	33	5530	33	4404	33
1521	15	2173	34	2705	34	3301	34	2536	20	2536	20
2640	30	6602	11	7884	18	5516	15	2640	10	3405	15
5895	18	67	16	1547	30	6582	11	7640	11	5654	30
8351	11	572	10	1970	9	471	9	7765	30	8816	34
1476	2	3463	18	1393	15	801	10	553	36	1212	10

表 2　各コミュニティのユーザとトピックのトップ 10

	C1	C2	C3	C4	C5	C6
(トピック)						
C1		3.35E-04	3.00E-04	3.62E-04	3.55E-04	5.03E-04
C2			3.63E-04	4.88E-04	3.62E-04	3.38E-04
C3				4.16E-04	4.68E-04	5.16E-04
C4					3.71E-04	6.34E-04
C5						4.77E-04
C6						
(ユーザ)						
C1		0.0342	0.0341	0.0338	0.0338	0.0330
C2			0.0343	0.0338	0.0340	0.0333
C3				0.0339	0.0341	0.0336
C4					0.0337	0.0331
C5						0.0334
C6						

表 3　コミュニティ同士の比較

956		847		4145		695		4045		2296	
トピック	確率	トピック	確率	トピック	確率	トピック	確率	トピック	確率	トピック	確率
2	0.144	7	0.119	33	0.331	6	0.246	22	0.160	6	0.0976
34	0.141	35	0.112	15	0.122	5	0.136	38	0.112	10	0.0918
33	0.132	21	0.105	37	0.093	30	0.126	3	0.085	25	0.0861
22	0.100	20	0.085	35	0.086	24	0.087	9	0.082	7	0.0746
38	0.072	36	0.066	26	0.077	16	0.085	8	0.080	31	0.0660
1	0.063	0	0.063	8	0.075	15	0.053	15	0.078	4	0.0617
10	0.039	16	0.058	19	0.045	38	0.051	35	0.056	19	0.0516
28	0.039	23	0.054	13	0.039	28	0.041	0	0.051	2	0.0502
12	0.037	6	0.046	32	0.039	37	0.028	10	0.046	12	0.0473
15	0.035	32	0.044	9	0.027	34	0.026	25	0.046	11	0.0430

表 4　ユーザのトピックトップ 10

1		2		4		16		28		33	
単語	確率	単語	確率	単語	確率	単語	確率	単語	確率	単語	確率
暴力	0.00304	新宿	0.00429	冷却	0.01040	開始	0.00220	花火	0.00239	英語	0.00207
抗議	0.00258	発表	0.00272	プール	0.00783	予約	0.00163	試合	0.00228	社会	0.00193
市民	0.00207	公開	0.00265	装置	0.00554	韓国	0.00157	大会	0.00225	学校	0.00185
機動	0.00198	予約	0.00189	燃料	0.00547	大変	0.00154	風呂	0.00178	人人	0.00170
大事	0.00196	部屋	0.00182	停止	0.00536	全員	0.00153	言葉	0.00177	テレビ	0.00167
警察	0.00194	風呂	0.00160	使用	0.00486	テレビ	0.00151	ドン	0.00177	発売	0.00166
ダンナ	0.00194	開催	0.00158	東電	0.00363	結構	0.00149	大変	0.00171	結構	0.00166
大変	0.00181	大変	0.00151	事故	0.00342	開催	0.00147	韓国	0.00165	最初	0.00152
女性	0.00180	上映	0.00151	連絡	0.00260	元気	0.00146	来週	0.00158	数学	0.00150
結構	0.00176	女性	0.00148	状況	0.00253	リプ	0.00142	結構	0.00154	女性	0.00149

表 5　トピックの単語トップ 10

	トピック 1	トピック 2	トピック 4	トピック 16	トピック 28	トピック 33
1	抗議 活動	公開 新宿	プール 冷却	生放送 開始	花火 大会	発売 決定
2	抗議 行動	新宿 バルト	冷却 装置	チケット 予約	大会 コピー	数学 理科
3	官邸 抗議	ヱヴァ 公開	装置 冷却	チケット 発売	恋人 花火	公開 決定

表 6　トピックのラベリングトップ 3

499,2007.

[5] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D.Manning “Labeled LDA:A supervised topic model for credit attribution in multi-labeled corpora”,Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing(EMNLP’09),pp.248-256,2009.

[6] Daniel Ramage, Susan Dumais, and Dan Liebling “Characterizing Microblogs with Topic Models”,AAAI’10,2010

[7] Liangjie Hong, and Brian D.Davison “Empirical Study of Topic Modeling in Twitter”,1st Workshop on Social Media Analytics(SOMA’10),2010

[8] Masato Shirai, and Takao Miura “On Domain Independence of Author Identification”,IDEAL’11, 2011

[9] Mrinmaya Sachan, Danish Contractor, Tanveer A.Faruquie,

	C1	C2	C3	C4	C5	C6
1	一方 詐欺	リフォロー 是非	配信 開始	プール 冷却	抗議 活動	男子 仕方
2	一方 拒否	パチンコ 野球	全員 全力	行動 報道	抗議 行動	カップル 会話
3	一方 会話	相撲 リフォロー	チケット 発売	冷却 装置	官邸 抗議	カップル 注

表 7　コミュニティのラベリングトップ 3

and L. Venkata Subramaniam “Using Content and Interactions for Discovering Communities in Social Networks”,WWW’12,2012