Vol.2014-DBS-159 No.14
Vol.2014-IFAT-115 No.14
2014/8/2

一般ユーザの観点に基づくTwitterからの人物関係の可視化

西村 章宏[1]　土方 嘉徳[1]　三輪 祥太郎[2]　西田 正吾[1]

概要：マイクロブログサービスの 1 つである Twitter では，その時々で話題になる政治家や芸能人など有名人に関する一般ユーザの発言を豊富に得ることができる．さらに近年では，SNS から得られる評判情報をマーケティングやその他サービスに応用しようという試みが活発に行われている．そこで本研究では，Twitter から得られる評判情報のうち，一般ユーザの有名人に関する発言とその発言を行ったユーザのプロフィール等に着目する．これらの各情報源から得られるデータに対し，抽出の妨げとなるノイズへの前処理を経て，一般ユーザの観点が反映された特徴量であるトピックの抽出を行う．そして得られたトピックの分布を元に人物の類似関係を獲得し，それを基に各人物を平面上に配置することで，人物関係の可視化を行う．この可視化結果に対しては使用した情報源毎に妥当性と発見性に着目して特徴の分析を行う．

1. はじめに

Web からあらゆる情報を容易に得ることができるようになった現代，ユーザは通学や通勤途中，家事の最中，友人との会話の最中など，日常生活の様々なシーンにおいて情報をパソコンやスマートフォンを使って獲得している．このようなシーンで獲得したい情報は，仕事や勉学の問題を解決するためのものだけでなく，ふと気になった些細な内容であることも多い．このような些細な気になる事柄の 1 つとして，有名人に対する興味や関心を取り上げる．これらは，例えばテレビや Web ページを見ている時や，友人と会話をしている時に，ふと目にした・耳にした有名人に対するものである．具体的な例を挙げると，

- 最近ニュース番組でよく耳にする名前であるが，その人がどういった人物なのかあまり知らない．
- 友人間でよく話題になる芸能人だけど，その人についてよく知らず，会話の内容についていけなかった．
- 自分が支持している政治家に対して，世間一般の印象がどのようなものか気になった．

といった場面が想像できる．このような際にユーザの多くは，wikipedia やタレント名鑑のようなオンライン百科事典を利用し，年齢や所属といった客観的事実からその人物を把握していると思われる．しかし，その有名人が世間一般ではどのように捉えられているかを表す印象や感想といった客観的ではない情報（以下，一般ユーザの観点に基づく人物情報）は，上記のような情報源からは得ることができない．上記のようにふと気になった有名人については，客観的な事実に関する情報だけでなく，世間一般の印象や評判についても知りたいところである．この一般ユーザの観点に基づく人物情報が多く存在するサイトとしては，ソーシャルネットワーキングサイト（SNS）や掲示板が考えられる．これらのサイトを巡回し，他の一般ユーザが投稿した多数の文章に目を通すことで，一般ユーザが有名人に対して抱いている印象や感想を知ることができる．ただし，そのためには膨大な文章を読む必要があり，自力で一般ユーザの観点に基づく人物情報を得ようと思うと多大な時間と労力が必要になる．そこで本研究では，SNS における一般ユーザの有名人に関する発言から自動で一般ユーザの観点に基づく人物情報を抽出し，これを可視化して提示することを目指す．

今回我々は，マイクロブログの 1 つで，一般ユーザの発言が豊富に存在する Twitter を対象に一般ユーザの発言を収集する．Twitter では投稿の単位となる 1 つの発言は tweet と呼ばれ，日本国内において日平均で約 8 千万の tweet が投稿されている [14]．tweet はそのユーザの日常に関する内容であることが一般的だが，他にもその時々で話題になっているニュースや有名人に関する内容の tweet も多く見られる．このため，有名人に関する一般ユーザの発言を得るためのデータソースとしても Twitter は適している．

本研究では，一般ユーザの観点に基づく人物情報の 1 つとして，一般ユーザ目線での有名人間の類似度の抽出に焦点を置く．我々は，A. 有名人に関して言及している tweet，B. 有名人に関して言及している tweet を行ったユーザのプロフィール の 2 点に着目して，それぞれの有名人間の類似度を計算する．また比較対象として，一般ユーザの観点

[1] 大阪大学大学院 基礎工学研究科
[2] 三菱電機株式会社 先端技術総合研究所

が反映されない C. 有名人本人の tweet を用いる．類似度の元となる特徴量の算出には Latent Dirichlet Allocation（LDA）[1] を利用し，有名人毎のトピック分布を求める．そして，推定したトピック分布間の類似度求め，多次元尺度構成法により 2 次元平面上に可視化して提示するシステムを提案する．

2. 関連研究

Twitter 上の tweet を利用して情報抽出や意見要約を行う研究には様々なものがある．中でも，商品やテレビ番組などのアイテムに対する評判情報を分析し，マーケティングやその他のアプリケーションに応用しようという研究は多く見られる [10], [12]．一方，アイテムだけでなく人物に対する評判情報を扱った研究も近年増えてきている．Meng らは，あるエンティティ（アイテムや人物などの総称）に関する Twitter 上での意見要約を行った [5]．彼らは，あらゆるエンティティに関しての tweet 群から，ハッシュタグに注目してクラスタリングにより N 個のトピックを見つけ，それぞれのトピック毎に極性も考慮した要約を行った．また，Twitter 以外の SNS の情報を利用した研究には次のようなものもある．Park らは，実名公開型の SNS である Cyworld における政治家への一般ユーザのコメントから，セマンティックネットワーク解析および感情分析を用いてその政治家への集団感情を分析した [7]．Guy らは，社内 SNS における 9 つのソーシャルメディアから他者への評価を元にユーザ間の類似度を求めている [4]．

評判情報の対象ではなく，その発信源となったユーザの特徴を調べることも興味深いことである．古賀らは Twitter において，ある対象ユーザのリツイートやフォロー関係などを文書として LDA を行い，ユーザの興味や嗜好に関する潜在トピックに注目し，ユーザ間の類似度を測りユーザ推薦を行った [6]．藤本らは，ユーザの Web 閲覧行動には潜在的なトピックが存在すると仮定し，サーバに残ったアクセスログの URL を階層型 URL 辞書で集約した後 LDA を行い，可視化により抽出したユーザ層を提示した [9]．

人物関係の情報を提示するための可視化技術に注目すると，松尾ら [8] の研究がある．彼らは学会における人間関係を自動的に抽出し，その関係を表すネットワークを可視化することで複雑な関係や研究者のクラスタを直感的に分かるように提示している．

本研究は，一般ユーザの対象人物に関する発言だけでなく，その発言を行ったユーザの特徴（すなわちプロフィール文）に注目した分析も行っている．

3. 提案システム

我々は Twitter 上の情報から，一般ユーザの観点に基づく人物情報を抽出し，その結果を可視化して提示するシステムを提案する．このシステムは大きく 3 つのモジュールから構成されている．

1. 前処理：tweet 中に含まれるノイズに対する処理および形態素解析を行う．
2. 特徴量算出：人物毎のトピックに基づく特徴量を求める．
3. 可視化：人物間の類似関係を 2 次元平面上に可視化する．

3.1 情報源選択

人物情報の情報源として何を用いるかによって，人物に対して注目する観点が変わってくると思われる．我々は一般ユーザの観点に基づく情報源として次の A と B を，比較対象として C を用いる．

A. 有名人に関して言及している tweet

対象有名人 i の名前を含む tweet 群に対して，前処理を行って得られた単語素性集合を文書 d_i，文書 d_i の集合を D とする．この入力データは，一般ユーザの対象有名人に対する率直な発言であることが多い．そのため，一般ユーザの対象有名人に対する印象・感想という観点で有名人間の比較を行うことができる．

B. 有名人に関して言及したユーザのプロフィール

対象有名人 i の名前を含む tweet（上記情報源 A）を行ったユーザの集合を作成する．この集合中の各ユーザのプロフィール文[*1]に対して，前処理を行って得られた単語素性集合を文書 d_i，文書 d_i の集合を D とする．この入力データは，その tweet を行ったユーザの社会的情報や趣味・興味に注目している．そのため，対象有名人に関心を持っているユーザ層の違いという観点で有名人間の比較を行うことができる．

C. 有名人本人の tweet

対象有名人 i が発信した tweet に対して，前処理を行って得られた単語素性集合を文書 d_i，文書 d_i の集合を D とする．これは対象有名人が自ら発した tweet であるため，対象有名人がどう見られたいかを反映していると思われる．そのため，一般ユーザの観点は反映されていないといえる．

3.2 前処理

収集した tweet には，一般ユーザの有名人に関する発言以外に，外部のサイトの宣伝や blog の記事名が含まれた文章（以降，ノイズテキスト）が存在する．これらが多く含まれてしまうと一般ユーザが発信する印象・感想に関する情報の抽出が妨げられる．上記のようなノイズテキストは，URL を伴いかつ複数の tweet により参照されることが多い．そこで，URL を伴う tweet に注目し，2 つ以上の tweet に共通して出現するテキスト部分を発見し，除去することにする．以下に具体的な除去方法を示す．

処理内容

[*1] ユーザが自由に記入できる 160 文字以内の自己紹介文.

(1) URL を含む tweet の URL より前のテキスト全てと後ろのテキスト全てを抽出し，ノイズテキスト候補集合（集合 A）を作成する．また，空の集合 R を用意する．
(2) 集合 A から文字列（要素 e）を 1 つ選択し，その文字列の中心 N 文字を切り取る（文字列の中心位置から，N/2 文字分手前の文字列と，N/2 文字分後ろの文字列を抽出する）．ただし，文字数が N 文字未満の場合は，その要素 e は棄却されて次の要素の判定へ移る．なお，N の値は一般語を抽出してしまわない範囲で小さい方が好ましく，4 章の実験では 6 に設定している．
(3) 要素 e を除く集合 A 中から，(2) で切り取った文字列を文中に含むものを探す．存在する場合，そのマッチした要素の集合 (集合 B) を作りノイズテキストの同定 (4) へ進む．存在しない場合，この要素 e を棄却し次の要素の判定 (2) へ移る．
(4) ノイズテキストの同定を行う．まずはノイズテキストの左端を見つける．切り取る範囲を要素 e の中心 N 文字から左側に広げ，集合 B 中の各要素と部分的にマッチするか調べる．この操作を集合 B 中のどの要素ともマッチしなくなるまで続け，ノイズテキストの左端を特定する．同様に，右端も探索する．
(5) 特定したノイズテキストを集合 R に追加し，集合 A の次の要素の判定 (2) へ移る．すべての要素に対して判定を終えた場合，(6) へ進む．
(6) 集合 R の要素を文字数の多い順番にソートし，オリジナルの tweet 集合中の tweet において各要素とマッチする文字列を除去する．この除去処理を行った tweet 集合を以降用いる．

上記の処理とは別に全ての tweet から，URL とハッシュタグの除去も行う．また，URL を含んでいなくとも tweet のテキストが長文で完全一致するものが複数ある場合，これらはスパムであることが多いため，tweet の文字数が M 文字以上[*2]でテキストが完全一致するものは除外する．

上記ノイズ処理を情報源 A,C に対して行う．その後，各情報源のテキストに対して MeCab[13] を利用して形態素解析を行い，テキストを単語単位に分解する．そして，有名人毎に単語の品詞が名詞・形容詞・動詞と判定されたものを bug-of-words 表現で表し，これを文書 d_i とする．この d_i の集合を文書集合 D として以降用いる．

3.3 特徴量算出

前節で作成した文書集合 D を入力とし，各人物（文書 d_i と対応している．以降，添字の i を省略して d と表す）のトピックに基づく特徴量を求める．文書毎の特徴量を求める手法としては，単語の出現頻度と稀少度を考慮した TF-IDF を算出する方法が一般的によく用いられている．ただしこの方法では，文書中の完全一致する単語しか考慮されておらず，類似した意味の語の影響を反映させることができない．そこで，文書の背後に存在するトピックを考慮したソフトクラスタリングを行う手法であるトピックモデルを用いる．我々はトピックモデルとして，近年注目されている Latent Dirichlet Allocation（LDA）を使用し，文書毎のトピック分布を推定する．

3.3.1 LDA

LDA は，Blei らにより提案された確率的トピックモデルである [1]．トピックモデルとは，ある文書（N_d 個の単語からなるトークン列）$\mathbf{w^d} = (w_1, w_2, \ldots, w_{N_d}) \in D$ が単一または複数のトピックに属する単語から構成されているという仮定をおき，その文書を構成するトピックの比率（トピック分布）$\Theta^d = (\theta_1, \theta_2, \ldots, \theta_{|T|})$ と，それらのトピック毎の単語生成確率（単語分布）$\Phi^t = (\phi_1, \phi_2, \ldots, \phi_{|W|})$ に基づき確率的に文書を生成するモデルである．ただし，文書集合を D，各文書のトークン数を N_d，全文書中に出現する全語彙の集合を W，トピックの集合を T とする．以下に LDA の生成過程を示す．

1. トピック毎に，ディリクレ分布 $Dir(\beta)$ から Φ^t を生成．
2. 文書毎に，
   a. ディリクレ分布 $Dir(\alpha)$ から Θ^d を生成．
   b. 文書中の各トークン毎に，
      i. 多項分布 $Mult(\Theta^d)$ から，トークン w_i^d にトピック z_i^d を付与．
      ii. 多項分布 $Mult(\Phi^{z_i^d})$ から，トークンとなる新たな単語を生成．

入力として与える必要があるのは，文書集合 D，トピック数 $|T|$，反復回数，Θ と Φ の事前分布であるディリクレ分布のハイパーパラメータ α と β である．ここで，α と β は一般にベクトルであるが，Griffiths に従いすべての要素を同じ値に設定する [2]．なお，我々は与えられた文書集合 D からパラメータ Θ，Φ の推定を行うために，Griffiths らの Gibbs Sampling による手法 [2] を用いた．

3.4 可視化

前節で求めた人物毎の特徴量を用いて，人物間の類似関係を 2 次元平面上に可視化する．可視化手法として，距離尺度を表現する一般的な手法である多次元尺度構成法（MDS）を用いて結果の比較を行う．ただし，MDS には特徴量である確率分布をそのまま用いることはできないので，分布間の非類似度を表す行列を算出し，これを用いる．

3.4.1 JS 情報量

確率分布の類似度を測る手法としては，比較対象の一方の確率分布の値に 0 が存在しないことを制約条件に持

[*2] 我々は tweet の文字数上限の 3 割（42 文字）以上に設定した

春名風花	宮迫博之	松本人志	スギちゃん	山本太郎	ローラ	剛力彩芽	西川貴教	GACKT	きゃりーぱみゅぱみゅ
鬼龍院翔	喜矢武豊	宇治原史規	伊集院光	篠田麻里子	平野綾	中川翔子	安倍晋三	東国原英夫	太田順也（ZUN）
橋下徹	堀江貴文	孫正義	香川真司	ダルビッシュ有	東浩紀	乙武洋匡	茂木健一郎	田原総一朗	前山田健一（ヒャダイン）

表 1 対象有名人一覧

つ Kullback-Leibler divergence（KL 情報量）や，2 つの平均的な確率分布までの KL 情報量の平均を求める Jensen-Shannon divergence（JS 情報量）などがある．本研究では，比較の対称性を満たす JS 情報量を用いて類似度の比較を行う．ここで，対象 A と B のトピック分布 Θ^A, Θ^B が与えられたとき，JS 情報量は次のようにして計算される．

$$D_{JS}(A,B) \;=\; \frac{1}{2}\left(\sum_{k=1}^{|T|}\theta^A_{t_k}\log\frac{\theta^A_{t_k}}{\theta^R_{t_k}}+\sum_{k=1}^{|T|}\theta^B_{t_k}\log\frac{\theta^B_{t_k}}{\theta^R_{t_k}}\right)$$

ただし，$\theta^R_{t_k}=\frac{1}{2}\left(\theta^A_{t_k}+\theta^B_{t_k}\right)$ である．この D_{JS} は，値が小さいほど対象間の類似度が高いことを意味する．

3.4.2 多次元尺度構成法

多次元尺度構成法（MDS）は，2 者間の距離尺度を維持した配置を求める方法であり，主成分分析とは異なり配置した空間における軸には意味的な解釈が存在しない．MDS は入力に要素間の距離行列（非類似度行列）が必要であり，3.4.1 により求めた行列を入力とする．また，今回は最も基本的な古典的多次元尺度構成法と呼ばれる手法を用いる．

4. 実験

3 種類の情報源に対する可視化結果の特徴の分析を行う．分析する特徴としては，提示結果がある程度事実に則しており一般的に納得できるものであるか（妥当性），人物間の意外な関係や集団を見つけられるか（発見性）という 2 点に着目する．なお，本実験では可視化結果のスクリーンショットを示し，その中で発見されたグループ（密に集まった有名人集合）に注目するが，そのグループはスクリーンショット上で四角で囲んで，また (a)〜(e) の記号を振って示すことにする．各グループを本文中で参照するときには，(a) のように表記する．

4.1 データセット・パラメータ

データセット

Twitter 公式 API を利用して，表 1 の対象有名人（合計 30 人）に対してそれぞれ情報源 A〜C のデータを収集した．情報源 A は，対象有名人の名称・呼称[*3]をクエリとして検索で得られた 2013 年 9 月 2 日から 8 日までの 1 週間分の tweet である．情報源 B は，情報源 A に含まれる tweet を行った各ユーザの 2013 年 9 月時点でのプロフィールである．情報源 C は，対象有名人の 2013 年 9 月時点から上限数の 3200 まで遡った tweet である．

[*3] Wikipedia に登録された名称に加え，Wikipedia の本人記事中に記載された愛称・略称も用いる

パラメータ

LDA のトピック数 $|T|$ は，[11] の手法により形成されたクリーク数である 17 と設定した．LDA のハイパーパラメータは $\alpha=50/|T|$，$\beta=0.1$，反復回数は 1000 とした．

4.2 情報源毎の結果

情報源 A は対象人物に対する発言内容のトピックに注目しており，このため時事ニュースや人物に対する現在の世論の影響を強く受ける傾向がある．実際に今回得られたトピックは，表 2 に示すようなものが見られた．トピックの種類としては，好意的・批判的といったポジティブ・ネガティブを表すものや，アイドル・俳優・お笑い芸人・政治・ゆるキャラなど話題に関連するグループ，時事ニュースがトピックとして見られた．妥当性に関しては，図 1 の下部（図 1-(d)）に政治家・評論家・実業家が集まっており，中央やや上部（図 1-(b)）にスポーツ選手，上部（図 1-(a)）に俳優・アイドル・芸人を中心とする芸能人が集まる結果となっている．(a) の部分に関して，一部に芸能人ではないネット上の有名人が含まれているものの，この配置は多くのユーザにも理解されるのではないかと思われる．発見性に関しては，中央部（図 1-(c)）にバラエティ番組にも出演するが世間からは知識人と捉えられている知識系芸能人が集まっており，興味深いといえる．

情報源 B は対象人物に言及したユーザのプロフィールであり，これを用いれば同じユーザ層に関心を持たれている人物同士が近くに配置されることが期待される．このユーザ層はトピックとして抽出されており，これを表 3 に示す．トピックの種類としては，主にユーザが興味のある対象が得られている他，学生や BOT のようなユーザ層そのものも得られている．妥当性に関しては，図 2 の上部（図 2-(a)）に政治家・評論家・実業家が集まっており，中央部（図 2-(c)）に芸能人が集まる結果となっている．情報源 A では周囲から分離していたスポーツ選手の集まりは，はっきりと分離していないものの互いに近い位置に配置されている．これらの配置は多くのユーザにも理解されるのではないかと思われる．発見性に関しては，情報源 A と同様に中央やや右上（図 2-(b)）に，知識系芸能人が集まっている．また，中央やや下部（図 2-(d)）にはネット上で有名な人物の集まりが表れている点も興味深い．最下部（図 2-(e)）には V 系（ヴィジュアル系）バンドであるゴールデンボンバーの二人が他の芸能人や音楽関係者とは大きく隔離した位置にあり，彼らに関心を持つユーザは特異であることが分かる．加えて，情報源 B の特徴として，四角で囲

トピック解釈	好意的	批判的	男性有名人	AKB	ゆるキャラ	政治（政府）	政治 (大阪維新)	映画「ガッチャマン」
上位単語	てる	はるかぜちゃん	山田孝之	卒業	熊本	安倍総理	橋本	誕生日
	する	菜々緒	城田優	似る	可愛い	いる	蓮舫	剛力
	好き	大人	真島ヒロ	板野友美	(笑)	日本	市長	可愛い
	見る	叩く	映画	笑	くまモン	総理	維新	綾野剛
	歌う	匿名	かっこいい	AKB	ふる	麻生太郎	批判	ゴーリキー
	可愛い	批判	イケメン	篠田麻里子	グッズ	首相	大阪	実写化

表 2 主なトピック（情報源 A：有名人に関して言及している tweet）

トピック解釈	ジャニ系	アイドル	V 系	アニメ・ゲーム	政治・経済	スポーツ	学生	BOT
上位単語	嵐	AKB	金爆	アニメ	日本	サッカー	好き	紹介
	NEWS	推す	ゴールデンボンバー	好き	政治	野球	フォロー	名言
	組	前田敦子	V 系	ゲーム	反対	香川真司	音楽	面白い
	二宮	篠田麻里子	ギルド	ボカロ	脱原発	選手	気軽	よろしくお願いします
	相葉	SKE48	シド	漫画	TPP	MLB	野球	ニュース
	潤	NMB48	己龍	声優	日本人	SAMURAI	高校	2ch

表 3 主なトピック（情報源 B：有名人に関して言及したユーザのプロフィール）

んだ人物が集中している部分が情報源 A がよりも密集している傾向が見られる．例えば，図 2-(a) の政治家・評論家・実業家の塊や，図 2-(b) の知識系芸能人の塊は，情報源 A よりも密集していることが分かる．図 2-(d) のネット上で有名な人物の集合も密な配置となっている．このように，情報源 B は情報源 A よりも集団がより密になる傾向があり，グループの発見が容易になる利点があるように思われる．さらに，得られた集団の種類は情報源 A に比べて豊富であり，より多くの発見が可能であることを示している．

情報源 C は対象人物本人の tweet に注目しており，一般ユーザの観点は反映されていない．トピックの種類としては，政治・モデル撮影といった有名人本人の職業に関するトピックの他，ライブイベントの告知やブログ更新の告知といったトピックが見られたが，それ以外のトピックは解釈できないものが多い結果となった．得られた可視化結果は図 3 のようになっている．情報源 A,B と同様に，政治家・実業家（図 3-(a)），芸能人（図 3-(c)）で分かれる傾向が現れた．また，スポーツ選手や V 系の人物同士も近くに配置されている．ただし，外れ値と思われるものもいくつかあり，剛力彩芽（tweet 数が極端に少ないため）や中川翔子（アニメ，ゲームの話題が多いため）などが目立って現れた．加えて，(c) の部分を詳しく見ると，ネット上の有名人や伊集院光・松本人志といった知識系芸能人（図 3-(b)）に含まれた方がより良い人物まで含まれており，情報源 A,B と比べて妥当性はやや低い結果となっている．発見性に関しては，中央部（図 3-(b)）に作家・評論家と知識系芸能人の集まりが見られ，情報源 A と同等といえる．

以上の結果をまとめたものを表 4 に示す．これまで述べてきたとおり，有名人の塊が意味のある分け方になっているかどうかを表す妥当性に関して，情報源 A と情報源 B の両方で優れた結果を得られていると判断する． 興味深い新たな塊を見つけることができたかどうかを表す発見性に関しては，情報源 B が最も優れており，情報源 A は情報源 B に比べるとやや劣る結果となった．一方，情報源 C は情報源 A と比較して妥当性に関して劣る結果となった．また，情報源 B においては集団がより密になる傾向があるため，グループの見つけやすさという点でも優れている．

5. おわりに

本稿では，Twitter から得られる情報を元に，一般ユーザの観点に着目した人物間の類似関係の可視化を行うことを目指した．我々はこの目的を達成するために，多くのノイズが含まれたテキストに対して前処理を行い，各情報源から得られるトピックに基づいた人物毎の特徴量を算出し，これを元に 2 次元平面上での配置を決定した．また，可視化結果に対して使用した情報源毎に妥当性と発見性に着目して特徴の分析を行った．分析の結果，我々が用いた一般ユーザの観点に基づく情報源は事実に基づく妥当性を保ちつつ，知識人として捉えられている芸能人やネット上で人気の有名人といった容易に思いつかない発見も可視化結果に反映させられることを示した．また，情報源の中では，有名人の名前を含んだ tweet の著者のプロフィール文が最も妥当性が高く，なおかつ多くの発見が行えることが分かった．今後は，人物同士だけでなく，人物とアイテムといった異なるクラス間の一般ユーザの観点に基づく関連性の抽出を行う研究や，それらの関連性を元に推薦を行うシステムの作成を検討していく予定である．

謝辞 この研究は，三菱電機株式会社先端技術総合研究所との共同研究である．

参考文献

[1] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, The Journal of Machine Learning Research, Vol.3, pp.993-1022 (2003)
[2] Griffiths, T.L. and Steyvers, M.: Finding Scientific Top-

情報源	A	B	C
主なトピック	話題内容, 感情極性	ユーザの趣味・興味	有名人の仕事内容, 広報
妥当性	○	○	△
発見性	△	○	△

表 4 情報源毎の結果まとめ

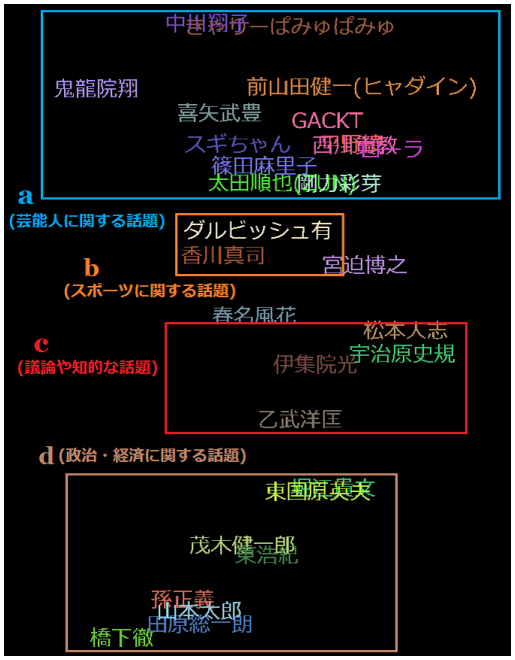

図 1 情報源 A（一般ユーザの発言内容）

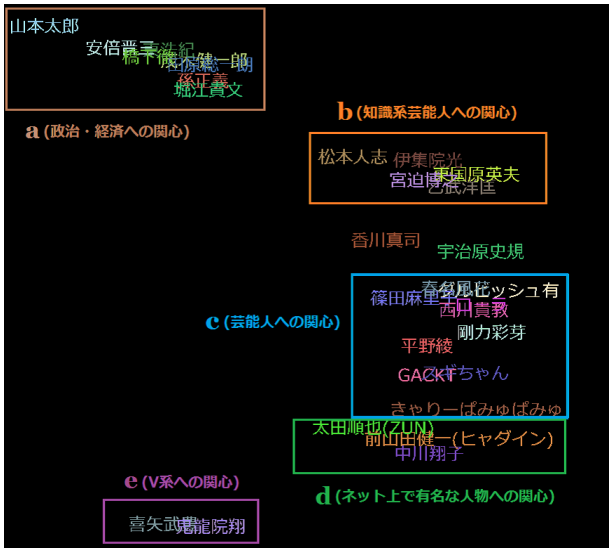

図 2 情報源 B（関心を持つユーザ層）

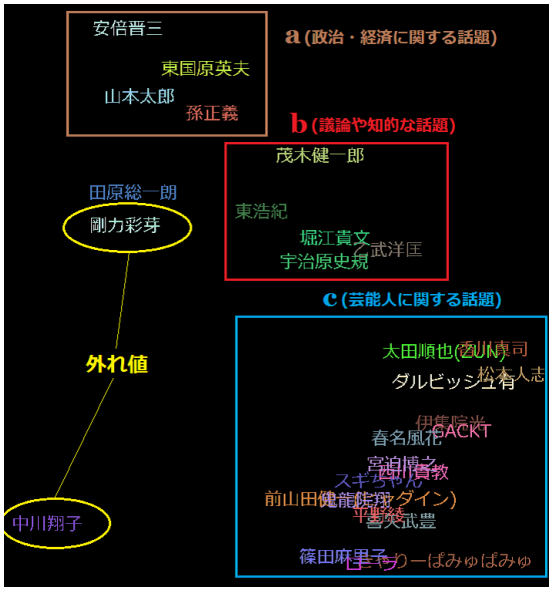

図 3 情報源 C（有名人本人の発言）

ics, Proc. of National Academy of Sciences, Vol.101, pp.5228-5235 (2004)
[3] Hofmann, T.: Probabilistic Latent Semantic Indexing, Proc. of ACM SIGIR, pp.50-57 (1999)
[4] Guy, I., Jacovi, M., Perer, A., Ronen, I. and Uziel, E.: Same Places, Same Things, Same People? Mining User Similarity on Social Media, Proc. of ACM CSCW, pp.41-50 (2010)
[5] Meng, X., Wei, F., Liu, X. and Zhou, M.: Entity-centric topic-oriented opinion summarization in twitter, Proc. of ACM SIGKDD, pp.379-387 (2012)
[6] Koga, H. and Taniguchi, T.: Developing a user recommendation engine on twitter using estimated latent topics, Proc. of HCII (the 14th international conference on Human-computer interaction), pp.461-470 (2011)
[7] Park, S.J., Lim, Y.S., Sams, S., Sang, M.N. and Park, H.W.: Networked Politics on Cyworld: The Text and Sentiment of Korean Political Profiles, The Journal of Social Science Computer Review, Vol.29, No.3, pp.288-299 (2011)
[8] 松尾豊, 友部博教, 橋田浩一, 中島秀之, 石塚満: Web 上の情報からの人間関係ネットワークの抽出, 人工知能学会論文誌, Vol.20, No.1, pp.46-56, (2005)
[9] 藤本拓, 秋永和計, 栄藤稔: 潜在トピックモデルを利用したユーザプロファイリング技術, NTT DoCoMo テクニカル・ジャーナル, Vol.19, No.3, pp.37-41 (2011)
[10] 池田和史, 服部元, 松本一則, 小野智弘, 東野輝夫: マーケット分析のための Twitter 投稿者プロフィール推定手法, 情報処理学会論文誌コンシューマ・デバイス＆システム, Vol.2, No.1, pp.82-93 (2012)
[11] 芹澤翠, 小林一郎: 潜在的ディリクレ配分法に基づくトピック類似度を考慮したトピック追跡, データ工学と情報マネジメントに関するフォーラム (2011)
[12] 山本祐輔, 浅井洋樹, 上田高徳, 秋岡明香, 山名早人: テレビ番組に対する意見をもつ Twitter ユーザのリアルタイム検出, データ工学と情報マネジメントに関するフォーラム (2013)
[13] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, 入手先 ⟨http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html⟩, (参照 2013-02-18)
[14] ビッグローブ: 5 月の Twitter 利用動向を発表 ⟨https://www.biglobe.co.jp/pressroom/release/2014/06/140609-a⟩, (参照 2014-06-09)