Vol.2019-DBS-169 No.4
Vol.2019-IFAT-136 No.4
2019/9/10

Wikipediaのカテゴリ情報を用いたツイート発信者の特徴表現

中谷 将佳史[1,a)]　湯本 高行[1,b)]　礒川 悌次郎[1]　上浦 尚武[1]

概要：近年，SNS の普及に伴い，情報を素早く発信，入手することが可能となっている．代表的な SNS の一つである Twitter では，リツイート機能などにより，拡散されている情報を容易に得ることができる一方で，個々の発信者に関する情報を得る機能は提供されておらず，個別にツイートをさかのぼる必要がある．このため，利用者が他の発信者について，その特徴を素早く判断することができないといえる．そこで，発信者別のツイート集合から，発信者の特徴表現を作成する方法を提案する．特徴表現の作成にあたって，ツイート中に出現する名詞に注目する．名詞が表すエンティティを Wikipedia の記事に紐付け，記事毎のカテゴリ情報を集計し，ツイートの特徴表現を作成する．Wikipedia の記事やカテゴリ情報を用いることで，大規模なツイート収集を行わず，最新の語に対応した表現の作成が可能である．次に，発信者のすべてのツイート表現を集計し，発信者の特徴表現を行う．

1. はじめに

近年，Twitter をはじめとする SNS の普及に伴い，様々な情報を素早く発信したり，入手することが可能となり，様々な角度からその特徴に関する研究がなされている [1], [2]．SNS が既存の情報源と異なる点として，情報の受信者が次の情報の発信者となりうることが挙げられる．様々な最新の情報を多くの人間に素早く伝えられるという利点があり，災害時においてこうした特性を活用する研究なども行われている [3]．その一方で，誤った情報や不確実な情報が伝播しやすいという問題を抱えている [4]．実際に誤った情報が伝播した例として，災害が発生した際に「被災地域にライフラインの断絶が起こる」というツイートが拡散した，などが挙げられる．

SNS においては，各ユーザが今までどのような情報を発信してきたのかを一目で確認できるような機能が提供されていない．例として Twitter を取り上げると，あるユーザがどのような情報を発信してきたか知りたい場合，過去のツイートを直接見て判断する必要がある．このため，情報を受け取ったユーザが情報の真偽を確かめたり，拡散を行うか否かは，各ユーザーの受け取った情報そのものに関する知識や，発信者に関する事前知識によって判断されているといえる．

そこで，本研究では，ユーザが情報を拡散する際の判断基準の一つとなるような，各ユーザの特徴表現の作成を目的とし，これまでに発信された情報に基づき，発信者の特徴を分析，可視化して提示する手法について述べる．

[1] 兵庫県立大学
[a)] ei19a016@steng.u-hyogo.ac.jp
[b)] yumoto@eng.u-hyogo.ac.jp

2. 関連研究

関連研究について以下に述べる．Wikipedia のデータを言語リソースとして活用する研究では，いくつかのタスクについて既存の手法を上回る精度を発揮している [5], [6], [7]．

ツイートを対象とした話題推定の手法では，LDA[8] などのトピックモデルを適用する研究が主流である [9], [10]．本研究の特徴として，トピックモデルを用いず，Wikipedia のカテゴリ情報に基づいて特徴表現を行っている．特徴分析の際に Wikipedia のカテゴリ情報を用いることで，大規模なデータセットを用意する必要がなく，新語や専門用語等を含めた幅広いトピックに対応するとともに，ツイートのように比較的短い文からなり，文量や語の頻度に関する情報が十分でないテキストデータの解析を可能にしている．

3. Wikipediaのカテゴリ情報による特徴表現

3.1 概要

本節では，与えられた文書から Wikipedia のカテゴリ情報に基づく特徴表現を作成する手法について述べる．Wikipedia には，各記事に対し分野別にまとめた索引（カテゴリ）が付与されている．各記事のカテゴリを参照することで，関連する記事やカテゴリについての情報を得ることが可能である．提案手法では，各記事からさかのぼって

到達可能なカテゴリ集合を記事毎の特徴表現とする．

Wikipedia のカテゴリ情報に基づく特徴表現の作成手法についての概要を述べる．本手法では，ツイートのようにごく短い文章からなり，語の頻度に関する情報が十分でない文書を解析することを目的とする．文書の特徴を表現する手法として，文書に特有の単語の傾向から判断するなどの方法が考えられる．しかし，発信者別のツイート集合のように，それぞれの文書に十分な数の語が含まれていない場合，既存の手法が有効に機能しない場合が多い [10]．そこで，文中の名詞と対応付けた Wikipedia の記事を多数のカテゴリ情報へと変換し，頻度の情報を補うとともに，適当なカテゴリを選択することで，文書の特徴表現を行う．具体的な処理の流れを**図 1** に示す．図の矢印に付加されている 1 対 1（多）のラベルは，1 つの参照元オブジェクトから単独（複数）の参照先オブジェクトが得られることを示す．例として，1 ツイートから 5 個の名詞が取得される，などの場合が考えられる．

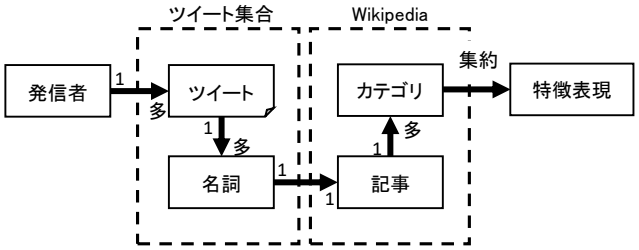

図 1 提案手法の流れ

まず，文単位での名詞抽出を行い，抽出された名詞に対応する Wikipedia の記事を決定し，作成した特徴表現を割り当て，単語別の特徴表現を作成する．次に，文中に出現する単語間におけるカテゴリ情報の重なりから，文の話題を表すカテゴリ情報を抽出し，それぞれの文について特徴表現を作成する．最後に，各文の特徴表現の重なりを用いて，文書全体の特徴表現を作成する．

3.2 テキストの前処理と名詞抽出

形態素解析エンジン MeCab[*1] と，最新語を採録している辞書 NEologd[*2]，また独自に作成した Wikipedia の記事タイトルからなるユーザ辞書により，各文から名詞の抽出を行う．このとき，以下の単語は抽出の対象外とする．

- 「名詞」のうち，「サ変接続」，「副詞可能」，「接尾」，「形容動詞語幹」，「ナイ形容詞語幹」，「非自立」，「数」，「接続詞的」，「特殊」，「代名詞」に該当するもの

3.3 名詞別の特徴表現

抽出された名詞に対し，対応記事からさかのぼって得た Wikipedia のカテゴリ情報を整理し，特徴表現として付与する処理の流れを以下に示す．以下では，記事からさかのぼって得られる Wikipedia カテゴリ構造の一部を指して周辺カテゴリという．

3.3.1 周辺カテゴリの取得

抽出された名詞 w に対して，名詞と Wikipedia の記事名とが完全一致する記事 p を割り当てる処理について示す．このとき，周辺カテゴリを正確に得るため，参照された記事がリダイレクトされている場合は，リダイレクト先を一致する記事 p として割り当てる．

次に，得られた記事 p から h ホップさかのぼって得られる周辺カテゴリ集合 C を，$C = \{c_1, c_2, ..., c_{|C|}\}$ とすると，周辺カテゴリ間のリンク構造は，

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1|C|} \\ a_{21} & a_{22} & \cdots & a_{2|C|} \\ \vdots & \vdots & \ddots & \vdots \\ a_{|C|1} & a_{|C|2} & \cdots & a_{|C||C|} \end{pmatrix} \quad (1)$$

なる隣接行列 A で表される．ここで，A の行が下位カテゴリ，列が上位カテゴリに対応し，c_i から $c_j (1 \leq i \leq |C|, 1 \leq j \leq |C|)$ へのリンクが存在する (c_i から見て c_j が上位カテゴリである) とき，$a_{ij} = 1$，存在しない場合は $a_{ij} = 0$ とする．また，C に含まれるカテゴリを決定する過程で，カテゴリ間のリンクがループする場合はそのリンクを無視して集計している．

3.3.2 周辺カテゴリ構造の簡略化

Wikipedia のカテゴリは複雑なネットワーク構造をもつ [5]．このため，語句の属するカテゴリを判別することが困難である．そこで，各カテゴリと記事との間の最大距離に基づき，周辺カテゴリ間のリンク構造を整理し簡略化を行う処理について示す．C に含まれる各カテゴリを，記事からの最大ホップ数を用いて以下のように分類する．

$$C_i = \{c | c \in C \text{ のうち記事 } p \text{ から見た最大ホップ数が } i\} \quad (2)$$

ここで，$1 \leq i \leq h$ である．(2) 式により，周辺カテゴリ集合 C は，

$$|C| = \sum_{i=1}^{h} |C_i| \quad (3)$$

を満たす，互いに素な集合 C_1 から C_h に分割できる．

C_1 から C_h を用いて，隣接行列 A を次のように区分行列に区分けする．

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1h} \\ A_{21} & A_{22} & \cdots & A_{2h} \\ \vdots & \vdots & \ddots & \vdots \\ A_{h1} & A_{h2} & \cdots & A_{hh} \end{pmatrix} \quad (4)$$

それぞれの小行列 A_{kl} は，$|C_k| \times |C_l|$ 行列であり，C_k を下位カテゴリの集合，C_l を上位カテゴリの集合 (ただし，

*1 https://taku910.github.io/mecab/
*2 https://github.com/neologd/mecab-ipadic-neologd

$1 \leq k \leq h$，$1 \leq l \leq h$) とした場合の各カテゴリ間のリンク構造を表す．

次に，周辺カテゴリ間のリンク構造の簡略化を行う．記事 p から見た各カテゴリへの最大ホップ数を，各カテゴリの抽象度と考える．抽象度の低いカテゴリが高いカテゴリの上位に現れる経路を以下のように削除する．$k \geq l$ を満たす A_{kl} について，

$$A_{kl} = 0_{|C_k|,|C_l|} \tag{5}$$

とする．ただし，$0_{|C_k|,|C_l|}$ は $|C_k| \times |C_l|$ の零行列を表す．

(5) 式によって簡略化した隣接行列 A' は，

$$A' = \begin{pmatrix} 0 & A_{12} & A_{13} & \dots & A_{1h} \\ 0 & 0 & A_{23} & \dots & A_{2h} \\ 0 & 0 & 0 & \dots & A_{3h} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \tag{6}$$

と表せる．$h = 5$ のとき，簡略化した周辺カテゴリ間のリンク構造は**図 2** のようになる．

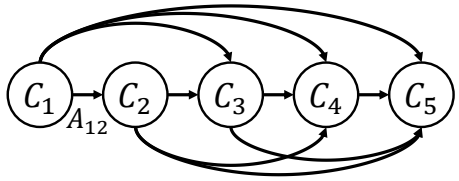

図 2 簡略化された隣接行列の表すグラフ

3.3.3 カテゴリスコアの付与

簡略化した周辺カテゴリ構造を用いて，各カテゴリの重要度を表すスコアを決定する．より記事に近く，多くのカテゴリに紐付いているカテゴリに高いスコアを与える．すべての周辺カテゴリにスコアを与えておき，最終的な特徴表現に用いるカテゴリの絞り込みは後に行う．各カテゴリについて，下位カテゴリのスコアを上位カテゴリのスコアの流入によって決定する．C_h に属するカテゴリ c_i（ただし，$1 \leq i \leq |C_h|$) にそれぞれ対応するスコアを s_i とする．各カテゴリに対し，スコアの初期値として以下のようにスコア行列 S_h を与える．

$$S_h = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{|C_h|} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{7}$$

(7) 式を用いて，C_{h-1} から C_1 のそれぞれに対応するスコア行列 S_i を，

$$S_i = \sum_{j=i+1}^{h} A_{ij} S_j \tag{8}$$

として計算する．また，記事 p から得られる周辺カテゴリ集合 C の各カテゴリに対応するスコア行列 S を，

$$S = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{|C|} \end{pmatrix} \tag{9}$$

とする．ここで，1 つの記事が持つスコアの合計が 1 になるように，

$$S' = \begin{pmatrix} s'_1 \\ s'_2 \\ \vdots \\ s'_{|C|} \end{pmatrix} = \frac{S}{\sum_{i=1}^{|C|} s_i} \tag{10}$$

として正規化する．

3.4 文単位での特徴表現

作成した単語別の特徴表現を文単位で集約し，文ごとの特徴表現を得る処理について以下に示す．ある文 t から名詞集合 W，記事集合 P_t が得られるとき，それぞれの記事からみたカテゴリについて，(10) 式より，スコア行列としてそれぞれ $S'(p_1), S'(p_2), ..., S'(p_{|P_t|})$ が得られる．ここで，記事 p から得られる周辺カテゴリ集合を $C(p)$ とし，$C(p)$ に対応するスコア行列を $S'(p)$ とする．

ある記事 p にあたる名詞 w が文中に存在するとき，記事 p から得られる周辺カテゴリ情報の $C(p)$ には不要なものが含まれている場合がある．例として，記事「リンゴ」からはカテゴリ「果物」,「バラ科」,「青森県の象徴」などが得られるが，「リンゴ」が出現する場合に必ずしも「バラ科」や「青森県の象徴」といった意味合いを含んでいるとはいえない．そこで，同じ文中に出現する他の名詞を用いて，必要なカテゴリの絞り込みを行う．

文 t 内において有効なカテゴリは，少なくとも 2 つの記事に含まれるカテゴリとし，次のように定義する．2 つの記事 p_i，$p_j (1 \leq i < j \leq |P_t|)$ について，その積集合 $C(p_i) \cap C(p_j)$ を用いて，有効なカテゴリの集合 $C(t)$ を

$$\begin{aligned} C(t) &= \bigcup_{1 \leq i < j \leq |P_t|} C(p_i) \cap C(p_j) \\ &= \{c_{t_1}, c_{t_2}, ..., c_{t_{|C(t)|}}\} \end{aligned} \tag{11}$$

と表す．また，$C(t)$ に含まれる各カテゴリに対応するスコア行列 $S(t)$ を，

$$S(t) = \begin{pmatrix} s_{t_1} \\ s_{t_2} \\ \vdots \\ s_{t_{|C(t)|}} \end{pmatrix} \tag{12}$$

と表す．

ここで，$S(t)$ に含まれるカテゴリ c_{t_k} のスコア $s_{t_k} (1 \leq$

$k \leq |C(t)|$) は，$c_{t_k} \in C(p_i) \cap C(p_j)$ を満たすすべての i，j (ただし，$1 \leq i < j \leq |C(t)|$) の組み合わせについて，c_{t_k} に対応するスコアの総和とする．

記事の場合と同様に，一つの文から得られるスコアの正規化を行う．

$$S'(t) = \frac{S(t)}{\sum_{k=1}^{|C(t)|} s_{t_k}} \tag{13}$$

3.5 文書の特徴表現

それぞれの文について得られた特徴表現を文書全体にわたって集約し，文書を表す重要なカテゴリを特定する．ある複数の文からなる文書 T を，

$$T = \{t_1, t_2, ..., t_{|T|}\} \tag{14}$$

と表す．それぞれの文に対してスコア行列を (13) 式によって計算する．各文から得られたカテゴリ情報を集計し，文書の特徴表現を行う．文書 T から得られるカテゴリの集合を $C(T)$ として，次のように定義する．

$$C(T) = \bigcup_{i=1}^{|T|} C(t_i) = \{c_{T_1}, c_{T_2}, ..., c_{T_{|C(T)|}}\} \tag{15}$$

で表す．文単位での集計と同様に，それぞれのカテゴリの文書内スコア s_{T_k} は，対応するカテゴリ c_{T_k} が含まれる集合 $C(t_i)$ に対応するスコア行列 $S(t_i)$ に含まれるスコア s_{t_i} の総和とする．それぞれのカテゴリの重要度はスコアが大きいほど高くなる．

Wikipedia には多くのカテゴリが含まれているが，一部のカテゴリは非常に広範な記事やカテゴリ間に横断的に出現する．例として，「社会」や「文化」といったカテゴリは非常に多くの記事から到達可能なカテゴリであり，文書の特徴を表現できるカテゴリとはいえない．このようなカテゴリを除去するため，TFIDF の考え方を基に，カテゴリ IDF を定義し，スコアの補正を行う．

本手法では，解析の対象となる文書以外の収集を最小限としたい．そのため，IDF の計算に用いる文集合は，解析対象となる文書とは別に収集している．無作為に選んだ文からなる文集合 T_{IDF} から得られるカテゴリ集合を $C(T_{IDF})$ とする．カテゴリ c について，c が出現した文の数を $count(c)$ とおく．$count(c)$ を用いて，カテゴリ c のスコアにかかる補正である $IDF(c)$ を，

$$IDF(c) = \log_{10} \frac{|T_{IDF}| + 1}{count(c) + 1} \tag{16}$$

とする．文書 T のスコア行列 $S(T)$ に対し，スコアの補正を行う．カテゴリ c_{T_i} に対応するスコア s_{T_i} を，(16) 式を用いて以下のように補正したスコアを，$s_{IDF(T_i)}$ とおき，

$$s_{IDF(T_i)} = s_{T_i} \times IDF(c_{T_i}) \tag{17}$$

とする．(16) 式によるスコアが大きくなるカテゴリの情報を用いることで，文書 T について，その特徴を表現することができる．

4. 実験

4.1 実験の方法

4.1.1 特徴表現の定性的評価

短文から構成され，語彙の頻度情報が十分でない文書の例として，政治家などの 4 つのグループごとにそれぞれ 5 名ずつ計 20 名のうち数名の発信者別のツイート集合を取り上げ，提案手法を適用し，カテゴリを集約した結果について，(17) 式により，スコアの上位 10 件を用いて，定性的な評価を示す．この時，ベースラインとして，語の頻度に基づく方法との比較を行う．

4.1.2 MRR による特徴表現の定量的評価

提案手法が各グループの特徴表現にどの程度有効であるかを平均逆順位を用いて検証を行う．平均逆順位については，各発信者のツイートを見て判断した，ツイート内容に合致するカテゴリを著者のうちの 1 名が主観的に選び，最も高い順位を出現順位とする．平均逆順位 MRR は，(18) 式で表され，出現順位が高いほど 1 に近づく．

$$MRR = \frac{1}{|U|} \sum_{u \in U} \frac{1}{r_u} \tag{18}$$

ただし，U はグループ内の発信者全体を表し，u が各発信者を表す．r_u が発信者 u の特徴に合致するカテゴリが初めて出現した順位を表す．ただし，合致するカテゴリが得られなかった場合には $r_u = 0$ となる．4.1.1 項における実験と同様に，ベースラインとの比較を行う．ベースラインでは，提案手法と同様に，著者のうちの 1 名が主観的に判断したツイート内容に深く関わる語とその出現順位を用いる．出現数が等しい場合は，最も高い順位を用いて評価している．

4.2 データセット

本節では，実験において使用した各種データについて述べる．形態素解析の辞書ファイル mecab-ipadic-NEologd は 2019 年 1 月 31 日に作成したものを使用する．Wikipedia については 2019 年 1 月 20 日にダンプされたファイルを使用する．また，Wikipedia の記事タイトルのうち，NEologd に採録されていないものをユーザ辞書として追加している．Wikipedia の編集に関わる一部のカテゴリについて，独自に作成したストップワードを用いて使用するカテゴリを制限している．特徴表現の対象となる発信者別の 20 名のツイート集合について，2019 年 1 月 31 日から 2019 年 2 月 7 日にかけて収集したものを使用する．ここで，ツイートの収集対象とした発信者を選択する際には，本人確認が行われているアカウントを選択した．**表 1** にそれぞれグループ別のツイート発信者の一覧を示す．

表 1 ツイート発信者一覧

政治家	企業	公共機関	芸能人
安倍晋三	スターバックス	tenki.jp	有吉弘行
小池百合子	セブンイレブン	防衛省・自衛隊	松本人志
枝野幸男	マクドナルド	東京都交通局	三村マサカズ
玉木雄一郎	USJ	日本銀行	堀江貴文
志位和夫	セコム	JAXA	吉高由里子

また，TFIDF の考えに基づく各カテゴリのスコア補正に用いる無作為なツイート集合として，2014 年から 2016 年に収集されたツイートの中から無作為に選択した 10,000 件のツイートを用いた．

4.3 実験結果

4.3.1 特徴表現の定性的評価

提案手法について，数名のツイート集合の特徴表現と，語の出現数に基づく方法を比較する．枝野幸男氏のツイートについて，新しいものから 100 件を用いた場合について，結果を**表 2** に示す．

表 2 の場合では，提案手法が有効に働く例を示している．提案手法では「政治」に関わるより詳しいカテゴリが多く得られている他，ベースラインでは単語「立憲民主党」が得られているのに対し，提案手法ではカテゴリ「政治団体」が得られているなど，背景的なトピックについて取得できていることが分かる．100 ツイートから得られた情報について，ベースラインでは計 725 語，語彙は 456 種類となった．

表 2 枝野氏の特徴表現の比較

順位	ベースライン	出現数	提案手法	スコア
1	皆さん	23	政治	6.43
2	立憲民主党	17	政党	5.33
3	よろしくお願いします	9	社会制度	5.13
4	政治	9	社会	4.88
5	野党	9	政治団体	4.00
6	状況	7	人間関係	3.94
7	辺野古	7	社会科学	2.96
8	候補	6	政治運動	2.87
9	党	6	文化人類学	2.82
10	枝野	6	人文科学	2.77

同様に，有吉弘行氏のツイート 100 件から特徴表現を作成した場合について，結果を**表 3** に示す．

表 3 には，抽出できた単語が非常に少ない場合について示している．ベースラインにおいて抽出された語は計 208 語，語彙は 165 種類となり，抽出数が少なく，頻度に関する情報はほとんど利用できないといえる．ベースライン，提案手法ともに，発信内容と関連性が窺えないカテゴリが多く現れている．従って，取得できた記事数が少ない場合には，提案手法が有効に機能しない場合があるといえる．

次に，宇宙航空研究開発機構 JAXA の公式ツイッターから最新の 100 ツイートを用いて，特徴表現の比較を行った結果を**表 4** に示す．

表 3 有吉氏の特徴表現の比較

順位	ベースライン	出現数	提案手法	スコア
1	田中	10	感情	2.01
2	ラジオ	8	娯楽	1.49
3	最高	7	人文科学	1.38
4	ー	5	放送	1.26
5	あと	3	広告	1.01
6	スマブラ	3	ラテン文字	1.00
7	顔	3	十年紀	0.78
8	D	2	年	0.65
9	U	2	ラテン語	0.64
10	ゲーム	2	ものまねタレント	0.62

表 4 より，ベースラインでは上位の単語について出現頻度が比較的高くなっている．JAXA が発信するツイートが宇宙に関する分野に偏っているため，「衛星」などの単語が出現しやすいことが原因と考えられる．提案手法では，政治家の場合と同様に，こうした単語から「宇宙開発」や「天文学」カテゴリといった背景的な分野を推定できていることが分かる．

表 4 JAXA の特徴表現の比較

順位	ベースライン	出現数	提案手法	スコア
1	衛星	30	宇宙開発	24.57
2	1 号	25	天文学	22.52
3	宇宙	25	宇宙	21.07
4	技術	24	宇宙空間	16.93
5	4 号機	21	対象	10.90
6	イプシロンロケット	21	物理学	10.53
7	JAXA	16	科学	10.21
8	あと	12	技術	10.04
9	はやぶさ 2	10	社会	10.02
10	映像	10	空間	8.37

4.3.2 MRR による特徴表現の定量的評価

それぞれの発信者の特徴に合致するカテゴリが初めて出現した順位を用いて，グループごとに (18) 式を用いて平均逆順位 MRR と，さらに平均順位として MRR の逆数を算出し，結果を**表 5** に示す．また，すべてのグループについて平均逆順位と平均順位の平均値とを併せて示し，**表 6** に示すベースラインとの比較を行う．

表 5 ベースラインの平均逆順位

グループ名	平均逆順位	平均順位
政治家	0.73	1.37
企業	0.59	1.70
公共機関	0.85	1.18
芸能人	0.70	1.43
全体	0.72	1.40

表 6　提案手法の平均逆順位

グループ名	平均逆順位	平均順位
政治家	0.64	1.56
企業	0.64	1.56
公共機関	0.67	1.49
芸能人	0.31	3.23
全体	0.57	1.77

提案手法では，政治家，企業，公共機関のグループでは，およそ 2 位までに関連するカテゴリが得られていることが分かる．また，企業以外のすべてのグループにおいてベースラインが優れている結果となり，特に芸能人の場合では，ベースラインと比べ大きな差が見られた．

このような結果となった理由として，提案手法では「役職」や「協働」などの話題として現れにくいと考えられるカテゴリに非常に到達しやすいため，無作為なツイートから作成したカテゴリの IDF による補正を加味しても，実際の話題によらず上位に現れてしまう傾向があり，結果として特徴表現として適切なカテゴリの順位が下がってしまったことが原因と考えられる．また，芸能人のツイートからの特徴表現においては，ツイートからあまり多くの単語が得られず，特徴表現に適するカテゴリ数自体が少なかったと考えられ，その結果，上記の傾向が更に強く表れたことで，正確なカテゴリ情報を得られなかったことが大きくベースラインとの差がついた原因と考えられる．

5. まとめと今後の課題

提案手法では，ツイート中から抽出した名詞に対し，Wikipedia のカテゴリ情報を利用することで，名詞をカテゴリへと抽象化し，得られた情報を集約することで，ツイート発信者の特徴を表現する手法について研究を行った．

定量的評価の結果として，提案手法では，政治家，企業，公共機関に属するツイートについて，直近の 100 ツイート程度の，語の出現頻度に関する情報が少ない文書から，各発信者の特徴に合致するカテゴリを得ることが可能であるといえる．一方で，芸能人のように，そのツイートの多くが近況や雑談で占められるような発信者については，特徴を表現するカテゴリを得ることが難しいといえる．

今後の課題として，カテゴリ構造から不要なカテゴリ情報の除去を行う，曖昧さ回避先の決定を行う処理を実装する，対応記事が発見できない未知の語に対し，記事を推定する機能を実装する，などが考えられる．また，カテゴリに対して，単純なスコアの大小による重要度の表現のみならず，カテゴリ間の関係を考慮してクラスタリングを行うなど，特徴表現の可視化の観点についても考慮する必要があると考えられる．

謝辞　本研究は JSPS 科研費 JP17K00429 の助成を受けたものです．ここに記して謝意を表します．

参考文献

[1] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pp. 591–600. ACM, 2010.

[2] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65. ACM, 2007.

[3] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pp. 851–860. ACM, 2010.

[4] 梅島彩奈, 宮部真衣, 荒牧英治, 灘本明代. 災害時 twitter におけるデマとデマ訂正 rt の傾向. 研究報告 データベースシステム (DBS), Vol. 2011, No. 4, pp. 1–6, 2011.

[5] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 233–242. ACM, 2007.

[6] 白川真澄, 中山浩太郎, 原隆浩, 西尾章治郎. Wikipedia のカテゴリグラフ解析による語句の確率的分類とその応用. 情報処理学会論文誌データベース (TOD), Vol. 5, No. 3, pp. 51–63, 2012.

[7] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, Vol. 7, pp. 1606–1611, 2007.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[9] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *European conference on information retrieval*, pp. 338–349. Springer, 2011.

[10] 佐々木謙太朗, 吉川大弘, 古橋武. Twitter におけるユーザの興味と話題の時間発展を考慮したオンライン学習可能なトピックモデルの提案. 情報処理学会論文誌数理モデル化と応用 (TOM), Vol. 7, No. 1, pp. 53–60, 2014.