

## ツイート内容のカテゴリ推定に基づくユーザの嗜好分析

日大生産工(学部) ○丘野 みなみ 日大生産工 関 亜紀子

## 1 まえがき

SNSなどで大量の情報がやり取りされる中、ユーザの趣味・嗜好に合った情報にフィルタリングする技術が必要とされている。例えば、Twitterのフォロー機能を利用し、興味のある情報を発信している人をフォローすると、欲しい情報を得ることができる。しかし、本当に欲しい情報と共に、フォローした人の普段の行動のツイートなど、特に必要としない情報も得ることになる。

そこで、個人の興味のあるカテゴリの情報だけを提供するためのフィルタリング手法を検討する。情報やユーザのカテゴリライズに関する研究には、カテゴリライズ対象となるユーザ自身のツイートをWikipediaのカテゴリに対応させることにより、ユーザのカテゴリライズを行うという例がある<sup>1)</sup>。

本研究では、個人の興味のあるカテゴリを提供する。Twitterユーザのログを収集し、ユーザのツイート内容を単語の出現頻度や関連性などからカテゴリ推定し、推定結果から趣味嗜好を分析する。

## 2 辞書更新の提案

提案するカテゴリ推定とは、ユーザのツイートに含まれるキーワードの特徴から、政治やアイドルなど興味があるカテゴリを推定することである。例えば、『国民が何の政策を重視するか自分で決めるべき。』とツイートされたとき、人間の目では「国民」や「政策」といった単語があることから政治についてのツイートであると判断することができる。

このようなカテゴリの推定を機械化するには、カテゴリの推定基準となるキーワードを集めた「辞書」が必要となる。しかしながら、日々、話題となる人物や出来事が変化するSNSを対象に、完全な状態の「辞書」を事前に持つことはほぼ不可能に近いといえる。また、多くのカ

テゴリに対応できるようにするには、それだけの数の辞書が必要になる。そこで、提案する方式では、初期状態としてカテゴリごとの簡易的な辞書を用意し、新しいツイートが出現するたびに、新規の単語を辞書に登録することで辞書を更新する。用意する辞書は、人名辞書と名詞辞書の2つである。人名辞書には、政治家やアイドルなどの人名とそのカテゴリに登録する。名詞辞書は、カテゴリごとに頻出する名詞をキーワードとして登録する。なお、今回は、「政治家」、「アイドル」、「役者」、「スポーツ選手」、「歴史上の偉人」の5つのカテゴリについて、辞書を用意している。

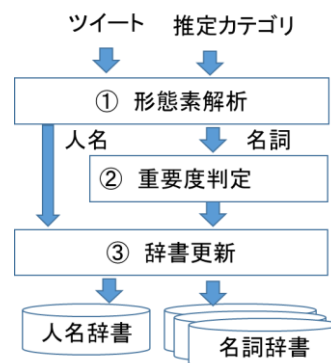


図 1 辞書更新の手順

一連の辞書を更新するためのアルゴリズムを図1に示す。①の形態素解析では、1回にツイートされた一連の文字列を文法的に意味のある単位である形態素に分割し、それぞれの形態素の品詞等を判別する。②の重要度判定では、①で得られた形態素のうち、名詞のみを対象とし、各形態素の重要度をtf-idf法<sup>2)</sup>により計算する。③の辞書更新では、重要度が5.00以上のキーワードを名詞辞書に追加する。また、ツイートに人名が含まれる場合は、人名辞書に、人名と推定されたカテゴリ属性に登録する。

表 1 名詞辞書の中身

| カテゴリ名           | 単語(名詞)                                                                                                                                                      |
|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 政治<br>(36語)     | 政治家、政治、人、国会、国会議員、熟練政治家、家、国民、選挙、官僚、原理主義者、自民党、日本、交渉担当者、人々、欧州委員長、カルトキチガイ政治家、女性政治家、政治団体、政治運動、国、知事、自分、政治学者、大統領制、政治指導者、臨時国会会期、政治利用、政治経験、漫談家、被害者、奨学金、政治的批判、天皇家、運動家 |
| アイドル<br>(15語)   | 嵐、横アリ、ジャニーズ、嵐にし、VS嵐、亜嵐、隊No.、人、o、メンバー、嵐翔、嵐電、RT嵐、嵐NEWS、ファン                                                                                                    |
| 役者<br>(14語)     | サムライせんせい、サムライ先生、神木隆之介主演、神木隆之介枠、神木隆之介着用、神木隆之介タイプ、神木隆之介サマーウォーズ、神木隆之介time、サムライ、立花瀧役神木隆之介、ドラマ、神木隆之介山田孝之新井浩文、映画、名                                                |
| スポーツ<br>(21語)   | 千葉ロッテ、プロ野球、ロッテ、高校野球、野球、プロ、プロ野球選手、選手、ロッテロッテ千葉ロッテ、プロ野球ファン、投手、西野七瀬勇士、失点西野勇士、ロッテ西野勇士試合、マー、抑え西野勇士、ドラマチックプロ野球、プロ野球PRIDE、プロ野球速報、プロ野球ファンガチ勢、審判                      |
| 歴史上の偉人<br>(18語) | 新選組、新撰組、土方歳三資料館、龍馬暗殺、坂本龍馬暗殺、土方歳三巡り、坂本龍馬とわ、by坂本龍馬、坂本龍馬像、人、土方歳三コス、あたり土方歳三、散華土方歳三、土方歳三墓、鬼、桜鬼、土方歳三記念館、写真                                                        |

表 2 人名辞書の中身

| カテゴリ名            | 人名                                         |
|------------------|--------------------------------------------|
| 政治家<br>(2語)      | 鳥越、鳥越俊太郎                                   |
| アイドル個人名<br>(5語)  | 二宮和也、二宮、翔、相葉、櫻井翔                           |
| 役者名<br>(7語)      | 神木隆之介、錦戸亮、神木、錦戸、森本レオ、亮、赤西仁                 |
| スポーツ選手名<br>(2語)  | 西野勇士、満井秀章                                  |
| 歴史上の偉人名<br>(11語) | 土方歳三、坂本龍馬、勝海舟、坂本龍馬、土方、斎藤一、歳三、海舟、近藤勇、龍馬、坂本、 |

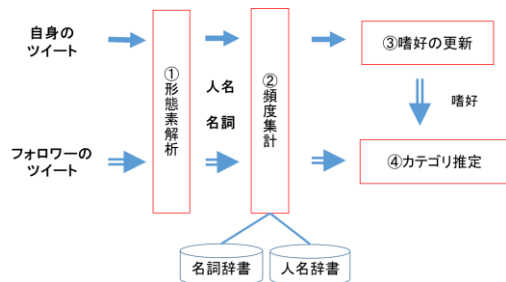


図 2 ツイートの解析から表示までの手順

### 3 カテゴリ推定の手順

提案するカテゴリの推定手法を図2に示す。ここでは、自身のツイートに基づく嗜好の更新手順と、フォロワーのツイートに対するカテゴリの推定から表示までの手順を示している。

自身のツイートに対する手順は、①の形態素解析で人名と名詞を取得する。②の頻度集計で①で得た名詞と人名をもとに、名詞と人名それぞれの辞書に基づきカテゴリの単語の出現頻度を算出する。③では嗜好の更新を行う。ツイートによって嗜好は変化するため、カテゴリの推薦結果は毎回更新する。

フォロワーのツイートに対する手順は、①②の処理ののち、④で自身の嗜好とフォロワーのツイートをもとにカテゴリの推定を行う。

### 4 カテゴリ推定の評価検証方法

提案手法を検証するために、テスト用データとして、Twitterのログを100件用意し表3に示すカテゴリ構成の仮想的なユーザ10人分のツイートを作成する。用意したデータは、予めカテゴリが自明なツイートを各カテゴリから20件ずつ収集する。ここでは、政治家は「民進党所属の政治家」、アイドルは「嵐」、役者は「ドラマ『サムライせんせい』に出演していた俳優」、スポーツ選手は「野球選手」、歴史上の偉人は「幕末に活躍した人」とする。また、評価用のデータとして、政治家は「民進党+自民党」、スポーツ選手は「野球+サッカー」を用意し、検証する。

表 3 ツイート作成状況の一例

|      | 政治家 | アイドル | 役者  | スポーツ選手 | 歴史上の偉人 |
|------|-----|------|-----|--------|--------|
| ユーザ1 | 80件 | 5件   | 5件  | 5件     | 5件     |
| ユーザ2 | 5件  | 70件  | 15件 | 5件     | 5件     |
| ユーザ3 | 0件  | 0件   | 90件 | 5件     | 5件     |
| ユーザ4 | 10件 | 10件  | 10件 | 60件    | 10件    |
| ユーザ5 | 15件 | 5件   | 10件 | 10件    | 60件    |
| ユーザ6 | 50件 | 40件  | 5件  | 5件     | 0件     |

### 5 まとめと今後

本稿ではカテゴリ推薦に用いる辞書更新の手順とツイートに対するカテゴリ推定の方法について検討を行った。

辞書の登録方法の手順や評価検証から、今後はテストデータを用いた検証を行い、結果を評価していきたい。

#### 「参考文献」

- 1) 胡 寅駿, 谷田 泰郎: Wikipediaのカテゴリ情報を用いたTwitterユーザの関心分野の抽出, 電子情報通信学会技術研究報告, NLC 113(213), pp 17-21, (2013/9)
- 2) 宮崎 将隆, 川端 豪: Wikipediaページへのtfidf法の適用, 情報処理学会研究報告, Vol.2009-SLP-77, No.2, pp. 1-6 (2009/7)