Twitterのプロフィールを使ったコンテンツの話題抽出

高橋 文彦　　山下 達雄
ヤフー株式会社

{ftakahas, tayamash}@yahoo-corp.jp

1　はじめに

SNS，特にTwitterでは日々膨大な量のツイートが投稿されている．このような状況に対して我々はユーザーに興味がある情報を提示する応用を念頭に研究を進めている．そのユーザの興味の対象を「コンテンツ」と呼ぶこととする．コンテンツは例えばアイドルグループやロックバンド，野球球団などである．本稿ではコンテンツごとに話題になっているツイートを抽出する方法を提案し，その評価を行う．

このような問題に対するアプローチとして，教師あり学習の手法を用いて，入力をツイート，出力をコンテンツの話題か否かの２値とした分類問題として解く方法が考えられる．しかしコンテンツの話題は時間で変化するため学習データの劣化や，他のコンテンツに対応するたびに学習データを作る必要がありスケールしない．そこで我々はTwitterのプロフィールを使用して，話題のツイートを抽出する方法を提案する．手法の概要を図１に示す．Twitterのプロフィールにはユーザーの興味があるコンテンツについて書かれることが多い．この特徴を利用して，プロフィールを使ってコンテンツのファンを自動で抽出し，ファンと全体のユーザーでのツイートに対する反応の差を使ってコンテンツの話題のツイートを抽出する．

2　関連研究

Twitterから情報を抽出する研究が進んでいる．Sriramら [3] は，ツイートのタイプの分類を行うために，Bag-of-Wordsの他にユーザーの情報とテキストの特徴を素性に追加することを提案した．また，SNSの投稿などの情報からプロフィールを推定する研究もある．Liら [1] は，Twitterのユーザーの属性 (配偶者，職業，学歴など) を推定するために他のSNSの情報をDistant Supervisionの枠組みで活用している．こ

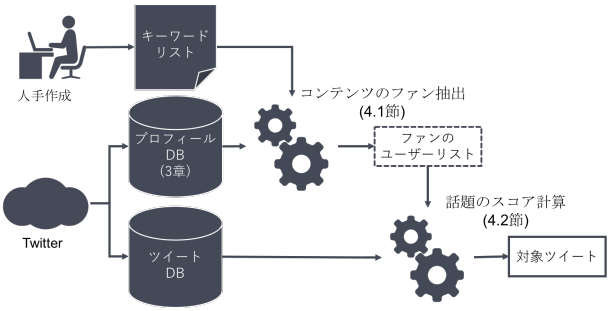

図 1: コンテンツで話題になっているツイートを抽出する手法の概要

の他にも，性別や年齢，出身地を推定する研究 [2] が進められている．本研究はこれらの研究と異なり，ツイートの分類のために間接的にユーザーのプロフィールを扱っている．

3　プロフィール

Twitterに登録できるプロフィール情報[1]を表１に示す[2]．名前以外の項目は任意入力欄である．FacebookやGoogle+などのSNSに比べて項目が少なく，フリーテキストで記入することが特徴的である [1]．本稿ではプロフィール情報の自己紹介をプロフィールと呼ぶ．

3.1　プロフィールの分析

プロフィールとツイートの比較を通してプロフィールの特徴を分析する．2017年3月にツイートしたユーザーを収集し，そのうち空欄でないプロフィールを対象に分析する．分析対象になったプロフィールは205,184件であった．また，2017年12月1日から3日の全ツイートからサンプリングした115,872件のツイートを

[1] 2017年12月現在
[2] 記載例は実際のプロフィール情報を参考に作成．

表 1: Twitter に登録できるプロフィール情報

項目	説明	必須項目	記載例
名前	表示される名前，ID とは異なる	○	やっくん
自己紹介	経歴や興味のあるコンテンツについてフリーテキストで記入	-	社会人です。ギターが趣味です。ワンオク ☆ 広瀬すず ☆ ヤクルトスワローズ ☆ 山田哲人 ☆ 青木宣親
場所	所在地など	-	東京都
ホームページ	自身のホームページの URL など	-	

表 2: プロフィールとツイートの比較

	プロフィール	ツイート
平均文字数	67.44	34.42
固有名詞占有率	18.70%	8.29%
機能語占有率	20.71%	33.87%
記号占有率	24.97%	15.94%

分析対象にする．統計情報を表 2 に示す．固有名詞占有率・機能語占有率・記号占有率は，プロフィールとツイートを形態素解析し，全形態素のうち固有名詞・機能語・記号の占める割合を計算したものである[3]．形態素解析器は MeCab を使用し辞書には mecab-ipadic-NEologd[4][4]を使用した．固有名詞は品詞が‘固有名詞’の形態素，機能語は品詞グループが，‘副詞’・‘助詞’・‘助動詞’・‘接頭詞’・‘連体詞’のいずれかの形態素，記号は品詞グループが‘記号’の形態素とする．

占有率を比較するとツイートに対してプロフィールは，固有名詞と記号が多く機能語が少ないことがわかる．実際に確認すると，表 1 の例のようにプロフィールは文で書かれるよりも単語の羅列で書かれることが多く，羅列された単語のデリミターとして記号が多用されていることがわかる．

平均文字数を比較すると，ツイートよりプロフィールの方が 1.96 倍長い．ツイートがそれ単体ではプロフィールに比べて情報が少ないと言える．ツイートの場合，その前後の出来事や投稿されたツイートと一緒に見ることを想定して投稿されていたり，画像や動画と共に投稿されることがあるため，平均文字数が少なくなっていると考えられる．これに対してプロフィールは，SNS という性質上，ユーザーが興味のあるコンテンツをわかりやすく示すことで他のユーザーと繋がりやすくしていると考えられる．また Twitter ではプロフィールがユーザー検索の検索対象フィールドのため，SEO 対策としてわかりやすくコンテンツを記載しているとも考えられる．

[3]例えば固有名詞占有率であれば，固有名詞の形態素数を C，全形態素数を N とした時，$\frac{C}{N}$ で計算する．

[4]v0.0.5

4 コンテンツのツイート抽出

本研究は，あるコンテンツについて興味があるユーザーに対して抽出したツイートを提示する Web サービスに応用することを想定している．このため抽出するツイートは，コンテンツに関連し，かつコンテンツのファンが好むものを抽出したい．例えばネガティブなツイートや批判的なツイートは対象にならない．

先述したように，教師あり学習の枠組みでの話題判定では学習データの劣化やスケールの難しさがある．本研究ではこのような課題からツイートに対するラベル付きデータを用いない方法で抽出を行う．Twitter ではツイートに対して他のユーザーが‘いいね’や‘リツイート’，‘リプライ’などの反応を起こす．この反応を利用してツイートを抽出する．コンテンツに関する話題は，ファンとそれ以外のユーザーで反応が異なると予想される．これに対して，コンテンツに関係しない話題はファンとそれ以外のユーザーで反応は変わらないと考えられる．例えば，アイドルグループ「嵐」というコンテンツに関する話題は,「嵐」のファンならリツイートやいいねの数が多くなると予想される．一方で「嵐」コンテンツに関係のない「地震」が起きた場合は,「嵐」のファンかどうかに関わらずリツイートやいいねの数が多くなると考えられる．

この特徴を利用して，次の 2 段階でツイートを抽出する．

1. プロフィールを使用して，コンテンツのファンを抽出
2. 抽出したファンと全体のユーザーの反応の違いを使って話題のツイートを抽出

ここで，ファンはプロフィールにコンテンツに関して記載しているユーザーと定義する．また，ツイートに

対する反応はリツイートを使用する．

4.1 ファンの抽出

コンテンツについてプロフィールに書かれているユーザーをファンとして抽出する．3.1 節の分析で，プロフィールには単語が羅列して書かれやすいことがわかっているので，シンプルなキーワードマッチで判定を行う．コンテンツに関するキーワードリストと NG ワードリストを用意して，キーワードがプロフィールに含まれるかつ，NG ワードがプロフィールに含まれないユーザーをファンとして抽出する．

キーワードリストと NG ワードリストの例を表 3 に示す．キーワードには表記揺れや関連する単語を選択する．NG ワードにはキーワードリストにマッチしたプロフィールのうち，誤ってマッチしたプロフィールを除外するように単語を選ぶ．例えば，コンテンツ「東京ヤクルトスワローズ」は球団の名前だが，その通称である「ヤクルト」をキーワードに選び，この際に企業名「ヤクルト」を意図するプロフィールにもマッチするため,「ヤクルト本社」や「水戸ヤクルト」を NG ワードとして追加する．

新たにコンテンツを追加する場合，キーワードリストはクラウドソーシングなどを使って作成できる．また NG ワードリストも，キーワードリストを使った抽出結果を見て作成するため，技術者でなくても作成できる．この方法は辞書を使ったシンプルな文字列マッチングなので，Aho Corasick 法などで高速に判別ができる．

4.2 コンテンツの話題抽出

コンテンツのファンとそれ以外のユーザーでのリツイート数の分布の差を使って，話題の度合いをスコアリングをする．分布の差の計算にはカイ二乗値を使い，話題の度合い $fanScore$ を式 1 で定義する．$n_{fa,rt}$ はファンのリツイート数，$n_{fa,tw}$ はファンのツイート数，$n_{ge,rt}$ は全ユーザーのリツイート数，$n_{ge,tw}$ は全ユーザーのツイート数である．ファンと全ユーザーでのツイート数に対するリツイート数の割合（リツイート割合）を計算し，ファンのリツイート割合が全ユーザーのリツイート割合と異なると $fanScore$ の値が高くなる．

$$fanScore = \chi^2 = N\frac{V}{E} \tag{1}$$

$$N = n_{fa,rt} + n_{fa,tw} + n_{ge,rt} + n_{ge,tw} \tag{2}$$

$$V = (n_{fa,rt}n_{ge,tw} - n_{fa,tw}n_{ge,rt})^2 \tag{3}$$

$$E = (n_{fa,rt} + n_{fa,tw})(n_{fa,rt} + n_{ge,rt}) (n_{fa,tw} + n_{ge,tw})(n_{ge,rt} + n_{ge,tw}) \tag{4}$$

$fanScore$ を計算する対象は，ファンのツイートに限らず全てのユーザーのツイートを対象とする．$fanScore$ のスコアリング結果を用いて，閾値以上のツイートをコンテンツの話題として抽出する．

5 実験

実際の Twitter のデータを用いて本手法の評価を行う．

5.1 実験設定

本手法で抽出したツイートがコンテンツ特有の話題である割合（正解率）を他の抽出方法と比較することで評価する．比較する抽出方法は，全てのツイートを対象にファンのリツイート数をスコアにして，スコアが高いツイートを抽出する．この方法およびスコアを $freq$ と呼ぶ．

漫画やタレント，球団など 5 つのコンテンツを対象に評価する．話題がある日ない日があると考えられるため，1 つのコンテンツにつき 3 日分ずつ抽出する．コンテンツと日ごとに，$fanScore$ と $freq$ それぞれでスコアの高い上位 10 件のツイートを抽出する．ファンの抽出は 3.1 節の分析と同じデータを用いる．また話題の抽出は，2017 年 3 月 1 日から 8 日の間でコンテンツごとにランダムに 3 日分選択して 1 日ごとに集計し抽出した．各コンテンツについて詳しい評価者に評価を依頼し，コンテンツ特有の話題であるかどうかを，正しい，間違っている，判断できないのラベルをつける．

5.2 評価

コンテンツごとに ‘正しい’ ラベルの付いた割合を図 2 に示す．全てのコンテンツで $freq$ より $fanScore$ の方が特有の話題が多かった．

$fanScore$ で抽出したツイートは，コンテンツの公式ユーザー（メンバーや選手）のツイートや，ユーザーが

表 3: キーワードリストと NG ワードリストの例

コンテンツ	キーワードリスト	NG ワードリスト
東京ヤクルトスワローズ	ヤクルト, スワローズ	ヤクルト本社, 水戸ヤクルト
ONE PIECE	ワンピース	ファッション, コーデ
AKB48	AKB, ゆきりん, まゆゆ	元 AKB, 公式ライバル

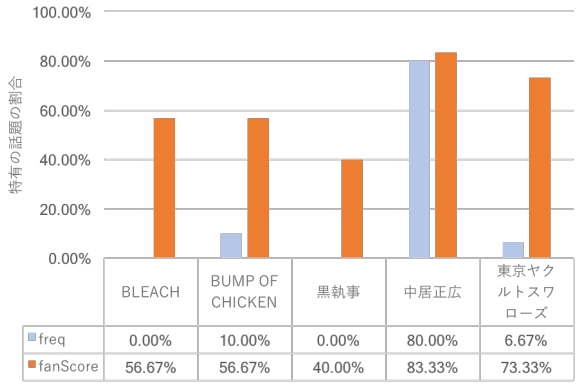

図 2: コンテンツ特有の話題の割合

書いた絵や，ライブのレポートなどが多かった．この中にはキーワードを含まないツイートも多く抽出された．例えば，ロックバンドである「BUMP OF CHIKEN」コンテンツでは，ツイートの本文には「やっと色塗りできた」としか書かれていないメンバーを書いた絵のツイートが抽出された．この他にも，動画や画像を見なければそのコンテンツと判断できないツイートが抽出できた．こういったツイートの抽出はツイート本文だけを使った方法では難しい．

一方で，$fanScore$ がコンテンツ特有の話題ではない (誤って抽出した) ツイートには，関連コンテンツのツイートを抽出しているものが多かった．例えば，「中居正広」コンテンツでは所属グループの元メンバー「草薙剛」を話題の中心にしたツイートが含まれていたり,「東京ヤクルトスワローズ」コンテンツでは他の球団の話題のツイートが含まれていた．これはコンテンツのファン層が被っていることが原因である．しかし，これらのツイートはファンのリツイート数 $n_{fa,rt}$ が比較的少なかったため，$n_{fa,rt}$ に閾値を設けてフィルタ処理するなどで改善が見込める．

対して $freq$ では，コンテンツに関係ない話題のツイートが多かった．このことから，ファンの情報だけでなく，全体のユーザーの反応と比較することが有効なことがわかる．

6 おわりに

Twitter のプロフィールを使用して，コンテンツの話題のツイートを抽出する方法を提案した．本手法は，公式ユーザーのツイートやキーワードを含まない動画や画像のツイートも抽出することができる．本研究の成果は，ヤフーのサービス「Yahoo!リアルタイム検索」[5]で応用されている．

本稿ではコンテンツの話題の粒度としてツイートを採用したが，実際に抽出したツイートは同じ話題に関するものが多い．これを解消するため，今後は話題をまとめ上げて要約する技術の検証を進めたい．

参考文献

[1] Jiwei Li, Alan Ritter, and Eduard Hovy. Weakly supervised user profile extraction from twitter. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 165–174, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[2] Delip Rao and David Yarowsky. Detecting latent user properties in social media. In *Proc. of the NIPS MLSN Workshop*, 2010.

[3] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pp. 841–842, New York, NY, USA, 2010. ACM.

[4] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会 (NLP2017), pp. NLP2017–B6–1. 言語処理学会, 2017.

[5] https://promo-search.yahoo.co.jp/realtime/wadainow/