Twitter 上のコミュニティの可視化
Visualization of Communities in Twitter

小薗　絢介
Shunsuke Kozono
法政大学情報科学部コンピュータ科学科
E-mail: shunsuke.kozono.9n@stu.hosei.ac.jp

Abstract

Many people are using Twitter to communicate with others who share interests and also to search for news that they want to know. On the other hand, many companies are using Twitter to promote their products and services. It is better for such companies to be able to know about people who are interested in their products and services and about their locations for effective marketing. This paper proposes a method for the visualization of communities of Twitter users. The method first identifies communities of Twitter users, which is done by using tf-idf weighting and cosine similarity. Tf-idf weighting is used to decide communities' words, and cosine similarity is used to compare companies' tweets with users' tweets. After that, the method visualizes communities of users by plotting the users on the map of Japan. The details of the visualization can be zoomed in to reveal the members and the populations of communities. This paper also shows the experimental results of the visualization of four communities (Baseball, Soccer, Sumo, and Golf) in Twitter.

1. はじめに

Twitter [1]は、投稿の気軽さや，自分の嗜好に合ったユーザとのコミュニケーションのしやすさなどから，近年多くの人々に利用されている．また，企業や自治体などが Twitter を宣伝活動や情報発信目的で利用するケースも増加している．中でも，一般のユーザをフォローし，意見やユーザの興味を収集しているケースも存在し，ユーザの把握が大切になっている．

本研究は，Twitter 上のユーザのコミュニティを分析し、可視化することを目的とする．これによって嗜好に応じたコミュニティがどのような地域で繁栄しているかを可視化して探ることや，地域ごとの嗜好がどの程度異なるかを把握することを可能にする．

本研究では，企業などの公式アカウントで多く使用された特徴的なワードをコミュニティワードとし，そのコミュニティワードをつぶやいているユーザをコミュニティの所属者とする．このようなコミュニティの判別を tf-idf 法とコサイン類似度によって行う手法を提案し、その実験の結果を示す．

2. 関連研究

Pennacchiotti と Gurumuthy [2]は，LDA を用いてユーザをトピックの混合物として表現し，類似のユーザを Kullback Leibler (KL)やコサイン類似度を用いて推薦している．山本 [3]は tf-idf などを用いてテレビ番組に対する意見をリアルタイムに検出している．この研究では，テレビ番組表のテキストと番組上の字幕テキストに対して形態素解析を行い，tf-idf 値と最後の出現から現時点までの時間を考慮して特徴単語を抽出しユーザの検出を行う手法と，Twitter ユーザのツイートに対して Labeled LDA を用いて特徴単語を抽出しユーザの検出を行う手法を合わせたハイブリッドな方法を提案した．

3. 準備

tf-idf 法とコサイン類似度について述べる．

3.1. tf-idf 法

tf-idf 法は文書における特徴単語を決定付けるための単語の重み付けを算出する [4]．以下に算出方法を示す．

$$\mathrm{tf}(t,d) = \frac{n_{t,d}}{\sum_{s\in d} n_{s,d}}$$

$$\mathrm{idf}(t) = \log\frac{N}{\mathrm{df}(t)} + 1$$

$$\mathrm{tfidf}(t,d) = \mathrm{tf}(t,d)\cdot \mathrm{idf}(t)$$

tf (Term Frequency)はそれぞれの単語の文書内における出現頻度を表す．ある単語tの文書dでの出現回数$n_{t,d}$に文書dに出現する全ての単語の数$\sum_{s\in d} n_{s,d}$を割ることで算出できる．

idf (Inverse Document Frequency)はそれぞれの単語がいくつの文書で共通して使われているかを表す．文書総数Nをある単語tが出現した文書の数$\mathrm{df}(t)$で割ったものの対数をとる．

この tf，idf の数値を掛け合わせたものは tf-idf と呼ばれ，この数値が高いほど単語tが文章dにおける特徴単語である可能性が高い．

3.2. コサイン類似度

コサイン類似度は文書間の類似度を測るために使用される [5]．その値は文書の特徴を表すベクトル同士の成す角度の大きさを表現し，1 に近いほど類似し，0 に近いほど類似していない．これは以下により算出される．

$$\cos(\vec{q},\vec{d}) = \vec{q}\cdot\vec{d} = \sum_{i=1}^{|V|} q_i d_i$$

Supervisor: Prof. Hiroshi Hosobe

4. 提案手法

本研究で提案する手法は，企業などの公式アカウントに類似するユーザを発見することでコミュニティを特定し，地図上に可視化するものである．

4.1. 利用するデータ

本研究で利用するデータを表 1 に示す．コミュニティとして扱われるアカウント（公式アカウントと呼ぶ）については，コミュニティを区別するためにユーザ ID を，コミュニティの特徴単語を見つけるためにツイートを利用する．一方，コミュニティに振り分けられるアカウント（ユーザアカウントと呼ぶ）については，ユーザを区別するためにユーザ ID を，コミュニティの所属可否を決定するためにツイート内容とプロフィール情報を，可視化するためにツイート時刻とツイート位置情報を利用する．

表 1　利用するデータ一覧

	公式アカウント	ユーザ アカウント
ユーザ ID	○	○
ツイート内容	○	○
ツイート時刻	×	○
位置情報	×	○
プロフィール 情報	×	○

4.2. 判別方法

4.2.1. コミュニティの特徴単語の抽出

コミュニティの特定のために，公式アカウントの特徴単語を利用する．公式アカウントを複数用意し，それぞれでつぶやかれたツイートの中から形態素解析によって名詞を抽出し，それぞれの名詞の tf-idf 値を算出する．

4.2.2. ユーザアカウントの所属可否の決定

ユーザアカウントのコミュニティ所属可否を決定するために，プロフィール文とツイート文で利用されている名詞を抽出し，tf-idf 値を算出した後，コミュニティとのコサイン類似度を計算する．なお，プロフィール文はツイート回数分利用する．これは，特定のジャンルだけでなく，多くのジャンルをつぶやいているユーザでも検知できるようにするためである．

4.3. 可視化

複数のコミュニティの分布を日本地図上にプロットし，コミュニティを可視化する．各地域の特徴を発見できるように拡大縮小を可能にしており，時間経過での変化を見られるように月別での表示を可能にしている．

5. 実装

提案手法の実装について述べる．実装に使用したプログラミング言語は Java である．

5.1. ツイートデータの取得

Twitter からデータを取得する際に，Twitter4J [6]と呼ばれる Java 用の Twitter API ライブラリを使用した．API の取得制限があるため，パブリックストリームを用いて取得を行った．データの取得期間は、2014 年 7 月 10 日から 2014 年 11 月 20 日までである．

5.2. 形態素解析

本研究では，ツイート内容での名詞をコミュニティワードとして扱うため，名詞のみを抽出する．そのために、Kuromoji [7]による形態素解析を実装した．辞書機能に関しては標準搭載のものを利用した．

5.3. tf-idf 法

tf-idf 法は 3.1 項で記載したものを実装した．実験におけるコミュニティワードを抽出する際には，文書数Nを公式アカウントとユーザアカウントの数の合計として計算を行う．また，計算に含まれる単語の品詞は名詞のみであり，名詞に分類される中で数字のみのもの（11, 9, 2014 など），記号が使われるもの，専門用語や英単語とは考えられない英語文字列（http，co, XxWL など）は，手動で排除している．

サッカーに関する情報をツイートしたアカウント(@SoccerKingJP)の tf-idf 値上位 5 位を表 2 に示す．

表 2 tf-idf 値例

単語	tf-idf 値
代表	0.06111
試合	0.03409
移籍	0.03314
日本	0.03086
情報	0.02976

5.4. コサイン類似度

コサイン類似度は 3.2 項で記載したものを実装した．この式は正規化された単位ベクトルに適用できるものであり，今回の実験においてはそれぞれの単語の tf-idf 値を掛け総和を算出している．比較する文書は，公式アカウントのツイートと，緯度経度情報のあるユーザアカウントのプロフィール文とツイート内容の両方である．

5.5. 可視化

複数のコミュニティを色別に分け，それらのコミュニティに所属するユーザを日本地図上にプロットすることで可視化を実現している．

6. 予備実験

実験におけるコミュニティ所属の決定材料であるコサイン類似度については，どの程度が類似といえるのかが明らかでなかったため，予備実験を行った．

予備実験での公式アカウントは表 3 の通りである．

表 3　公式アカウント一覧

ジャンル	アカウント ID	アカウント名	使用ツイート数
サッカー	@SoccerKingJP	サッカーキング	9111
野球	@SponichiYakyu	スポニチ野球記者	1733
相撲	@sumokyokai	日本相撲協会公式	1336
ゴルフ	@ALBA_golfnews	ゴルフ情報 ALBA.NET	10
ゴルフ	@GDO_news	GDO 編集部	24
ゴルフ	@pargolf_jp	パーゴルフ	28
ダミー	後述		

ユーザアカウントは表 4 のように分類した．なお，サッカーに属するユーザを A1, A2，野球に属するユーザを B1, B2，相撲に属するユーザを C1, C2，ゴルフに属するユーザを D1, D2，ダミーに所属するユーザを E1, E2 とした．ダミーとはこの実験における野球，サッカー，相撲，ゴルフに属さないユーザのことである．このコミュニティはこれらユーザの集合である．

表 4　ユーザアカウントの分類

ジャンル	ユーザ数
サッカー	2 (A1, A2)
野球	2 (B1, B2)
相撲	2 (C1, C2)
ゴルフ	2 (D1, D2)
ダミー	2 (E1, E2)
合計	10

6.1. 結果

各ユーザとコミュニティのコサイン類似度は表 5 のようになった．なお，小数点第 6 位以降は切り捨てている．

表 5　予備実験結果

ユーザ	サッカー	野球	相撲	ゴルフ
A1	**0.00667**	0.00295	0.00130	0.00210
A2	**0.01199**	0	0.00105	0.00072
B1	0.00065	**0.00550**	0.00060	0.00011
B2	0.00222	**0.00236**	0.00026	0.00116
C1	0.00016	0	**0.02015**	0
C2	0.00011	0	**0.01363**	0
D1	0.00006	0	0.00004	**0.02509**
D2	0.00209	0.00064	0	**0.00434**
E1	0.00064	0.00018	0.00438	0.00042
E2	0.00110	0.00654	0	0.00465

6.2. 考察

A1, B1, B2, D2 の類似度が低い要因としては，所属ジャンルに関するツイートが少ないことが挙げられる．E2 における相撲以外の類似度が高くなった要因としては，サッカー，野球，ゴルフでよく使われる「プロ」という表記が多く使われていたこと考えられる．

7. 実験

予備実験で得られたコミュニティ決定に最適と考えられるコサイン類似度を利用し，実験を行う．今回の実験では，コミュニティ所属として扱うコサイン類似度を 0.01 以上とした．対象となる公式アカウントは予備実験と同様である．

適合率・再現率を評価するため，各ユーザを表 6 の通り分類した．本実験では複数のコミュニティに所属しているユーザも存在する．なお，今回のユーザは位置情報を含むツイートをしているユーザのみとしている．

表 6　本実験ユーザ分類

ジャンル	ユーザ数
サッカー	9
野球	15
相撲	11
ゴルフ	17
サッカー・野球	1
サッカー・相撲	1
野球・相撲	8
野球・ゴルフ	15
サッカー・野球・相撲	2
サッカー・野球・ゴルフ	1
ダミー	**10**
合計	**90**
サッカー総数	**14**
野球総数	**42**
相撲総数	**22**
ゴルフ総数	**33**

7.1. 結果

実験での適合率・再現率は表 7 の通りとなった．なお，小数点第 4 位で四捨五入を行っている．可視化結果を図 1 に示す．

表 7　本実験適合率・再現率

	適合率	再現率
サッカー	0.429	0.214
野球	0.968	0.714
相撲	0.714	0.227
ゴルフ	0.8	0.121

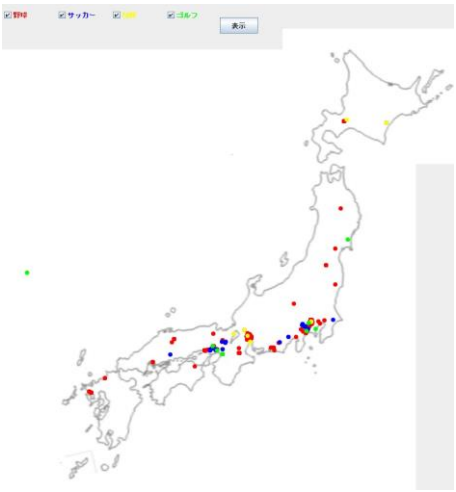

図 1　可視化結果

7.2. 考察

本実験では，野球の適合率と再現率が高く，相撲，ゴルフにおいても適合率は高い値となった．一方，サッカーの適合率と再現率，相撲，ゴルフの再現率は低い結果となった．再現率の低い要因としては，各ユーザがプロフィールにスポーツ名を記載しているが，踏み込んだ話題（試合内容，選手名など）を記載していることが少ないことが挙げられる．一方，野球においては，球団名を記載していることが多かったことと，ツイート取得期間がオールスター戦以降の優勝争い期間であり，踏み込んだ話題が多かったことにより良い結果になったと考えられる．

コミュニティの特徴単語を抽出する際の名詞は、Kuromoji 標準搭載の辞書を元にしたが，サッカー関連でよく使われる「日本代表」，「日本国内」と，野球関連でよく使われる「日本シリーズ」がそれぞれ「日本」と「代表」，「日本」と「国内」，「日本」と「シリーズ」に分割されて扱われており，「日本に来た」などと同じ扱いとされてしまったため，誤検知の原因となった．また，選手名などは公式ではフルネームおよび正式名称で記載することが多いが，ファンは愛称や苗字のみまたは名前のみで記載することが多いため，通常の辞書ではユーザを取得できないケースが生じると考えられる．これらの問題は，コミュニティの規模や一般性に応じて考えていく必要がある．また，今回の実験で利用した手法は同じ意味でもひらがな，カタカナ，漢字などの違いで別の単語として扱われるため，この点に関しても課題が残った．

可視化結果は，プロチームやプロ選手だけでなく，部活としてやプライベートとしての趣味があるため，プロチーム所在地や大会開催地などに関連した特色のある変化を得ることが出来なかった．この点では，テーマ設定が良くなかった，一般ユーザアカウントの選択が良くなかったのではないかと考えられる．また，ツイートごとに可視化をしているために一般ユーザの中でもツイート数が多いユーザが多く反映されてしまうので可視化としてはあまり良いものとはいえない．これらユーザに対して平均的な位置を推定し，可視化する手法であればよいのではと考える．

今回は緯度経度情報がついているユーザのツイート情報を可視化に利用したが，経度緯度情報が記載されているユーザは，Cheng らの研究によると全体のツイートの 0.42%しか存在しない [8] [9]．また，Twitter のプロフィール部分にある現在地の記載は，事前調査の結果，有用に使えるものがさらに少ないことが判明した．今回の実験でも 1 ユーザあたりの位置情報ツイート数は少なく，ユーザの位置情報特定は難しい課題といえる．

8. おわりに

本研究では，コミュニティをユーザのツイートとプロフィール情報をもとに tf-idf 法とコサイン類似度を用いて振り分け，可視化を行った．実験としてサッカー，野球，相撲，ゴルフのコミュニティの判別，可視化を行った．判別においては，野球における適合率と再現率が高くなったが，サッカー，相撲，ゴルフでは再現率がかなり低い結果となった．また，可視化においては，選択したユーザなどの要因から特徴を見出すことは出来なかった．コミュニティ所属可否の条件の精密化，単語の揺れと位置情報ツイートの数の問題への対応が今後の課題である．

文　献

[1] Twitter, Inc, "Twitter," https://twitter.com.

[2] M. Pennacchiotti and S. Gurumurthy, "Investigating topic models for social media user recommendation.," in *Proc. WWW2011*, 2011, pp. 101-102.

[3] 山本祐輔, "テレビ番組に対する意見を持つ Twitter ユーザのリアルタイム検出," 早稲田大学基幹理工学研究科, 2013.

[4] "TF-IDF で文章内の単語の重み付け," http://blog.takuti.me/2014/01/tf-idf/.

[5] 赤澤康幸, "コサイン類似度," http://www.cse.kyoto-su.ac.jp/g0846020/keywords/cosinSimilarity.html.

[6] "Twitter4J - A Java library for the Twitter API," http://twitter4j.org/ja/.

[7] Atilika, "Kuromoji," http://www.atilika.com/ja/products/kuromoji.html.

[8] 奥村学, "マイクロブログマイニングの現在," 電子情報通信学会技術研究報告, vol. NLC2011, no. 59, pp. 19-24, 2012.

[9] Z. Cheng, J. Caverlee and K. Lee, "You are where you tweet: A content-based approach to geo-location twitter users.," in *Proc. CIKM'10*, 2010, pp. 759-768 .