2N-5

Twitterのリスト機能を用いたユーザの特徴抽出

奥川 巧† 倉門 浩二†† 大石 哲也‡ 越村三幸‡‡ 藤田博‡‡ 長谷川 隆三‡‡
†九州大学工学部電気情報工学科 ††九州大学大学院システム情報科学府
‡九州大学情報基盤研究開発センター ‡‡九州大学大学院システム情報科学研究院

1 はじめに

近年，インターネット上でコミュニケーションを促進する場としてSNS(Social Network Service) が普及している．その中でもここ最近爆発的に利用登録者数を増大させ有名となったのがTwitterと呼ばれる，2006年7月に開始されたマイクロブログサービスである．ユーザはツイートと呼ばれる140字以内の文章を投稿する．また，あらかじめ登録したユーザのツイートを読むことができる．この登録をフォローという．

Twitterの利用の仕方として，他人をフォローすることにより知りたい情報を集めるというものがある．しかし，新たに誰かをフォローしようとしたとき，そのユーザが自分の知りたい情報を持っているか，その情報に関連したツイートをするかを一目で判断するのは難しい．現状として，ユーザがどのような人物であるか，何に興味があるか，などを知りたい場合は自己紹介の欄を見るか，過去のツイートを見るしかない．しかし，自己紹介に興味や趣味を書いてない人もおり，過去のツイートを遡って趣味などを判断できる文章を見つけるのも手間がかかる．ユーザの特徴が一目でわかるような機能があれば，このような手間をかけずに自分がフォローしたいユーザかどうかを判断することができる．

ところで，Twitterにはユーザをグループ化したリストを作る機能がある．リスト機能を使うと，特定のユーザのツイートだけを表示させることができる.「友人」,「bot」,「バイク好き」,「福岡の人」など様々なテーマでリストを作ることが可能である．つまり，リストにはそこに含まれているユーザの特徴が表れていると言える．

本研究ではTwitterのリスト機能を用いてユーザの特徴を抽出する手法を提案する．Twitterのリスト機能を用いた研究については，リスト名を用いてユーザの属性抽出や特徴語抽出を行った榊らの研究が挙げられる [1]．我々はリスト機能の内，特にリスト中のユーザとその自己紹介文に着目して特徴を抽出することを目指す．

2節では提案手法，3節では実験について述べる．

2 提案手法

リストとはあるユーザが他のユーザをカテゴリに分けたものと我々は考えた．リスト中のユーザには共通して属するカテゴリがあり，このカテゴリはユーザの特徴ともいえる．また，Twitterにおいてユーザ自身の情報が最も良く反映しているのは自己紹介の欄である．そこで，ユーザの特徴を抽出するには，そのユーザが含まれているリストの中の他のユーザの自己紹介文を見て共通点を見つけ出せばよい．

まず，特徴を抽出したいユーザが含まれるリストを取得し，それらのリストの中でそのユーザ以外のユーザの自己紹介文を取得する．次に，取得した自己紹介文から名詞を抜き出し，出現頻度の高い順にランキングする．このランキングの上位に位置するものがユーザの特徴を表すと考えた．

しかし，ユーザ一人に特徴が一つとは限らない．一つの特徴に関する語の全てが他の特徴に関する語より格段に頻度が高かった場合，この方法では一つの特徴に関する語だけが上位を占め，他の特徴を見つけることができない可能性がある．そこで，特徴を抽出したいユーザが含まれるリストをクラスタリングすることにした．リスト群を特徴ごとの集合に分けるためである．

クラスタリングを加えた手法について説明する．はじめに，特徴を抽出したいユーザが含まれるリストを取得し，ユーザを素性としてクラスタリングする．次に，クラスタリングによってできたリストの集合（クラスタ）がどのような特徴をもとに集まったかを特定する．まず，クラスタ内のリスト中のユーザの重み付けをし，クラスタにおけるユーザの重要度を決める．そして，ユーザの自己紹介文中の名詞を抜き出し，名詞にそのユーザの重み付けの値を与える．最後に，名詞を値の降順でランキングして，上位に位置する語を特徴とする．つまり，クラスタ内で重要なユーザの多くが自己紹介で使う名詞ほど特徴としてふさわしいということである．

以下にクラスタリング，重み付け，名詞のランキングについて詳しく述べる．

2.1 リストのクラスタリング

特徴を抽出したいユーザが含まれるリストをTwitterから取得し，リストの中のユーザを素性としてクラスタリングする．クラスタリングには潜在的ディリクレ配分法 (Latent Dirichlet Allocation， LDA) を用いる [2]．潜在的ディリクレ配分法はテキストマイニングの手法の一種である．文書を単語の集合と考えて，単語ごとにトピックを割り当て，文書集合の背後に潜む潜在的なトピック構造を推定するものである．ここでは，

文書をリスト，単語をそのリストに含まれているユーザとする．

2.2 クラスタごとのユーザの重み付けの設定

ユーザの重み付けには索引語重み付けのエントロピーを用いる．

クラスタの総数を n，全クラスタにおけるユーザ i の出現頻度を F_i とする．また，あるクラスタ j でのユーザ i の出現頻度を f_{ij} とし，ユーザ i のエントロピーを g_i とすると

$$g_i = 1 + \frac{1}{\log n} \sum_{j=1}^{n} \frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i} \quad (1)$$

で表わすことが出来る．

あるクラスタ j 内のユーザ i の重みは $f_{ij}g_i$ で表され，少数の特定のクラスタに頻繁に出現するユーザほど大きい値が与えられる．

2.3 名詞のランキング

クラスタごとに名詞をランキングし，クラスタごとの特徴語を抽出する．まず，クラスタ内の全ユーザの自己紹介文を Twitter から取得し，名詞のみを抜き出す．上で求めたユーザ i の重みを V_i とし，名詞 k がユーザ i の自己紹介文に含まれている場合 E_{ik} を 1，含まれていない場合 E_{ik} を 0 とする．また，クラスタ内のユーザ数を m とすると，名詞 k に与えられる評価値 S_k は $\sum_{i=1}^{m} E_{ik}V_i$ となる．S_k の値について降順に並べることで名詞をランキングする．

3 実験

3.1 実験方法

今回は女子プロレスラーで元参議院議員の神取忍氏（ユーザ名：kandorishinobu）を提案手法に沿って特徴を抽出する．

クラスタリングにおいては，クラスタ数を 2 から 4 まで試し，尤度が一番高かったクラスタ数を採用する．

自己紹介の名詞は一般名詞，固有名詞，Wikipedia の見出し語のみを抜き出している．また，事前に 2 万ユーザの自己紹介文から名詞を抜き出し，それを出現数でランキングしたものの上位の語の中から特徴を表す語になりえないものを主観で判断してピックアップしており，実験での名詞のランキングにおいてピックアップした語が含まれていた場合はそれを除いている．

適切にクラスタに分かれているかということに関する評価方法として，結果として出た各クラスタのランキングの上位 10 語を手動により同じクラスタ数に分け，精度と再現率を求めた．

次に，結果として表 1 にクラスタごとの名詞のランキングと，クラスタごとの精度と再現率を示す．表の太文字の名詞は実行結果と手動で分類した結果，同じクラスタであったものである．

3.2 実験結果

表 1: 神取忍氏のクラスタごとの名詞のランキングとその精度と再現度

順位	クラスタ 1	クラスタ 2
1	**プロレス**	**衆議院議員**
2	**プロレスラー**	**自民党**
3	日本	**参議院議員**
4	ブログ	**民主党**
5	**出身**	**政治**
6	**女子プロレスラー**	**日本**
7	**女子プロレス**	**参議院**
8	**格闘技**	**議員**
9	**プロレス団体**	**政策**
10	**世界**	**生まれ**
精度	0.8	1.0
再現率	1.0	0.83

クラスタ 1 にはプロレス関係が，クラスタ 2 には議員に関連した語がランキングの上位に位置し，神取忍氏の特徴が抽出できているといえる．また，クラスタごとの再現率と精度も高く，特徴ごとに明確なクラスタに分けられているといえる．

4 おわりに

本稿では Twitter のリスト機能からユーザの特徴を抽出する手法を提案した．今回の実験では精度，再現率共に高い結果であったが，ユーザによっては多くのクラスタに分かれても，クラスタ同士で似た単語が多く含まれ，精度，再現率共に低い場合もある．今後の課題としては，クラスタリングにおいて違う素性を採用することが挙げられる．

謝辞 本研究は科研費 (21500102) の助成を受けたものである．

参考文献

[1] 榊 剛史, 松尾 豊, “ ソーシャルブックマークとしての Twitter のリスト機能の応用 ”*The 24th Annual Conference of the Japanese Society for Artifcial Intelligence*, 3B3-2, 2010.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. “ Latentdirichlet allocation ” *Journal of Machine Learning Resarch*, 3:993-1022, 2003.