3P-2

マイクロブログを利用した Webサイトの閲覧者像と地域性の推定

西田 綾佑 † 高崎 隼 †† 平田 紀史 †† 白松 俊 †† 大囿 忠親 †† 新谷 虎松 ††
名古屋工業大学工学部情報工学科 † 名古屋工業大学大学院工学研究科情報工学専攻 ††

1 はじめに

マイクロブロギング・サービス (Microblogging Service，通称「マイクロブログ」) とはウェブログの一種であるが，近年莫大なユーザの増加が見られたサービスである．本研究では，Twitter[1]のユーザ登録の際に記入されるプロフィール欄と，Twitter の機能のひとつであるリツイートによる即時性・伝播性のある情報の拡散に着目する. 最新の Web ページに関して言及した即時性・伝播性にすぐれた Twitter のユーザ群の解析を行うことで，最新の Web ページの閲覧者像の推定をはかる．

さらに，Web ページが分布している地域を推定できれば，ユーザが持つ関心の傾向を地域の Web ページ毎に分析するために有用である．Wikipedia 日本語版[2]の全記事データを利用した入力文章を地方自治体へ割り振る分類器を作成し，これにより Web ページを分類し，地域性を推定する．利用する Wikipedia の記事データが 145,668 記事と膨大なため，効率化のため，大規模データの分散応用システムである Apache Hadoop[3]を利用し，そのシステム上で動作する機械学習ライブラリである Apache Mahout[4]による学習を行う．

大手ニュースサイトやブログの中には，ソーシャルブックマークサービスや SNS と連携したリンク投稿機能が設置されているものが存在する．これらはソーシャルボタンと呼ばれ，閲覧者の利用する外部サービスと連携して，関心を持った Web ページを手軽に友人・知人と共有することができる機能である．また，Twitter クライアントの中には，ブラウザから閲覧中の Web ページ URL を投稿する機能が備わっているものも少なくない．Twitter ユーザがある Web ページの URL を含む発言を取得できるということは，そのユーザが Web サイトを閲覧している可能性が高く，Web サイトの内容に興味がある可能性も高いと考えられる．

2 関連研究

岩木ら [1] は，Twitter の発言履歴とユーザ同士のつながりを元に，有用な記事の発見を行っている．あるクエリに対する Twitter 検索により，クエリがどのようなものであるかの推定を行うとともに，発言内容から感性辞書の作成を行った．さらに，ユーザ同士のつながりを follow しているかどうかからの判定ではなく，発言に対する返信とその内容からユーザ近似度の算出を行った．我々の研究では，クエリがどのようなものかを考慮せず，クエリを含む発言をしたユーザの集合に対して，解析を行っている点で異なる．榊ら [2] は，Twitter のリスト機能をユーザのタグ付けであると考え，リスト名によるユーザ属性の抽出と，同一リスト名に限定した特徴語の抽出を行った．リスト名というクエリを与えることで，クエリに関係するユーザの集合を取得することができる．Owen ら [3] は，RSS フィードのフィルタリングに Twitter を利用し，RSS から有益な記事を抽出することを目指した．多くの RSS フィードを購読している場合，有益な記事の発見には多くの時間を費やす．そこで，マイクロブログのタイムラインを利用したフィルタリングを行い，有益な記事の発見へとつなげた．発言履歴の解析を行っているが，我々の研究ではプロフィールを利用した解析を行う．

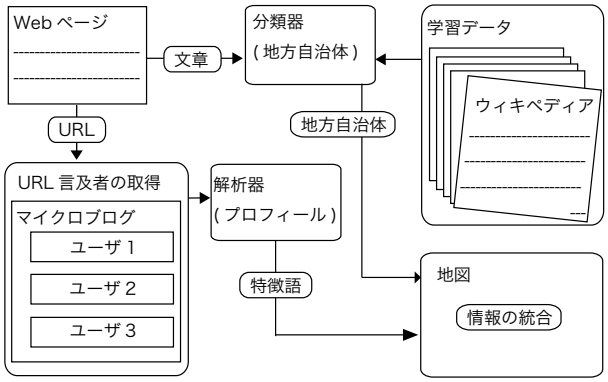

図 1: システム概要図

3 解析手法

3.1 マイクロブログを利用した閲覧者像の解析

ユーザごとに登録されたプロフィール情報には，ユーザの特徴を表す語が多く含まれることが期待される. そこで，Web ページ p の URL を含む発言をしたユーザ集合を U_p とし，特徴語の候補を s とする．語 s をプロフィールに含み，U_p に含まれるユーザの数を $DF(s,U_p)$ とする．また，尺度 $f(s)$ によって語 s を降順にソートした場合の s の順位を $rank\_s(f(s))$ とする．このとき，Web ページ p の閲覧者を表す特徴語集合 $S(p)$ を以下の式により求める．

$$S(p) = \{s; rank_s(\log DF(s,U_p)) \leqq \theta \ \} \quad (1)$$

本稿では $\theta = 11$ とし，11 位までの語を p の閲覧者の特徴語と見なす．

3.2 Wikipedia を利用した地域性の推定

閲覧者像推定対象となった Web ページについて，地域性の推定を行う．Web ページの文章を分類器へ入力し，文章の内容から Web ページを各地方自治体ごとへ分類する．Wikipedia に登録されている記事データについて形態素解析し，文書頻度を記録したデータ，145,668 記事についてのデータを利用し，コーパスを作成することでナイーブベイズ分類器を試作する．Wikipedia のある記事 d と地方自治体のカテゴリ C について，確信度と単語正規率，頻度，総単語数を元に分類器を作成することを考える．記事 d がカテゴリ C に分類される確率 $p(C|d)$ を，カテゴリの確率 $p(C)$ と単語生起確率 $p(w|C)$ から計算する．$confi(d,C)$ が記事 d がカテゴリ C に入るかどうかの確信度，$tf(w,d)$ を記事 d 中での単語 w の頻

Analysis of Twitter Users and Locations of Web pages.

Ryosuke NISHIDA, Shun TAKASAKI, Norifumi HIRATA, Shun SHIRAMATSU, Tadachika OZONO, and Toramatsu SHINTANI

Dept. of Computer Science and Engineering, Nagoya Institute of Technology. Dept. of Computer Science and Engineering, Graduate School of Enngeering, Nagoya Institute of Technology.

[1] http://twitter.com/

[2] http://ja.wikipedia.org/

[3] http://hadoop.apache.org/

[4] http://mahout.apache.org/

度，$totaltf(d)$ を記事 d に含まれる総単語数とする．CNB で補集合を使うことにより対処，$p(w|C)$ ではなく $p(w|\bar{C})$ を使う．式 2 がナイーブベイズ分類器に使用する式である．

$$p(w|C) = \frac{\sum_d confi(d,C) * tf(w,d)}{\sum_d confi(d,C) * totaltf(d)} \quad (2)$$

最も単純なナイーブベイズ分類器では，正解文書が少ないカテゴリには分類されにくいという問題点がある．そこで，データが多ければ多いほど精度が上がり，逐次学習可能である補集合ナイーブベイズ分類器を利用することを考える．補集合である $confi(d,C) = 0$ の記事 d が数多く存在するため，$confi(d,C)$ が 1 以上の記事は無視しても影響ない．そこで，以下の式 3 を用いて演算を行う．

$$p(w|\bar{C}) = \frac{\sum_{d \notin C} tf(w,d)}{\sum_{d \notin C} totaltf(d)} \quad (3)$$

補集合ナイーブベイズ分類器において，文書 d 分類時の地名 C のスコアは，式 4 となる．

$$score_d(C) = logp(C) - \sum_{w \in d} TF(w,d) logp(w|\bar{C}) \quad (4)$$

これにより，ある記事に対して自治体数. 47 都道府県, 1,788 市町村, 190 区の 2,025 件の階層的な分類を行う．

表 1: 解析対象の Web ページ

Web ページタイトルと URL
【A】【Apple ホームページ】[1]
【B】【Apple App Store】[2]
【C】【ダルビッシュ最速５億円…２４歳、イチロー超え】[3]
【D】【新潟で新米６００キロ盗まれる】[4]
【E】【漁場圧迫海人怒り 水域外で米軍演習通知】[5]
【F】【東京駅線路から火、ダイヤ乱れ８万２千人に影響】[6]

表 2: Web サイトを言及したユーザ群の特徴語

【A】		【B】		【C】	
特徴語	出力値	特徴語	出力値	特徴語	出力値
Mac	10.411	音楽	10.151	ニュース	8.197
デザイン	9.617	iPhone	10.008	フォロー	8.197
Apple	8.752	IT	9.841	リプライ	7.727
IT	8.752	0	9.746	情報	7.727
iPhone	8.465	仕事	9.523	配信	7.727
アプリ	8.465	会社	9.289	音楽	7.727
音楽	8.465	Apple	9.389	スポーツ	7.504
カメラ	8.282	Mac	9.389	Google	7.216
東京	8.059	Web	9.389	ツイート	7.216
京都	8.059	iPad	9.389	ビジネス	7.216
大学	8.059	デザイン	9.389	人	7.216

4 解析結果

表 1 のニュース記事について.【A】,【B】,【C】に関して閲覧者像の解析.【D】,【E】,【F】に関して地域性の解析を行った．閲覧者像の解析結果を表 2 に示す．Mac，Apple 等，Web ページから推測できる単語が多く抽出され，Web ページの内容と Twitter の嗜好が一致していることが推測できる．ほかに，音楽，カメラといった特徴語を抽出することができた．これらは Web ページの内容からでは推測できない特徴語であり，Web ページ内のみの解析からでは得られない閲覧者の特徴をマイクロブログから取得できたと考えられる．また，ニュース記事についてナイーブベイズ分類器による分類を行った結果を表 3 に示す．分類結果について，正解となる地方自治体を太字にて表示しておく．全地方自治体中の上位 10 件のランクにいずれの記事も分類がされている．大阪市，横浜市，名古屋市等，比較的都市部となる地方自治体に多く分類される結果となった．これは，学習に利用した Wikipedia のデータの量に偏りがあるためだと考えられる．

[1] http://www.apple.com/jp/
[2] http://www.apple.com/jp/mac/app-store/
[3] http://www.yomiuri.co.jp/sports/npb/news/20110106-OYT1T00655.htm
[4] http://sankei.jp.msn.com/region/chubu/niigata/100919/ngt1009190258000-n1.htm
[5] http://www.okinawatimes.co.jp/article/2011-01-06_13439/
[6] http://www.yomiuri.co.jp/national/news/20110106OYT1T00881.htm

図 2: インターフェース

表 3: 地方自治体への分類結果

分類結果					
【D】		【E】		【F】	
順位	自治体	順位	自治体	順位	自治体
1	大阪市	1	大阪市	1	大阪市
2	横浜市	2	横浜市	2	横浜市
3	**新潟市**	3	名古屋市	3	名古屋市
4	神戸市	**4**	**那覇市**	4	神戸市
5	名古屋市	5	京都市	**9**	**千代田区**

5 インターフェースの試作

地図上の自治体に割り振られた Web ページが地図上にマッピングされる．地図上に表示される吹き出しには，閲覧者像となる特徴語のリストと，特徴語を多く含んだユーザを表示している．

6 まとめ

本研究では，マイクロブログのプロフィール情報を用いた Web サイトの読者像の推定と，マイクロブログと Web サイトの内容を用いた位置情報の推定を行い，推定結果を地図上に表示するインターフェースを試作した．今後の課題として，補集合ナイーブベイズ分類器の精度の向上．応用として，閲覧者像と地方自治体への分類結果を利用した記事の推薦システムへの利用が考えられる．

謝辞 本研究の一部は，総務省による戦略的情報通信研究開発推進制度 (SCOPE) の支援を受けて行われた．

参考文献

[1] 岩木 祐輔，アダム ヤトフト，田中 克己，”マイクロブログにおける有用な記事の発見支援”，データ工学と情報マネジメントに関するフォーラム 2009 A6-6．2009.

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

[3] Owen Phelan, Kevin McCarthy, and Barry Smyth. ”Using twitter to recommend real-time topical news.” In Proceedings of the third ACM conference on Recommender systems, pp. 385 - 388, New York, New York, USA, 2009. ACM.