Twitter を用いた個人の特徴抽出と その情報提供ポータルサイト構築への応用

近藤 直人 [1]　　内田 理 [2]
[1] 東海大学大学院工学研究科情報理工学専攻
3bdrm004@mail.tokai-u.jp
[2] 東海大学情報理工学情報科学科
o-uchida@tokai.ac.jp

1. はじめに

近年，ユーザの興味を分析して有益な情報（パーソナライズされた情報）を提供するサービスが多数展開されている．例えば Gunosy[1]は，Twitter[2]や Facebook[3]への投稿内容に基づいてユーザの興味を推定し, 推定結果に基づいて情報を提供している．しかし, Gunosy が提供する情報はニュースサイトの記事やブログ記事等であり，画像や動画，音楽といった情報は提供されていない．SmartNews[4]はスマートフォン向けアプリで展開されている情報提供サービスである．提供する情報は，Twitter で数多くツイートされているニュース記事から選ばれている．Twitter アカウントを連携することにより，ユーザの投稿内容が情報提供に反映されるが，原則として提供される内容はニュース記事のみである．しかし，パーソナライズされて提供される情報は，ニュース記事に限定せず，動画や音楽，アプリケーションなどユーザのニーズに応じて様々なものであることが望ましいと考える．そのような背景のもと，我々はパーソナライズされた様々な情報を提供するポータルサイトの構築を目指している. Yahoo! [5]ではポータルサイトの構成をカスタマイズするサービスを提供しているが，Twitter などから個人の特徴を推定して画面構成を自動的にカスタマイズしたり，提供する情報を最適化する機能は有していない．本研究では，Twitter の投稿内容からユーザ個人の興味や特徴を推定し，はてなブックマーク[6]，及びニコニコ動画[7]からパーソナライズされた記事や動画を提供するポータルサイトの構築を行う．

2. 関連研究

胡ら[8]は Wikipedia[9]を用いて Twitter ユーザの関心分野の抽出をしている. Wikipedia はユーザによって編集や追加が可能な Web 百科事典の 1 つであり，Wikipedia 日本語版においては 2014 年 1 月現在で約 90 万の単語が登録されている．また，各単語はカテゴリの情報を持っており，9 万種類のカテゴリが存在しており，これらのカテゴリは階層構造となっている．胡らの手法は曖昧性のあるカテゴリ情報を取得出来ているが，ユーザの関心分野を抽出するための計算方法には改善の余地があると考えられる．

3. 提案手法

3.1. 概要

Twitter はリアルタイム性に大きな特徴を有する[10]ため，対象ユーザのツイートを解析することはユーザの「今」の興味や特徴を知るために有益と考えられる．そこで本研究では，個人の特徴抽出に Twitter を利用することとした．提案手法の概要を図 1 に示す．まず，対象ユーザの最新ツイートを 200 件分取得する．取得するのは，ツイート本文とツイート中に含まれる URL である. ツイート文を用いてユーザの特徴を推定する．推定されたユーザの特徴と，ツイート中に含まれる URL を用いて，ユーザにとって有益である情報を収集する．最後に，収集された情報をポータルサイト上に出力する．

3.2. ツイートの取得

対象ユーザの最新ツイートを 200 件分取得する．本研究では，ツイートの取得には Twitter API v1.1[11]を利用した．ツイート本文の他に，ツイート中に含まれる URL を抽出することによって, ユーザのツイートにおけるサイトの出現頻度を導出する．なお，以降の処理の前処理として，ツイート本文から改行やハッシュタグ，ユーザ ID などの部分を削除する．

3.3. 特徴の推定

ツイート文には通常の形態素解析器では未知語と判断される単語が多いという特徴があるため，本研究では，ユーザの特徴推定にはてなキーワード[12]を用いる．はてなキーワードとは Wikipedia と同様に, ユーザによって追加や編集が可能な Web 百科事典である．2014 年 1 月現在，約 40 万の単語が登録されている. はてなキーワードには 20 種類のカテゴリ（表 1）があり，各キーワードに 1 つ以上のカテゴリが付与されている. 3.2.で取得した対象ユーザの最新 200 ツイートの文中に含まれるはてなキーワードに登録された単語を抽出し，出現した単語の回数と属するカテゴリの頻度を求める．なお，はてなキ

ーワードの半数が“一般”のカテゴリに属しているが，特徴推定には不向きである場合がおおいため，提案手法では除外することとした．また，“はてな”と“はてなダイアリークラブ”のカテゴリは，はてなキーワードのサービスの特有のカテゴリであるため，これらも除外することとした．“一般”，“はてな”，“はてなダイアリークラブ”を除く 17 種類のカテゴリの頻度を求め，ユーザの特徴とする．あるユーザに対して導出したカテゴリ頻度の上位 5 件を表 2 に示す．また，それぞれのカテゴリに属し出現回数が多い単語の上位 5 件を表 3 から表 6 に示す（“音楽”カテゴリは全ての単語の出現回数が 1 であったため掲載は省略した）．

次に，抽出した URL のドメイン名より，ユーザがリンク付きツイートをしたサイトの頻度を求める．サイト出現頻度の例を表 7 に示す（表 7 は上位 5 件のみを示している）．なお，診断メーカー[13]とは，Twitter でつぶやくことができる診断をユーザが作成することができるサイトである．また，Pixiv[14]とはイラスト投稿サイトの一つであり，ニコニコ静画[15]はニコニコ動画のサービスの 1 つで，Pixiv と同様にイラスト投稿サイトである．この例から分かる通り，個人の特徴抽出にツイート中の URL は有益であると考えられる．

3.4. ユーザに提供する情報の収集

対象ユーザのツイート文 200 件から取得した，ユーザの特徴・単語の出現頻度に基づいてユーザに提供する情報を選択する．本研究では，はてなブックマーク，及びニコニコ動画から情報を提供することとした．両サービスとも API が提供されており，その API を用いて情報を収集する．ニコニコ動画は，日本国内において最大級の動画投稿サイトであり．2013 年 9 月で，ユーザ数は 3600 万人を超えている．ニコニコ動画内の各動画には，動画の特徴を表すタグ情報が付与されている．

はてなブックマークから情報を収集する際には，最も出現頻度が高いカテゴリと，そのカテゴリに属しており出現頻度の高い単語をクエリとして検索する．有用な記事を取得するために，10 人以上のユーザによってブックマークされた記事のみを新着順に取得することとした．

ニコニコ動画から情報を収集する際には，ニコニコ動画の 1 時間ごとのランキングを取得し，出現頻度の高いカテゴリに属する動画をマイリスト数の上位順に取得する．マイリストとは，ニコニコ動画におけるブックマーク機能であり，マイリストされた件数は各動画に情報として付与されている．ニコニコ動画において各動画に登録されているタグと，はてなキーワードのタグには違いがあるため，ニコニコ動画から情報を取得する際には，はてなキーワードのカテゴリを表 8 の様に置き換えて検索することとした．なお，該当カテゴリがない場合には，全てのカテゴリを合算したランキングで検索する．

3.5. 情報の提供

構築するポータルサイトの例を図 2, 3 に示す．取得した各サイトの情報を，カテゴリの頻度と各カテゴリにおける単語の出現頻度，ユーザのツイート中に含まれるサイトの頻度に基づいて出力する．本研究では，カテゴリの頻度が高く，そのカテゴリに属する単語が含まれる情報を提供されやすくした．

トップページでは図 2 にように取得された全ての情報を出力する．ユーザがツイートしたサイトの頻度が高いサイトの情報がページの上部に配置されるようになっている．また，ページの上部のボタンをクリックすることにより，取得する情報を選択できるようになっている（図 3）．これにより，ユーザが必要とする情報を選択しやすくなると考えられる．また，更新ボタンを押すことにより，ユーザのツイートが再取得されるため，ユーザの最新の特徴を抽出することができる．また，各サイトの情報も更新の度にアップデートされ，最新の情報を提供することが可能である．

4. まとめと今後の課題

本研究では，Twitter を用いて個人の特徴を推定し，その特徴を用いてパーソナライズされた情報を提供するポータルサイトを提案した．本研究では，はてなキーワードのカテゴリを利用して個人の特徴抽出を試みたが，カテゴリの種類数が 20 種類（実際に使用したのは 17 種類）と少ないため，ユーザの特徴推定の精度に問題が生じる可能性が考えられる．そこで今後，Wikipedia のカテゴリの利用を検討する．Wikipedia のカテゴリが階層構造を形成していることを利用すれば，より正確性の高いユーザの特徴推定が可能になると考えている．さらに，LDA（Latent Dirichlet Allocation）[16][17]のようなクラスタリングの技術を用いることで多様なカテゴリやタグにおいても対応できるようになると考えている．また，ポータルサイトの実装と評価を早急に実施したい．

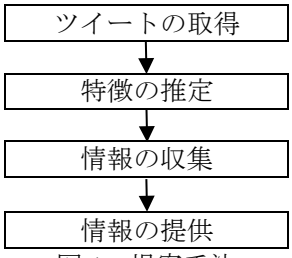

図 1．提案手法

表 1　はてなキーワードカテゴリ一覧

一般	読書
音楽	映画
ウェブ	動植物
テレビ	アニメ
ゲーム	マンガ
社会	地理
スポーツ	アイドル
食	コンピュータ
サイエンス	アート
はてな	はてなダイアリークラブ

表 2　カテゴリの頻度例

カテゴリ	頻度[%]
ゲーム	16.3
テレビ	9.7
音楽	8.7
ウェブ	8.7
食	8.7

表 3　“ゲーム”カテゴリに属する単語

キーワード	出現回数
オーブ	6
艦これ	4
PC	3
ビビ	3
麻雀	2

表 4　“テレビ”カテゴリに属する単語

キーワード	出現回数
RT	26
アキバレンジャー	1
ホウセイマイフレンド	1

表 5　“ウェブ”カテゴリに属する単語

キーワード	出現回数
ニコ生	5
(´・ω・`)	3
ry	3
アメーバ	2
テンプレ	2
ワロタ	2

表 6　“食”カテゴリに属する単語

キーワード	出現回数
マクドナルド	3
いちご	2

図 2.　ポータルサイト例（全てのサイト）

図 3.　ポータルサイト例（ニコニコ動画）

表 7　サイトの出現頻度の例

サイト	出現回数
Twitter	16
ニコニコ動画	10
診断メーカー	4
Pixiv	2
ニコニコ静画	1

表 8　カテゴリの置き換え

はてなキーワード	ニコニコ動画
音楽	音楽
動植物	動物・自然
テレビ	エンターテイメント
アニメ	アニメ
ゲーム	ゲーム
マンガ	アニメ
社会	政治
スポーツ	スポーツ
アイドル	エンターテイメント
食	料理
コンピュータ	ニコニコ技術部
サイエンス	科学
アート	エンターテイメント
その他	該当なし

参考文献

[1] Gunosy, http://gunosy.com/

[2] Twitter, https://twitter.com/

[3] Facebook, https://www.facebook.com/

[4] SmartNews, https://www.smartnews.be/

[5] Yahoo! JAPAN, http://www.yahoo.co.jp/

[6] はてなブックマーク, http://b.hatena.ne.jp/

[7] ニコニコ動画, http://www.nicovideo.jp/

[8] 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”, 信学技報, Vol.113, No.213, pp.17-21, 2013.

[9] Wikipedia, http://www.wikipedia.org/

[10] 奥村学, “ソーシャルメディアを対象としたテキストマイニング, 電子情報通信学会, Vol.6, No.4, pp285-293, 2013.

[11] Twitter Developers, https://dev.twitter.com/

[12] はてなキーワード, http://d.hatena.ne.jp/keyword/

[13] 診断メーカー, http://shindanmaker.com/

[14] Pixiv, http://www.pixiv.net/

[15] ニコニコ静画, http://seiga.nicovideo.jp/

[16] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp.993-1022, 2003.

[17] 山本修平, 佐藤哲司, “LDA を用いた実生活 Tweet の二段階抽出法”, DEIM Forum 2013, 2013.