Wikipedia のカテゴリ階層を利用した Twitter ユーザのカテゴライズ

放地 宏佳　鶴田 雅信　酒井 浩之　増山 繁
豊橋技術科学大学
{houchi,tsuruta}@la.cs.tut.ac.jp,{sakai,masuyama}@tut.jp

1　はじめに

Twitter[*1] は世界的に急成長を遂げているマイクロブログサービスであり，世界で 1 億 4,500 万ユーザ以上[*2]，日本でも 500 万ユーザ以上[*3]の登録ユーザが存在する．Twitter はこのような大規模ユーザで形成されるにも関わらず，SNS やブログといった他の CGM(Consumer Generated Media)[*4] に見られるユーザ分類機能が充実していない．例えば mixi[*5] には，コミュニティや mixi 同級生といったユーザ分類機能が存在しており，この機能を用いることでユーザ間の交流の促進 (ユーザ推薦) を行っている．

現在 Twitter には「おすすめユーザ機能 [*6] 」と呼ばれる，各カテゴリ別のおすすめユーザを推薦する機能があるが，日本人ユーザ向けのおすすめユーザ機能におけるカテゴリは「Web, キャラクター，ツイッター，ビジネス，ミュージシャン，政治，文化・スポーツ，芸能人」の 8 種類であり，カテゴライズの対象となるユーザ数も全部で 53 ユーザしか存在しない[*7]．これは Twitter の全体ユーザ数に対して，非常に少ないカテゴリ数，および，カテゴライズの対象となるユーザ数であるが，カテゴリ分けは Twitter の新規ユーザが初期にフォローするユーザを発見するために，有用な情報であると考えられる．

そこで本研究では，カテゴリ数，および，カテゴライズの対象となるユーザ数の増加を目的とし，自動的に Twitter ユーザをカテゴライズし，ユーザ分類する手法を提案する．提案手法では，カテゴライズの対象となるユーザのツイート[*8]を，Wikipedia のカテゴリに対応させることにより，ユーザのカテゴライズを行う．カテゴライズされた結果と Wikipedia のカテゴリ構造を用いることにより，階層的なカテゴリにユーザを割り当てる形のユーザ推薦機能を提供することができる．

1.1　Wikipedia について

Wikipedia は誰でも編集可能な巨大なウェブ百科事典である．新語が即時に反映され，また，すべての語に対してその語の属するカテゴリが存在する．各カテゴリは 1 つ以上の上位カテゴリに属する構造になっている (図 1)．

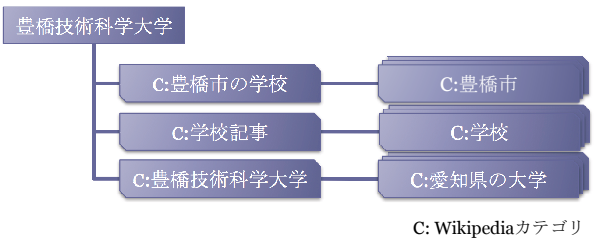

図 1: Wikipedia のカテゴリ階層

1.2　関連研究

Twitter ユーザのカテゴライズの研究として，榊ら [1] がリスト機能[*9]を用いたカテゴライズを行っている．この研究では，リスト機能を使いユーザをまとめ上げる際に付与されるリストの名称に注目し，自動的にカテゴライズを行っている．自動的にカテゴライズ

[*1] http://twitter.com/
[*2] http://www.blogherald.com/2010/09/03/twitter-reaches-125-million-users-300-thousand-apps/
[*3] http://japan.cnet.com/news/biz/20408610/
[*4] 消費者生成メディアと呼ばれ，インターネットなどを活用して消費者がメディア内容を生成していくもの
[*5] http://mixi.jp/
[*6] http:/twitter.com/#!/who_to_follow/interests
[*7] 2010 年 12 月 17 日現在
[*8] Twitter で投稿されたメッセージのこと
[*9] 各ユーザが自分と関わりのあるユーザ群をまとめ上げる機能

を行う点で本研究と類似しているが，分類方法，分類対象データともにリストを利用する点で本研究とは異なる．田中ら [2] はフォロー関係を用いて有用なユーザを分類する研究 (また，ユーザの分類情報からツイートの分類を行う研究) を，Weng ら [3] は Twitter のフォロー関係を用いて，影響度の高いユーザを発見するための手法を提案している．これらの研究は，ユーザを分類する点で本研究とは類似しているが，フォロー関係を用いる点という特徴がある．本研究では，ツイートの内容に注目しカテゴライズを行う．

Wikipedia の情報を用いた研究として，Banejee ら [4] が RSS，また，ATOM といったフィードのサマリーのような非常に短いテキストを高精度でクラスタリングするため，Wikipedia の情報を用いて素性の拡張を行う手法を提案している．短いテキストを分類する上での情報源として Wikipedia を用いる点で本研究と類似している．本研究では Wikipedia のカテゴリ情報を利用する．

2 提案手法

提案手法は，ツイートから得られたキーワードが属するカテゴリにユーザが属するという考えに基づき，そのキーワードをよくツイートするユーザをそのカテゴリに割り当てる．提案手法の概要を図 2 に示す．

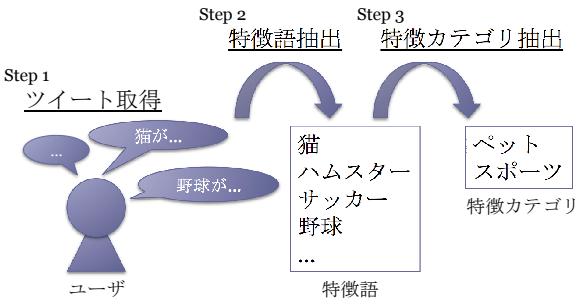

図 2: 提案手法の概要

図 2 の Step 1 では， Twitter REST API [*10] を利用し， ユーザのツイートを取得する．Step 2 では， ユーザのツイートから,「一般的ではなく，カテゴライズの対象となるユーザがよく利用する語」である特徴語を抽出する．Step 3 では， 特徴語と Wikipedia のカテゴリ階層を用いて,「一般的ではなく，具体的な意味をもつカテゴリ」である特徴カテゴリを抽出する．

[*10]本研究におけるユーザのツイートの取得には，statuses/user_timeline を用いた．
http://dev.twitter.com/doc/get/statuses/user_timeline

2.1 前処理

Twitter REST API (statuses/user_timeline) から取得したユーザのツイート集合，および，Wikipedia ダンプデータ[*11]に対して以下に示す前処理を行う．

- 特徴語，および，特徴カテゴリの頻度データの作成 (節 2.2 Step 2.4，および，節 2.3 Step 3)．
- Wikipedia ダンプデータを用いて，Wikipedia に掲載されている全ての記事名 (Wikipedia に掲載されている語，および，カテゴリ) 集合の作成．
- Wikipedia ダンプデータを用いて，Wikipedia に掲載されている全ての記事に対して上位カテゴリ名 (その記事が属しているカテゴリ名) 集合の作成．

2.2 特徴語集合の抽出

特徴語の定義を「一般的ではなく，カテゴライズの対象となるユーザがよく利用する語」として，特徴語集合の抽出を行う．特徴語集合の抽出のプロセスを以下に示す．

特徴語集合抽出アルゴリズム

Step 1 単一ユーザの今までのツイートを取得する．

Step 2 各ツイートに対して Step 2.1 から Step 2.4 の処理を行う．

Step 2.1 単一のツイートに対して，ツイートの正規化を行う (文頭の@username, 文中の RT,QT 以降の文，URL，ハッシュタグを取り除く)．

Step 2.2 Step 2.1 で得られた正規化されたツイートに対して，Unicode 正規化 (NFKC) を行う．

Step 2.3 Step 2.2 で得られた正規化されたツイートに対して，Wikipedia の記事名と一致する文字列 (以下「語」とする) を全て抽出し，出現回数 $count_w$ をカウントする．

Step 2.4 Step 2.3 で得られた語を割り当てられたユーザの総数を語の出現頻度 $freq_w$ とする (特徴語頻度データの作成)．

Step 3 Step 2.3 で得られた語に対して，$count_w > 2$, $1/freq_w > 0.005$ を満たす語をユーザの特徴語とする． □

[*11]http://download.wikimedia.org/jawiki/20101018/

2.3 特徴カテゴリ集合の抽出

特徴カテゴリの定義を「一般的ではなく，具体的な意味を持つカテゴリ」として，特徴カテゴリ集合の抽出を行う．特徴カテゴリ集合の抽出のプロセスを以下に示す．

特徴カテゴリ集合抽出アルゴリズム

Step 1 (節 2.2) で得られた各特徴語に対して，上位カテゴリ名集合を用い，その語の最上位カテゴリまでのパス集合を取得する (例えば，図 3 で特徴語「ネコ」に注目した場合，[ネコ,C:ネコ,C:ネコ科,..., 最上位カテゴリ],[ネコ,C:ネコ, ペット,..., 最上位カテゴリ],[ネコ,C:ペット,..., 最上位カテゴリ])．なお，ここで用いられるパスはある記事名を始点とし，上位カテゴリへの向きのみを持つ．

Step 2 Step 1 で得られた全ての特徴語の全ての最上位カテゴリまでのパスに対して，同一のカテゴリ (共通カテゴリ) が存在する部分を取得する (例えば，図 3 において [C:ペット])．

Step 3 Step 2 で得られた共通カテゴリを割り当てられたユーザの総数をカテゴリの出現頻度 $freq_c$ とする (特徴カテゴリ頻度データの作成)．

Step 4 Step 2 で得られた共通カテゴリに対して，最上位カテゴリからのパスの大きさ len_c, 同一共通カテゴリの数 $count_c$, カテゴリの出現頻度 $freq_c$ とし，$len_c/count_c > 2$, $1/freq_c > 0.005$ を満たす共通カテゴリをユーザの特徴カテゴリとする．□

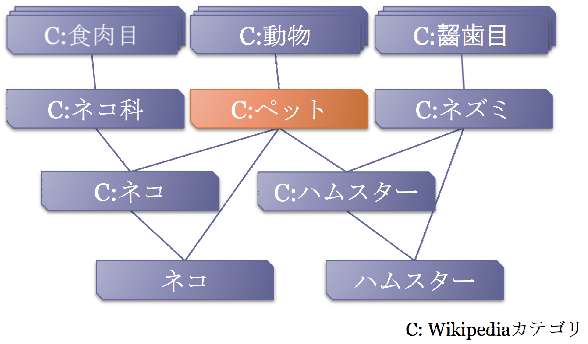

図 3: 共通カテゴリの抽出

3 評価実験

提案手法で得られた特徴カテゴリがユーザを分類するのに適切なカテゴリであるか否か調べるため，評価実験を行った．実験は，Twitter Streaming API (statuses/sample) [*12]を用いて 2010 年 11 月 23 日に発言を行った日本人ユーザをランダムに 1,000 ユーザ取得し，その後，ランダムに選択した 20 ユーザ (29,631 件の発言) に対して行った．各ユーザの発言数の最大は 2,000 件とした．

評価実験では, まず，提案手法によって出力された「特徴カテゴリ」,「特徴語から特徴カテゴリへのパス」, および,「特徴語を抽出したツイート」を評価者に提示する．評価者は提示された情報に対して，[特徴カテゴリ, 特徴語から特徴カテゴリへのパス], [特徴語から特徴カテゴリへのパス, 特徴語] の関係が適切であるか否か，特徴カテゴリがユーザのカテゴライズとして適切であるか否かを人手で判断した．評価項目が適切であるか否かを判断するため,「関係が適切である (取得された特徴カテゴリについてのツイートをしている)」「関係が不適切である (取得された特徴カテゴリとは異なるツイートをしている)」「選定された特徴カテゴリはユーザ分類に適さない」の 3 通りに分類した．評価者は工学系大学生 1 名である．

4 実験結果

評価実験を行った結果を表 1 に示す.「適切である特徴カテゴリ」の正解率は 0.49 であり,「ユーザ分類において不適切である特徴カテゴリ」を正解とした正解率は 0.78 であった．

表 1: 実験結果

取得された特徴カテゴリ数	88
適切である特徴カテゴリ数	43
不適切である特徴カテゴリ数	26
ユーザ分類において不適切である特徴カテゴリ数	19

5 考察

「適切である特徴カテゴリ」と判断されたものは,「スポーツ」や「コンピュータ」といったカテゴリが

[*12]http://dev.twitter.com/pages/streaming_api_concepts#sampling

多く，そのツイートには「サッカー」「野球」といった直感的にカテゴリが分かりやすい単語が含まれていた.「不適切である特徴カテゴリ」と判断されたものは,「物理学」「心理学」といった，専門用語が日常生活でも使われるカテゴリが多かった.「ユーザ分類において不適切である特徴カテゴリ」と判断されたものは,「英語の男性名」「各年のコンピュータゲーム」「言語別の語句」といった，Wikipedia 上の分類のために作られたカテゴリが多かった.

「物理学」という特徴カテゴリが取得された特徴語の例として反射，振動が挙げられる．反射や振動といった語は日常生活でも利用される語であり，また，この言葉を物理学の用語と捉えるのは不適切である．出現頻度以外の情報を用い，一般的な語の除去を行うなど，特徴語抽出プロセスの改善が必要であると考えられる.

「英語の男性名」という特徴カテゴリが取得されたツイートの例として「ニックの新曲『Just One Kiss ♪』」が挙げられる.「英語の男性名」と解釈された特徴語は「ニック」であり，特徴語として適切であると考えられるが，このツイートで言及している「ニック」は「ニック・ラシェル (歌手)」の事であるため，この「ニック」は「歌手」，または,「歌」といったカテゴリに選定されるべきであると考える．しかしながら,「ニック」は Wikipedia 上で「英語の男性名」という上位カテゴリしか持っておらず，またニックという省略した名前の情報から本名をたどることは不可能であると考える．提案手法の枠組みでは，Wikipedia 上の全てのカテゴリが，特徴カテゴリとして選定されうる．しかしながら,「英語の男性名」や「各年のコンピュータゲーム」といったユーザ分類において不適切なカテゴリについて，フィルタリングが必要であったと考えられる.

特徴語抽出がうまくいかないツイートについての考察を述べる．特徴語抽出がうまくいかないツイートとして図 4 に示す,「reply(あるツイートに対する返信), retweet(あるツイートの引用，もしくは引用返信).」，および，図 5 に示す,「リアルタイムなメディアに対しての実況 (テレビ番組，野球観戦等).」というものが挙げられる．reply や retweet はあるツイートに対する返答であることが多く，返答には特徴語が含まれない場合が多い．実況では感想や実況書き込みといったツイートが多く，特徴語が含まれない場合が多い．以上の理由により，適切な特徴語が取得出来なかったと考えられる.

・@**** ちょ！混ぜてｗｗｗ
・RT @****: 今日は『災厄』（講談社文庫）の発売日です。ｂｋ１への注文はこちら。 http://****
・あ、なるほど！それもやりたい！RT @****: @**** 出来れば、歴代のPDAのマガジンとかf^_^;)

図 4: reply, retweet の例

・ﾃﾝｼｮﾝｗ #agqr #syy
・なんかこの２人いい感じなような #FUJITV
・決戦じゃーーーーー！！気合を入れろーーーーー！！！

図 5: 実況の例

6 おわりに

本研究では，Wikipedia のカテゴリ階層を用いて，自動的に Twitter ユーザをカテゴライズする手法を提案した．評価結果は，特徴カテゴリとツイート関係の正解率 0.48，特徴語とツイート関係の正解率 0.78 となった．一般語の除去，特徴カテゴリとして適切であるカテゴリの選定を行うことで，正解率を向上させることができると考えられる．また，今回の手法ではユーザのツイートのみを利用するため，reply 先を取得しない，retweet の除去, ハッシュタグの除去を行った．しかしながら，reply 先や retweet の情報，ハッシュタグにも多くの情報があると考えられるため，今後はこれらを利用できるような手法を考案し，実装を行いたい.

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 リスト機能の応用”, 第 24 回人工知能学会全国大会, 2010.

[2] 田中 淳史, 田島 敬史, “twitter のツイートに関する分類手法の提案”,DEIM Forum 2010,2010.

[3] J. Weng, E. -P. Lim, J. Jiang, and Q. He, “ TwitterRank: Finding Topic-sensitive Influential Twitterers”. WSDM 2010, 2010.

[4] S. Banerjee, K. Ramanathan, and A. Gupta, “Clustering short texts using Wikipedia”, SIGIR 2007 Proceedings , 2007.