機械学習を用いたTweetの多カテゴリ分類

小坂 龍一† 青野 雅樹‡

†豊橋技術科学大学 工学部情報工学科

‡豊橋技術科学大学 大学院工学研究科 情報・知能工学系 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘1-1

E-mail: †kosaka@kde.cs.tut.ac.jp, ‡aono@tut.jp

あらまし Twitterでは，複数人をフォロー設定した場合，表示されるTweetが内容的に混在するという問題がある．そこで本研究ではTweetを収集し，機械学習によってTweetをカテゴリ分類するシステムを提案する．このシステムでは，カテゴリ分類済み文章群を用いて機械学習を訓練し，この機械学習を用いてTweetをカテゴリ分類する．機械学習にはSVMとRandom Forest法を用いた．また，文章のベクトルモデル化に用いる新たな素性を提案した．分類カテゴリには，“PC”，“政治”，“スポーツ”，“その他”の４カテゴリとし，多値分類の実験を行ったので，これを報告する．

キーワード テキストマイニング，Twitter，SVM，Random Forest

1. はじめに

インターネットが普及した近年，インターネット事業者が多様なWebサービスをユーザに対して提供している．最近ではWeb上においてユーザ同士の情報の双方向性を強化する動きが多い．その中でソーシャルネットワーキングサービス（Social Networking Service，以下SNSと呼ぶ）と呼ばれる社会的なネットワークをインターネット上で構築するサービスが提供されるようになった．SNSの形態の一つとして，マイクロブロギング・サービス（Microblogging Service，以下「マイクロブログと呼ぶ」）が挙げられる．マイクロブログとは，チャットとブログの中間的なサービスであり近年そのコンテンツは急増している．今回はその中でも代表的なマイクロブログである，Twitterを研究対象とする．

Twitter（ツイッター）とは2006年7月に開始されたマイクロブログサービスである．ユーザが投稿する1つの文章をTweet（ツイート）と呼びユーザ同士は互いにTweetを閲覧することができる．ユーザが他ユーザのTweetを購読することをフォローと呼び，自分のTweetを購読しているユーザのことをそのユーザのフォロワーと呼ぶ．Tweetの内容は140文字以内に限られユーザは文章を投稿する．

旧来のブログサービスでは，記事がカテゴリ別に分類されている場合が多い．しかし，Tweetはカテゴリ分類されていない．たくさんのユーザをフォローした場合，様々な話題のTweetが並ぶことになる．このようなときに，Tweetをカテゴリ分類できれば便利だと考えた．

今回の研究では，”PC”，”政治”，”スポーツ”，”その他”の４カテゴリを扱った．これらは，Tweetを収集した際に多くみられたものから選んだ．

2. 関連研究

豊富なテキスト情報とユーザ嗜好などを含んだTwitterに関する注目度が高くなったことで，Twitterに関する研究は増加している．しかし，Tweetをカテゴリ分類するという目的のものは少ない．

Sriramら[1]は，Tweetを目的別に分類するシステムを提案している．これは，Tweetをニュース（N），イベント（E），意見（O），取引（D），プライベートメッセージ（PM）の５種類に分類するものである．素性にはBag-Of-Wordsに加え，８個の新たな提案素性を使っている．分類器には，ナイーブベイズを使用している．評価実験の結果によれば，分類精度は，BOWのみの場合が70%程度で，提案素性を使用した場合が95%程度である．

本研究ではTweetの分類という点が共通しており，分類システムの骨格は本研究でも使える．しかし，分類カテゴリが異なるため，素性をそのまま使うことはできない．また，同様の理由で分類精度の比較もできない．

3. 提案システム

Tweetをカテゴリ分類するシステムを提案した．このシステムでは，カテゴリ分類済み文章群を用いて機械学習を訓練し，この機械学習を用いてTweetを分類する．

3.1. 分類までの流れ

Tweetを分類するまでの流れを述べる．

1. 分類器を訓練するための文章群を用意する．この文章群はカテゴリ分類済みである必要がある．
2. 訓練用文書群を形態素解析し，ベクトルモデルにする．ベクトルモデルにする際の素性には，BOW（Bag-of-words）と，文字種類の割合を使用した．各素性については，3.2節で詳しく述べる．
3. 訓練用文書群のベクトルモデルで，分類器を訓練する．
4. 分類対象Tweetも同様の手順でベクトルモデルにする．
5. 分類器によってTweetを分類する．

3.2. 各素性の説明

提案システムでは，文章をベクトルモデルにする際の素性に，BOW と，文字種類の割合を用いた．

- BOW

単語集合に存在する単語が，文章に出現するかどうかを素性とする．文章に出現する場合は１，出現しない場合は０とする．

単語集合は訓練用文書群を用いて，あらかじめ作成しておく．本研究では，各カテゴリに出現する名詞の，出現数上位 n 個を単語集合とした．

- 文字種類の割合

文章中の，ひらがな・カタカナ・漢字・数値のそれぞれの割合を素性とする．この素性の定義を式１に示す．なお，式１中の，char ratio は各文字種類の割合，char count は各文字の出現回数，length は文章の長さを示す．

$$\text{char ratio} = \left(\frac{char\ count}{length}\right)^2 \qquad (1)$$

この素性は，予備実験によって有効性が予想されたため，使用した．予備実験の結果を図１に示す．

予備実験では，訓練用文書群について，各カテゴリにおける文字種類の割合を算出した．この実験結果では，各カテゴリ間で文字種類の割合には差が現れており，この素性は Tweet の分類にも有効だと予想した．

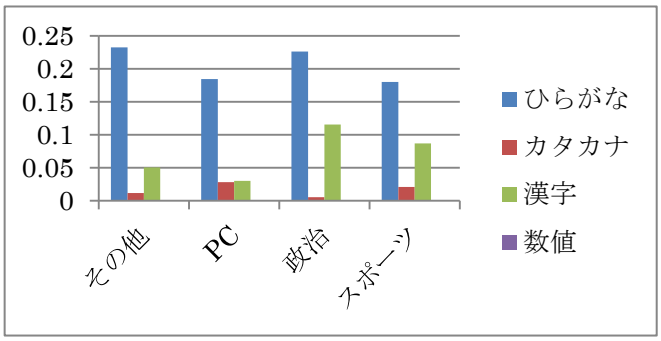

図 1　カテゴリ別の文字種類の割合

4. 評価実験

提案システムについて，評価実験を行った．評価実験では，訓練データと素性と分類器の違いによる，分類精度の比較を行った．

4.1. データセット

今回の評価実験では，Tweet と Yahoo!知恵袋の質問文を使用した．

4.1.1. Tweet

Tweet は分類対象データとして使っただけでなく，訓練データとしても使った．使用した Tweet の詳細を表１に示す．

Tweet の収集には Twitter Streaming API を使用した．収集された Tweet 全てを人手でラベル付けするのは，困難だと考えたので，いくつかの条件を設けて Tweet のフィルタリングを行った．そこから人手でラベル付けを行い，実験に使うデータを用意した．

Tweet のフィルタリングは次のようなものである。フィルタリングの過程を，図２に示す。

図 2　Tweet のフィルタリング

(ア) 日本語

今回は日本語の Tweet のみに限った．

(イ) 文字数の制限

今回は 100 文字以上の Tweet に限った．

(ウ) URL を含む Tweet

URL を含む Tweet は削除した．

(エ) RT やリプライを含む Tweet

他人の Tweet をコピーした Tweet（RT）や，誰かの Tweet への返事（リプライ）である Tweet は削除した．

以上のフィルタリングを行った後，人手によりラベル付けを行い，合計で 389 個の Tweet を準備した．

表 1 使用した Tweet の内訳

収集期間 ： 2011年10月27日 ～ 2011年10月29日

PC	政治	スポーツ	その他	合計
72	130	90	97	389

4.1.2. Yahoo!知恵袋の質問文

Yahoo!知恵袋の質問文は，訓練データとして使った．Yahoo!Japan から提供されているデータを使った．これは 2004 年 4 月から 2005 年 10 月のデータである．実験のために各カテゴリ 500 個，合計で 2000 個を準備した．質問文の内訳を次の表２に示す．

表 2　使用した Yahoo!知恵袋の内訳

PC	政治	スポーツ	その他	合計
500	500	500	500	2000

4.1.3. Yahoo!知恵袋のカテゴリ統合

Yahoo!知恵袋のカテゴリは，今回の比較実験で分類するカテゴリとは違う. そこで, Yahoo!知恵袋のカテゴリを統合し，実験に適した形に変更した．統合の一部を次の表３に示す.

表 3　カテゴリの統合

知恵袋カテゴリ	独自カテゴリ
パソコン、周辺機器	PC
ソフトウェア	
Macintosh	
インターネット	
政治、社会問題	政治
法律相談	
株と経済	
経済、経緯	
ボランティア、環境問題、国際協力	
国際情勢	
野球	スポーツ
サッカー	
スポーツ	
格闘技、武術	
ダイビング、サーフィン	
ゴルフ	
ヨット、ボート	
テニス、卓球	
ダンス、バレエ	
スキー、スノーボード	

4.2. 分類器

分類器には SVM と Random Forest を使用した。SVM についてはカーネルの選択を行った。

4.2.1. SVM カーネルの選択

SVM では，カーネルを選択できる．予備実験の結果が最も良かったカーネルを使用する事とした．次の表４に，予備実験の結果を示す．なお，訓練データは Yahoo!知恵袋，素性は BOW のみである.

表 4　予備実験の結果

カーネル	線形	RBF	シグモイド	ポリノミアル
分類精度	65.8%	66.6%	51.2%	50.9%

以上の結果から、カーネルには線形カーネルを使用した。

4.3. 実験結果

実験の結果を、表５に示す。なお、分類精度は式２により算出した。また、最も良かった結果について、分類の詳細な結果を表６に示す。

$$\text{分類精度} = \frac{\text{正解数}}{Tweet\text{総数}} \qquad (2)$$

表 5　実験結果

SVMによる分類

訓練データ:Yahoo!知恵袋の質問文

素性	BOW	文字種類の割合	BOW + 文字種類の割合
分類精度	66.58%	41.95%	68.89%

訓練データ:分類済みTweet

素性	BOW	文字種類の割合	BOW + 文字種類の割合
分類精度	62.80%	40.47%	64.70%

Random Forestによる分類

訓練データ:Yahoo!知恵袋の質問文

素性	BOW	文字種類の割合	BOW + 文字種類の割合
分類精度	63.32%	52.25%	52.85%

訓練データ:分類済みTweet

素性	BOW	文字種類の割合	BOW + 文字種類の割合
分類精度	62.54%	48.42%	51.10%

表 6　最も良かった結果

訓練データ　:Yahoo!知恵袋の質問文
素性　:BOW + 文字種類の割合
正解率　68.89% (268/389)

予想＼正解	PC	政治	スポーツ	その他
PC	**71**	0	34	39
政治	0	**127**	0	2
スポーツ	1	3	**56**	42
その他	0	0	0	**14**
正解率	98.61%	97.69%	62.22%	14.43%

5. まとめと今後の課題

5.1. まとめ

本研究では，提案システムにより 68%程度の精度で Tweet の分類を行う事ができた．また，提案素性の有効性も確認された．この節では，訓練データや素性による分類精度の違いなどについて考察を述べる.

5.1.1. 訓練データの比較

本研究では訓練データに Yahoo!知恵袋の質問文と，分類済みの Tweet を用いた．比較実験においては，Yahoo!知恵袋の質問文を訓練データに用いた方が，良い分類精度が出た.

これは，Yahoo!知恵袋の質問文と Tweet の文章の長さの違

いが原因だと考えられる. Yahoo!知恵袋の質問文は 1000 文字まで書けるが，Twitter では 140 文字までの制限がある．文章が長い方が，カテゴリの特徴が現れやすかったと考えた.

5.1.2. 素性の比較

本研究では２種類の素性を使った．BOW と文字種類の割合である．これらを比較した際，BOW のみの場合よりも，BOW に文字種類の割合を加えた場合の方が，良い分類精度が出た．この結果より，文字種類の割合は，分類に有効な素性だといえる.

5.1.3. 分類器の比較

分類器には SVM と Random Forest を使用した. 比較実験においては，SVM の方が良い結果が得られた.

5.1.4. カテゴリ別の正解率

表６の実験結果について，カテゴリ別の正解率は，『PC』と『政治』が 98%程度と高い正解率であった．しかし，『スポーツ』は 62%程度，『その他』は 14%と正解率が低い.

『スポーツ』の正解率が他のカテゴリに比べて低いのは，BOW や文字種類の割合が，スポーツを特徴付ける要素を表現することができなかったためだと考えた．正解率改善のためには，新しい素性を作る必要がある.

『その他』は，他のカテゴリに選ばれなかった Tweet であるため，他の３カテゴリの正解率を向上させることで，このカテゴリの正解率も向上させることができると考えた.

5.2. 今後の課題

今後の課題としては，提案システムの分類精度向上があげられる．この課題のために，今回使用しなかった分類器を使用した実験，Yahoo!知恵袋以外の文章群を訓練データに用いた実験，新しい素性の考案などをする必要がある.

参　考　文　献

[1] Twitter data API. http://apiwiki.twitter.com/ Twitter-API-Documentation.

[2] Twitter help keeping search relevant. http://help.twitter.com/forums/10713/entries/42646.

[3] Twitter help Twitter rules. http://help.twitter.com/forums/26257/entries/18311.

[4] Bharath Sriram, David Fuhry,Engin Demir, Hakan Ferhatosmanoglu,Murat Demirbas. Short Text Classification in Twitter to Improve InformationFiltering. SIGIR’10, July 19?23, 2010, Geneva,Switzerland.ACM 978-1-60558-896-4/10/07.

[5] Danesh Irani, Steve Webb, Calton Pu, Kang Li. Study of Trend-Stuffing on Twitterthrough Text Classification

[6] 辻村 浩 『Twitter API プログラミング』 初版 ワークスコーポレーション