3N-8

Retweetに着目した広がりやすいTweetの特徴分析

荒川　唯† 亀田　尭宙†† 相澤　彰子‡ 鈴木　崇史‡‡

東洋大学社会学部† 東京大学大学院情報理工学系研究科†† 国立情報学研究所‡ 東洋大学社会学部‡‡

1. はじめに

Twitterは2011年3月に発生した東日本大震災でソーシャルメディアの1つとして注目された．Twitterは140字以内で記事 (Tweet) を投稿するシンプルな構造であり，リアルタイム性と情報の強い伝播力をもつ．

中でも，Retweet (RT) は他のユーザのTweetを引用し，情報を伝播させるものであり，Twitter特有の重要な機能である．上記の震災時にも，災害情報やデマなど様々な情報が伝播されており，どのようなTweetがRTされやすいのかを検討することは，Twitter上での情報伝播のあり方を検討する上で重要な価値をもつ．

このような背景のもと，本研究では，4種類の特徴量と3つの分類実験を行うことで，RTされやすいTweetの特徴を分析する．著者推定等で有力とされている機能的特徴量 （cf., Stamatatos, 2007）と，Twitterに特有な特徴量を併用し，RTの大小，アカウントカテゴリー，両者をクラスとした分類実験を行うことで，RTされやすいTweetの特徴を分析する．これによって，Twitter上の情報伝播を理解する上で有用な知見を得る．

2. データ・分析手法

2.1. データ

4つのカテゴリー（有名人，芸能人，団体，キャラクター）を設定し，ツイナビ[1]を元に，各カテゴリーについてフォロワー数上位10アカウントを選択．TwitterAPIを用い，2011年9月7日，入手可能な全28,756Tweetを取得した．各Tweetは，（公式）RT数上位と下位，2グループに分類した．Tweet本文は，MeCab[2]用いて，形態素解析を適用した．

2.2. 特徴量

特徴量として，以下の三種類を利用する．

(1) 機能語（名詞-非自立，名詞-代名詞，接続詞，連体詞，助詞，助動詞）の相対頻度

(2) 品詞（名詞，動詞，副詞，形容詞，接頭詞，感動詞，接続詞，連体詞）記号，その他の割合

(3) Tweet特有の記法（@，RT (QT)，#），URLの の割合

(4) Tweetの情報提供上の役割（ひとりごと，宣伝，情報提供，会話）の割合

このうち，1については，MeCabの品詞タグを利用し，2については，マッチングで抽出する．3については，情報の役割のタグづけを各アカウントにつき半数をサンプリングし，[3] 第一著者が手作業で分類した．

2.3. 分類実験

上述の特徴量(1-4)全てを用いて，以下の3種類の分類実験を行う．

実験1. RTの大小2クラス分類，
実験2. アカウントのカテゴリー4クラス分類，
実験3. RT大小×アカウントカテゴリー8クラス分類．

分類実験には，ランダムフォレスト機械学習法（Breiman, 2001）を用い，精度，再現率，*F1*値を用いて評価する． ランダムフォレストの重要度計算（Breiman, 2001）によって，各分類に寄与の大きい特徴量を抽出する．

3. 結果と考察

3.1. 分類実験の結果

表1はランダムフォレストによる分類実験の結果である．結果，アカウントグループによる分類を行った実験2が全ての評価指標でもっとも高い値を示した．アカウントグループ，RTを合わせた実験3は分類性能が低く，アカウントの属性によらず，RTの大小2クラスで分類した実験1もそれほど高い値を示さなかった．

Analyzing the characteristics of tweets likely to be retweeted

† Yui Arakawa, Faculty of Sociology, Toyo University
†† Akihiro Kameda, Graduate School of Information Science and Technology, University of Tokyo
‡ Akiko Aizawa National Institute of Informatics
‡‡ Takafumi Suzuki, Faculty of Sociology, Toyo University

[1] twinavi.jp
[2] mecab.sourceforge.net

[3] 半数をサンプリングしたのは，手作業による分類を簡便化するためである．

表 1 分類実験の結果

	精度	再現率	*F1* 値
実験 1	39.97	40.00	39.96
実験 2	78.84	80.00	79.90
実験 3	13.94	16.25	—

3.2. 分類に寄与の大きい特徴量の分析

次に，分類に寄与の大きい特徴量について検討する．表 2 は，実験 1，2，3 それぞれの分類に帰与の大きい特徴量上位 30 に，特徴量(1-4)がいくつ含まれるか示したものである．実験 1 は，ほとんど機能語で占められている一方，実験 2 や実験 3 では，(2)品詞の割合，(3)@の有無，URL の有無や，(4)内容分析カテゴリーも含まれている．この結果から，アカウントグループの分類には，(2)Twitter 特有の記法，URL の有無や(3)Tweet の機能なども重要であると結論づけられる．一方，RT 数単独では，機能語が重要な要因であると解釈できる．

表 3 は，重要特徴量のうち(1)について，さらに，品詞ごとに分類したものである．表 3 から，すべての実験において助詞，助動詞が重要特徴量に含まれていることがわかる．また，接続詞は品詞の比率としては少ないものの重要特徴量には比較的多く含まれている．これら接続詞も分類に影響を与えていることが示唆される．

4. おわりに

本研究では，4 種類の特徴量と 3 つの分類実験を行うことで，RT されやすい Tweet の特徴を検討した．分類性能自体は，アカウントのカテゴリーによって分類した場合の方が圧倒的に高いものの，RT の大小によっても一定程度分類が可能であり，また，前者の特徴量には Twitter 独自の記法や URL，情報提供の役割など，独自の特徴量を多く含んでいるのに対して，後者の分類に有効な特徴量の多くは，機能語で占めれていることが示された．これらの特徴量は，RT のされやすさに関連すると示唆される．

今後は，RT する側の特徴を含め，さらにいくつかの特徴量を加えるとともに，他の分類手法を適用し，今回の研究結果の一般性を確認していきたい．

謝辞

本研究は，科学研究補助金若手研究(B)「計算文体論による多種メディアテキスト分析（研究代表者：鈴木崇史，研究課題番号：23700288）」および，国立情報学研究所公募型共同研究「多種テキストからのコミュニケーション・スタイルの抽出ならびにその分析と応用（研究代表者：鈴木崇史）」より，一部支援を受けています．ここに記して謝意を表します。

文献

Breiman, L. (2001) Random forests, *Machine Learning*, 45(1), 5-23.

Stamatatos, E. (2007) A survey of modern authorship attribution methods, *Journal of the American Society of Information Science and Technology*, 60(3), 538-556.

表 2 分類に寄与の大きい特徴量上位 30 の内訳

	(1)機能語	(2)品詞	(3)記法・URL	(4)役割
実験 1	29	0	0	1
実験 2	20	4	3	3
実験 3	19	5	2	4

表 3 上位特徴量における機能語内訳

	助詞	助動詞	接続詞	名詞-代名詞	名詞-非自立	連体詞
実験1	10	7	5	2	4	1
実験 2	5	8	3	0	2	1
実験 3	6	8	2	0	6	1