D-016

Twitter におけるスパムユーザの分別
Filtering Spammers on Twitter

中村 悠一† 山田 剛一† 絹川 博之†
Yuichi Nakamura Koichi Yamada Hiroshi Kinukawa

1. はじめに

ツイートと呼ばれるテキストメッセージを複数のユーザ間で共有しコミュニケーションを行う Twitter[1]と呼ばれるサービスが人気を集めている. Twitter は多くの利用者を集めているため, スパムの標的ともなっている. Twitter のスパムユーザには, 悪質なサイトへのリンクを含んだツイートを定期的に投稿するもの, 不特定多数のユーザにスパムをリプライするものなどがあり, これらの行動は一般ユーザと異なっていることが多い. 本研究では, スパムユーザと非スパムユーザのそれぞれの行動特徴などを機械学習することによりフィルタを生成してスパムユーザを分別する.

2. Twitter におけるスパムユーザ

Twitter 上のスパムユーザは, 自らの利益を目的としたサイトに一般ユーザを誘導するためリンクを含んだツイートを大量に投稿するなど, 一般ユーザと異なる行動を取ることがある. 機械学習による自動分別に利用するために, これらのスパムユーザの特徴を分析した. その結果特徴は大きく次の 2 種類に分けられ, 計 15 種を得た.

(1) フォロー関係における特徴・・・ 3 種
フレンド数, 一日のフォロー数など
(2) ツイートの投稿における特徴・・12 種
リンクの割合, ハッシュタグの割合など

3. スパムユーザの分別 実験

3.1 Twitter 上からのデータ収集

スパムユーザの分別を行う為に Twitter 上からユーザ情報とそのユーザの最新ツイートを上限 200 件ずつ収集した. 最終的に約 3.3 万ユーザ, 約 630 万ツイートを収集した. また, 収集したデータからランダムにユーザを抽出し人手でスパムユーザの分別を行うことで, 実験に用いる学習用データを構築した. 構築したデータは表 1 の通りである. 本研究では日本語を使用するユーザのみを対象としユーザを収集した.

3.2 学習アルゴリズムと評価指標

本研究では, スパムユーザを自動分別する為に機械学習の手法を用いた. データマイニングツールの Weka[2]とそれに内蔵されている決定木の学習アルゴリズムである J48[3]を用いて分類器を生成する. また, 10 分割交差検定法により精度と再現率, その調和平均である F 値を求める.

表 1. 学習用データ

非スパムユーザ	スパムユーザ	合計
1423	158	1581

それぞれの値の定義を以下に記す.

$$\text{精度} = \frac{N_{tp}}{N_{tp} + N_{fn}} \quad \text{再現率} = \frac{N_{tp}}{N_{tp} + N_{fp}} \quad \text{F 値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}}$$

N_{tp}: Spam/Normal ユーザを Spam/Normal ユーザとして分別した数
N_{fp}: Normal/Spam ユーザを Spam/Normal ユーザとして分別した数
N_{fn}: Spam/Normal ユーザを Normal/Spam ユーザとして分別した数

3.3 特徴の選別と分別結果

本研究では, 最も性能のよい決定木を求めるために次の方法で特徴の選別を行った. まず, 15 種全ての特徴の組み合わせ(2^{15}-1 = 32,767 通り)を用いて決定木の学習をそれぞれ 10 分割交差検定法で行う. この手順を, 学習データの並び順の影響を排除するためにランダムに並び替えた 10 通りの入力データで行い, それぞれの性能を求めた. そして, この 10 通りの F 値の平均が最も高くなる特徴の組み合わせを求めた. スパムユーザの分別に有効な特徴として選別されたものは次の 5 つであることがわかった.

特徴値 1 : リンク付きツイートの割合
スパムユーザは自らの利益を目的としたサイトに一般ユーザを誘導するためリンクを含んだツイートを大量に投稿する事がある. 以下の式の値が大きいほどその特徴が強く, スパムユーザである可能性が高いと考えられる.

$$LinkRatio = \frac{\text{リンク付きツイート数}}{\text{収集したツイート数}}$$

特徴値 2 :リンク付きの重複するツイートの割合
スパムユーザはツイートの投稿を自動化していることがあり, ツイートを重複して投稿することが多い. そしてそれらのツイートはリンクを含んでいることが多い. 以下の式の値が大きいほどその特徴が強く, スパムユーザである可能性が高いと考えられる. 重複するツイートがない場合は, −1 をとる.

$$DuplicateTweetsLinkRatio = \frac{\text{リンク付きの重複するツイート数}}{\text{重複するツイート数}}$$

特徴値 3 : 重複するリプライの割合
スパムユーザは不特定多数のユーザにリプライを送ることがある. そして, リプライの投稿は自動化されているため, 内容が重複することが多い. 以下の式の値が大きいほどその特徴が強く, スパムユーザである可能性が高いと考えられる. リプライがない場合は, −1 をとる.

$$DuplicateReplyRatio = \frac{\text{重複するリプライ数}}{\text{リプライ数}}$$

†東京電機大学大学院 未来科学研究科
Graduate School of Science and Technology for Future Life, Tokyo Denki University

特徴値 4 : 安全ドメインの割合
Twitter には, 特定の話題に関するニュースを定期的につぶやくボットなどが存在する. スパムユーザとこれらのボットを区別するため, スパム目的には利用できないサイトのドメインを安全ドメインとし, 収集したツイートに含まれる総リンク数における割合を調べる. 以下の式の値が大きいほど, 無害なボットである可能性が高いと考えられる. リンクがない場合は, −1 をとる.

$$SafeDomainRatio = \frac{\text{安全ドメインのリンク数}}{\text{総リンク数}}$$

特徴値 5 : 最大投稿間隔
スパムユーザはツイートの投稿を自動化していることが多く, ツイートの投稿間隔が短くなる傾向がある. また, 昼夜を問わずにツイートを行うスパムユーザも存在する. よって, ツイートの投稿間隔の最大値が短いほどスパムユーザである可能性が高いと考えられる.

$$MaxInterval = \max(\text{ツイートの投稿間隔})$$

上記 5 種の特徴値を利用し, 10 通りのデータで決定木を学習した際の精度, 再現率, F 値の平均を表 2 に, 生成された決定木を図 1 に示す.

表 2. 性能の平均

	精度	再現率	F 値
スパムユーザ	0. 834	0. 929	0. 879
非スパムユーザ	0. 992	0. 979	0. 985

4. Twitter 上のスパムユーザ分別システム

本研究では, Twitter のスパムユーザを分別するシステムを作成した. システムのインタフェースを図 2 に示す. 本システムは Web アプリケーションとして動作し, 以下のどちらかを指定しスパムユーザの分別を行う.
・指定した任意のユーザ
・ユーザが指定したキーワードでツイートを検索したときの, 検索結果の各ツイートを投稿したユーザ

また, 本システムは, 3.3 で生成した決定木を用いてスパムユーザの分別を行う.

5. 考察

本研究では, 3.3 で選別した特徴の組み合わせで決定木を学習しスパムユーザの分別を行った. その結果, スパムユーザの分別における精度が 0.834 , 再現率が 0.929 で, 精度がやや低い値となった. これは, リンクを含んだツイートを定期的に投稿するボットやツイートの投稿数の少ないユーザを誤検出した事が原因であると考えられる. ツイート本文の内容を考慮しなかったために, リンクを含むツイートの割合などの表層的な特徴の類似する非スパムユーザを誤検出してしまったためであると思われる. 今後の課題は, ツイート本文の特徴を導入し, 今回求めた特徴と組み合わせることでスパムユーザの分別精度を向上させることである.

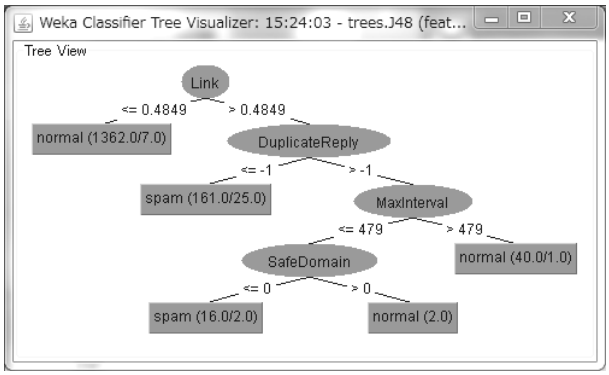

図 1. 生成された決定木

図 2. システムのインタフェース

6. おわりに

本研究では, Twitter 上から約 3.3 万ユーザ, 約 630 万ツイートを収集した. そして, 収集したデータからランダムにユーザを抽出し人手でスパムユーザ判定を行うことで実験に用いる学習用データを構築した. また, スパムユーザのフォロー関係の特徴とツイートの投稿における特徴を分析し, それらの中から決定木の学習における F 値の平均が最も高くなる特徴の組み合わせを選別した. 学習用データを用いて選別した特徴についての決定木を学習した結果, 精度が 0.834 , 再現率が 0.929 , F 値が 0.879 となった.

今後はさらなる性能の向上のためにツイート本文の内容を用いた特徴を導入する必要があると考えられる.

参考文献

[1] Twitter：http://twitter.com/
[2] Weka：http://www.cs.waikato.ac.nz/ml/weka/
[3] J48
http://weka.sourceforge.net/doc/weka/classifiers/trees/J48.html