

Twitter におけるスパムユーザフィルタの開発とその評価 A Spam Filter for Twitter and its Evaluation

中村 悠一†
Yuichi Nakamura

山田 剛一†
Koichi Yamada

絹川 博之†
Hiroshi Kinukawa

1. はじめに

Twitter [1] は多くの利用者を集めているため、スパムユーザ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、これらは一般ユーザが Twitter を利用する上で妨げとなる。スパムユーザの中には Twitter 社によってアカウントを停止 (サスペンド) されるものもあるが、そのカバー率は 36 % 程度とあまり高くない。これは、一般ユーザの誤検出を防ぐためであると考えられる。本研究では、Twitter 利用時にユーザが用いることを前提としたリアルタイムのスパムユーザフィルタの作成を目指している。既報告 [2] では、人手によって分別した学習データをもとに機械学習によってフィルタを作成した。しかし、この手法ではデータセットの作成に人手による確認を行う必要があった。

本論文では、データセットの作成を効率化するために、Twitter 社によってサスペンドされたユーザをスパムユーザとして学習データを作成した。このデータで学習した分類器を、人手で分別したスパムユーザを用いて作成したテストデータで評価し、Twitter 社にサスペンドされていないスパムユーザの検出においても本手法で作成した学習データが有効であるかを検証した。また、更なる精度向上のため特徴の追加と決定木、Naïve Bayes, サポートベクタマシン (SVM) について分別性能を実験評価した。

2. Twitter におけるスパムユーザ

2.1 スпамユーザの定義

Twitter 社は、スパムの定義として『有害なリンクを投稿 (マルウェアやフィッシングサイトへ飛ぶリンク)』や『重複ツイートを繰り返し投稿』など 7 個の基準を設け [3], これに違反するユーザのアカウントをサスペンドしている。本論文では、これらの基準を踏まえ『自らの利益を目的としたサイトに一般ユーザを誘導しようとするユーザ』という定義をもとにスパムユーザの分別を行った。

2.2 特徴の抽出

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に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、これらの行動から特徴を抽出しユーザの特徴を前回報告時の 15 種に加え、14 種導入した。その結果特徴は大きく次の 7 種類に分けられ、計 29 種を得た。以下、括弧内は新しく追加した特徴数を表す。

(1) フォロー関係における特徴・・・4 種 (2 種)

スパムユーザには、一方的に多数のユーザをフォローすることによりフォロー返しを狙うものがある。このため、フレンド数がフォロー数に比べて大きくなるなどの性質がある。これらの行動を検出するために、フレンド数、フォロー数、フレンド数とフォロー数の比率、一日あたりのフォロー数の特徴として導入した。

(2) URL における特徴・・・5 種 (3 種)

スパムユーザは自らの利益を目的としたサイトの URL を含んだツイートを大量に投稿する事がある。この行動を検出するため、URL 付きツイート率、重複ドメイン付ツイート率 (2 回以上出現するドメインの URL を含んだツイートの割合)、ユニークドメイン付ツイート率 (1 回しか出現しないドメインの URL を含んだツイートの割合)、URL 当たりの平均文字数を導入した。また、一般のボットの誤検出を防ぐためにニュースサイトなどスパム目的に利用できないサイトの URL を安全ドメインとしてその割合を求める安全ドメイン率を導入した。

(3) リプライにおける特徴・・・2 種 (0 種)

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、それらのリプライは URL などを含むことが多い。これらの行動を検出するためにリプライ率、URL 付リプライ率を導入した。

(4) ハッシュタグにおける特徴・・・2 種 (0 種)

スパムユーザの中にはスパムツイートが一般ユーザの目に触れやすいようにハッシュタグを付けるものがある。また、フォローを増やすことを目的として、#sogofollow など相互フォローを呼びかけるハッシュタグ (相互フォロータグ) を付けることがある。これらの振る舞いを検出するために、ハッシュタグ付ツイート率、相互フォロータグ付ツイート率を導入した。

(5) ツイートの投稿における特徴・・・5 種 (2 種)

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、投稿間隔が一定になるなど一般ユーザと異なる行動が見られる。これらの行動を検出するために、投稿間隔の分散、投稿間隔のバリエーション、最大投稿間隔、一日あたりの投稿数、一日あたりの投稿数の分散を導入した。

(6) ツイートの重複における特徴・・・3 種 (0 種)

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、同一のツイートを複数投稿することがある。この振る舞いを検出するため、重複ツイート率 (2 回以上投稿された同一内容のツイートの割合)、リンク付重複ツイート率、重複リプライ率を導入した。

†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

(7) その他の特徴・・・8 種 (7 種)

アカウントの生存期間, 最大ツイート長, 最小ツイート長, 平均ツイート長, ツイート長の分散, 語のバリエーション, 数字を含むツイート率を導入した. また, ベイジアンフィルタによって学習データからスパムワードを抽出し, それらのワードを含むツイートの割合 (スパムワード含有率) を特徴として導入した.

3. スпамユーザの分別実験

3.1 Twitter 上からのデータ収集

Twitter 上からユーザ情報とそのユーザの最新ツイートそれぞれ上限 200 件ずつ収集し, 約 3.3 万ユーザ, 約 630 万ツイートを収集した. また, 『自らの利益を目的としたサイトに一般ユーザを誘導しようとするユーザ』をスパムユーザの基準として人手によって分別しテストデータを構築した. さらに, 収集日から 1 年後に API を用いてデータセット内のすべてのユーザのアカウントを確認し, Twitter 社によってサスペンドされた 277 のユーザを同定した. このサスペンドされたものをスパムユーザとして, 学習データを構築した. 構築した学習・テストデータは表 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.

表 1 構築した各データセット			
	一般ユーザ	スパムユーザ	合計
学習	800	200 (200)	1000
テスト	600	150 (55)	750

※括弧内はサスペンドされたユーザ

3.2 学習アルゴリズムと評価指標

本研究では, スпам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 [4] とそれに内蔵されている SMO (SVM), Naïve Bayes, J48 (C4.5) の学習アルゴリズムを用いて学習データによって学習した分類器をテストデータで評価した. 評価指標には, 精度と再現率, その調和平均である F 値を用いる.

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

N_{tp} : Spam/Normal ユーザを Spam/Normal ユーザと分別した数

N_{fp} : Normal/Spam ユーザを Spam/Normal ユーザと分別した数

N_{fn} : Spam/Normal ユーザを Normal/Spam ユーザと分別した数

3.3 分別実験

SVM, Naïve Bayes, C4.5 を用いて, 学習データで生成した分類器をテストデータで評価した結果と 10 分割交差検定で評価した既報告の結果を表 2 に示す. また, SVM におけるサスペンドされたユーザとそうでないユーザの各々の再現率を表 3 に示す.

表 2 各分類器の評価結果

アルゴリズム	精度	再現率	F 値
SVM	0.929	0.960	0.944
Naïve Bayes	0.722	0.987	0.834
C4.5	0.934	0.753	0.834
既報告 (C4.5)	0.834	0.929	0.879

表 3 テストデータにおける各ユーザの再現率

テストデータ	サスペンド	非サスペンド
再現率	0.981 (54/55)	0.947 (90/95)

4. 考察

(1) SVM, Naïve Bayes, C4.5 の分別性能を比較した結果, SVM が F 値において最高となり高い分別性能を持つ分類器を作成できることが分かった.

(2) SVM 分類器においてスパムユーザであると誤判定された一般ユーザ (12 ユーザ) を確認したところ, すべてのユーザが自動的にツイートを投稿するボットであることが分かった. これらのユーザは URL 付ツイートや重複する内容のツイートを多数投稿する, 一日あたりのツイート数が多いなど, 特徴がスパムユーザと一致するために誤検出してしまったと考えられる. 今後一般のボットを分別する特徴の導入や安全ドメインの増加を行い誤判定を防ぐ必要がある.

(3) C4.5 分類器の分別性能が既報告時に比べて低下した. これは, 互いに独立でない特徴があったためであると考えられる. このため, 今後独立性を考慮した特徴の選択を行い C4.5 の性能を再度評価したいと考えている.

5. おわりに

本研究では, Twitter 社によってサスペンドされたユーザをスパムユーザとし構築した学習データによって SVM, Naïve Bayes, C4.5 などの分類器を学習した. それらの分類器を, 人手によって分別したスパムユーザからなるテストデータで評価した結果, SVM 分類器のスパムユーザの分別精度が 0.929, 再現率が 0.960, F 値が 0.944 となり最高の分別性能を示した. また, Twitter 社によってサスペンドされたユーザを 0.981 の割合で, サスペンドされていないスパムユーザに関しても 0.947 の割合で検出できることが分かった.

このことから, Twitter 社によってサスペンドされたユーザをスパムユーザとして構築した学習データを用いて機械学習した分類器で, Twitter 社にサスペンドされていないスパムユーザに対しても高い再現率で分別できることが分かった. 今後は, サスペンドされたユーザを自動収集する効率的な手法を考案し学習データの量を増やし, また一般ユーザに関しても同様に自動収集手法を考案することを目指す.

参考文献

- [1] Twitter : <http://twitter.com/>
- [2] 中村 悠一, 山田 剛一, 絹川 博之, “Twitter におけるスパムユーザの分別”, 情報科学技術フォーラム (2011)
- [3] Twitter 社のスパムの基準 : <https://support.twitter.com/articles/234686->
- [4] Weka : <http://www.cs.waikato.ac.nz/ml/weka/>