講演番号：137

LDAを用いたTwitterにおける スパムに乗っ取られたアカウントの検出

D-19

Detection of Taken Over Accounts in Twitter Using LDA

根本 凌介　　伊與田 光宏
Ryusuke NEMOTO　　Mitsuhiro IYODA
千葉工業大学　情報工学科
Department of Computer Science, Chiba Institute of Technology

1. はじめに

近年スマートフォンとインターネットの普及に伴い、ソーシャルメディアは急激な成長を遂げ、利用率は増加傾向にある。

ソーシャルメディアが発達していく中、Twitterは国内月間アクティブユーザ数が4,500万人を突破した。Twitter内でも、ユーザが悪質と感じるスパム行為が流行している。その際、乗っ取られから、ユーザ自身がスパムに加担するケースがあげられる。

2. 目的

本研究ではTwitterにおける乗っ取られたアカウントのツイートの本文に着目し、乗っ取られたアカウントの検出を行う。

3. 乗っ取られたアカウントの定義

Twitter[1]はスパム行為について、『ウェブサイトへの訪問数を増やしたり、無関係なアカウント、商品、サービス、イニシアティブなどに注目を集めたりするために、TwitterやTwitter利用者の快適性を操作したり損なったりすることを目的とした大量または過剰な行動を示します。』 と定義している。そのため本研究では、乗っ取られたアカウントの定義については、『スパムツイートを意図せず行ってしまう一般ユーザアカウント』と定義する。

4. 検出方法について

本研究では次の手順で検出を行う。

1. Twitterからツイートを取得
2. ツイートの本文を形態素解析
3. 特徴語を作成
4. トピックから分類器を作成
5. 分類器を用いてツイートの判別
6. 乗っ取られたアカウントの検出

特徴語の抽出についてはLatent Dirichlet Allocation(LDA)を用い、ツイートの判別にはサポートベクターマシン(SVM)を用いる。
コーパス（全対象文書）の生成確率を次に示す。

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \quad (1)$$

式中のw_nは単語の個数、z_nは単語が持っている話題、θ は文書が持っている話題分布、α はトピックの生起パラメータ、β はあるトピックkにおける単語の生起パラメータを示す。検出の流れを図1に示す。

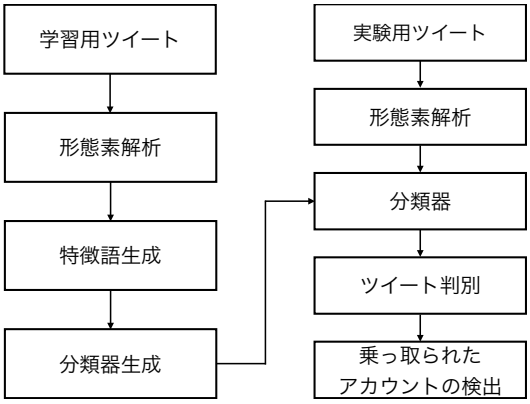

図 1 . 検出の流れ

5. 分類器生成

学習用ツイートから、特徴語を生成し、分類器生成する。学習用ツイートとは別に、スパム、非スパムを含む実験用ツイートを用意し、乗っ取られたアカウントの検出を行う。

6. 評価方法

評価方法として、分類器の評価では、ツイートの判別の精度、再現率、F値を算出する。乗っ取られたアカウントの検出では、ユーザのスパムツイートの検出回数にしきい値を決め、判別を行う。乗っ取られたアカウントの検出回数から評価を行う。

7. おわりに

本研究ではスパムツイートに用いられる特徴語を抽出を行い、スパムの判定を行い、乗っ取られアカウントの判別を行った。

参考文献

[1]Twitterルール(https://help.twitter.com/ja/rules-and-policies/twitter-rules)
[2]若井 一樹,佐々木 良一, "Twitterのスパム検知機能となりすまし検知機能の開発と評価", 情報処理学会論文誌, 2015