マイクロブログにおけるインタラクション構造及び変化に基づくリンク誘導型スパムユーザ検出

清水　翔太†　　豊田　正史††

† 東京大学 大学院情報理工学系研究科　〒 113–8656 東京都文京区本郷 7–3–1
†† 東京大学 生産技術研究所　〒 153–8505 東京都目黒区駒場 4–6–1
E-mail: †shimisho@tkl.iis.u-tokyo.ac.jp, ††toyoda@tkl.iis.u-tokyo.ac.jp

あらまし　マイクロブログが普及するにつれて、マイクロブログを媒体とするスパムも増加している。本研究では代表的なマイクロブログのひとつである Twitter を対象としたスパム検出に取り組んだ。スパムユーザは自身の Web ページへの誘導リンクを多くの一般ユーザの目に触れさせる工夫として複数のアカウント同士で相互にインタラクションを行うという傾向があり、そのようなアカウント周辺のインタラクションの構造に着目すると特異な構造が観測できる。加えて、スパムアカウントの特徴量の時間非依存性からインタラクション構造とその変化に関する特徴量を用いてスパムを判定する手法を提案する。2014 年 6 月から 7 月までの期間の 100 万人規模の実際の投稿からリツイートの多いユーザを候補ユーザとし、アカウントの投稿および挙動を実際に確認して正解データを人手で作成し、このデータセットを用いて実験を行い、手法の有効性を示した。

キーワード　マイクロブログ、ソーシャルグラフ、リンク誘導型スパム

1 はじめに

1.1 情報発信の敷居の低下とスパムの拠点の遷移

2000 年代後半頃、ブログサービスが人々に対してオンライン上への情報発信に対する敷居を低くする役割を果たした。かつては、情報の発信者と受信者は固定されていたがおよそ 10 年前、それらの区別が無くなり、誰もが気軽に情報を発信できる時代が到来した [4]。また、同じ頃に登場した mixi(注1) は人々のオンライン上での交流を促進するという役割を果たした。

近年では、スマートフォンなどの携帯端末機器の普及により、実世界での出来事や、ある瞬間に感じたことを気軽に投稿できるようになり、Twitter(注2) や Facebook(注3) といったソーシャルネットワーキングサービスでの短文記事や画像の投稿や、ユーザ間の交流が盛んに行われている。具体的にソーシャルネットワーキングサービスのユーザ数を示すと、Twitter では 2014 年 7 月から 9 月までの平均月間アクティブユーザ数は 2 億 8400 万であり、そのうちの約 80%がモバイルアクティブユーザであると報告されている [2]。Facebook では、2014 年 12 月の月間アクティブユーザ数 (MAUs) は 13 億 9000 万、そのうちモバイルアクティブユーザは 11 億 9000 万人であると報告されている [1]。

マイクロブログは企業や団体および個人が宣伝活動を行うことにも用いられている。一方で、Web の変遷に伴い悪意のあるユーザたちも活動の拠点を移してきた [5]。具体的には、複数アカウントを用いた不自然な宣伝活動や、権限を不正に取得した他ユーザーの投稿の操作が問題視されている。

スパムアカウントおよびその周辺アカウントのインタラクション構造は特異な構造を持つことが多い。スパムは多くのユーザに対して目的の URL を読ませて、リンク先に誘導したいという意図を持っているためである。また、時間の経過に伴い挙動が大きく変化するスパムも存在している。そこで本研究では、アカウント間のインタラクションの構造と時間変化に着目してスパムを検出することを試みた。

1.2 以降の本稿の構成

本稿の以降の流れは次の通りである。まず第 2 章でマイクロブログにおける、スパム検出や分析に取り組んだ過去の研究について紹介する。続く第 3 章で本研究で対象とした、代表的なマイクロブログである Twitter におけるスパムユーザの特徴について、ソーシャルメディアとしての Twitter の性質を踏まえつつ述べたあと、第 4 章で本研究でのスパムユーザ検出のために用いる手法について述べる。そして第 5 章で実験設定及びその結果について述べ、考察を行う。最後に第 6 章で本研究についてまとめ、今後の予定を述べる。

2 関連研究

SNS の普及に伴い、SNS に活動拠点を移す攻撃者も多く存在する。そのため、マイクロブログを用いたスパムに関する研究も多数行われている。本章ではスパムの検出に関連する研究を以下の観点から紹介する。

- 検出の対象（スパムユーザか、それともスパム投稿か）
- 用いた特徴量
- その特徴量の意図（どのようなスパムを検出したいか）

(注1)：https://mixi.jp
(注2)：https://twitter.com
(注3)：https://www.facebook.com

Stringhini らは Twitter のほか、Facebook や Myspace(注4) におけるスパムの挙動を分析する目的でのアカウントを複数個作成し、怪しいアカウントからの通知などを分析することを行った [11]。その分析をもとにして、投稿数や URL 数やフォロワー数などの基本的な特徴量を組み合わせて検出を行った。Benevenuto らは、基本的な特徴量とそれらの最大値や最小値や平均や中央値といった統計量のほか、スパムに見られる単語や投稿の傾向を特徴量とし、人手によるラベリングで正解データを作成し、SVM での分類問題に取り組んだ [6]。

McCord らの研究 [10] では、手動によるスパム判定方法は経験に依存するためスパムでないユーザがスパムであると判定されることを問題視し、自動でスパム判定を行う方法の必要性を指摘している。彼らは、ユーザベースの特徴量と内容ベースの特徴量を両方用いている。ユーザベースの特徴量は、Wang の研究 [12] で用いられている発話内容やアカウントの特徴量に加え、24 時間を 3 時間ごとに分割した 8 スロットの投稿回数の分布を特徴量に加えている。一般ユーザが就寝している時間帯にもスパムアカウントが投稿を行っているという傾向からこの特徴量を加えている。内容ベースの特徴量としては URL 数、メンション数、含まれる単語の傾向、wordweight、リツイート数、ハッシュタグ数を用いている。手法としては SVM、ナイーブベイズ、k 近傍法、ランダムフォレストを用いており、これらを比較してランダムフォレストが最も高い precision と F 値を出したと述べている。

インタラクションの構造に着目した特徴量を用いた研究のひとつとして Chen らによる研究 [8] がある。投稿間の類似度やフォロー/フォロワー比などの特徴量はスパムアカウントに容易にかいくぐられてしまうことを指摘し、スパム投稿のリツイートによる拡散構造に着目し、message-passing graph という枠組みを提案している。グラフ構造に対し、クラスタリング係数や推移性等の指標を特徴量として導入している。2014 年 5 月 14 日から 2014 年 7 月 15 日の 500 万投稿を収集し、スパムの正解/不正解の根拠は、凍結済みアカウントであるかどうか、認証済みアカウントであるかどうか、投稿やプロフィールやスクリーンネームに特定の単語が含まれているか（follow、followback 等）などをもとにしている。

Gao らの研究 [9] は Facebook および Twitter を対象にした。彼らはオフラインスパム分析で用いるキャンペーン検出手法を応用し軽い動作で検出可能なオンラインスパム分析手法を提案している。Facebook については、2008 年 1 月から 2009 年 6 月までの 1 億 8700 万投稿（およそ 350 万ユーザ）を収集し、Twitter では、*What the Trend*(注5) というウェブサイトから話題性の高いトピックを選び、そのキーワードを含む投稿を 1700 万収集している。トピックの選択期間は 2011 年 6 月 1 日から 2011 年 7 月 21 日としている。スパムであるかどうかの判定については、Facebook の場合では先行研究によって得られた結果を用い、約 20 万投稿をスパムと取り扱っている。Twitter では投稿に含まれる短縮 URL を確認し、短縮 URL サービスが規約違反で展開を停止している URL を投稿に含むものを有害であるとみなしている。特徴量に関しては、スパムである投稿やアカウントやコミュニティと、そうでないものについて特徴ごとに分布を作成し、投稿の文字数や単語数や平均単語長といった特徴量では識別が困難であるとした一方、ネットワークの特徴やインタラクションの特徴、投稿間隔や URL 数などの特徴からは識別が可能であると述べている。

本章で紹介した論文は、凍結されているユーザをスパムであるとしているものが多いが、本研究ではユーザが意図しないリンク誘導（広告、アフィリエイトブログ、アダルトサイトなど）を含んでいるかどうかをもとに、人手によるスパム判定を行っている。また、本研究では、時間的な挙動の変化を考慮してスパムの検出を試みている。

3 Twitter におけるスパム

本研究では、代表的なマイクロブログである Twitter に出現するスパムを対象とする。Twitter 上でのどのような行為がスパムに該当するかは公式に発表されている [3]。具体的には、自動ツールによる過剰な投稿やフォロー行為、不正利用を目的とした複数アカウントの取得などがスパム行為と判定されうる。

スパム行為の中でも、特に本研究ではリンク誘導型スパムと呼ばれるものを対象とする。リンク誘導型スパムとは、アダルトサイトやアフィリエイトサイトといった Web ページへ誘導することを企てるスパムのことをいう。このようなスパムの手口としては、多くのユーザに自分の投稿を見せるために、複数のアカウントを利用しているものが存在している。

本章ではまずソーシャルメディアとしての Twitter の特徴を説明し、ユーザ間で行われるインタラクションについて説明をする。インタラクションにはメンションと リツイートとがあり、リンク誘導型スパムは主にリツイートを使っていることを述べる。そして、リツイート行動の回数に関する情報をもとにしてアカウントを収集しサンプルセットを構成し、それらの中からリンク誘導型スパムを選び出す。そして、選ばれたスパムがどのようなものであるかを述べる。

3.1 Twitter の特徴

Twitter のユーザは、簡単な操作で短文記事を投稿することができ、その記事中に画像や Web ページへのリンクを付加させることもできる。また、ユーザたちは、興味や関心などに基づきユーザを選択し、読む投稿を制御することができる。このユーザを選択する行為をフォローと呼ぶ。Twitter のホーム画面には、フォローしたユーザたちの投稿が最新のものから順番に上から表示される。この投稿の並びをタイムラインと呼ぶ。

3.1.1 Twitter におけるインタラクション

本節では Twitter の特徴のひとつであるインタラクションと呼ばれる投稿について述べる。インタラクションによって、他ユーザと交流したり、他ユーザの投稿を引用したりすることなどができる。インタラクションには二種類存在し、メンションと呼ばれるものと、リツイートと呼ばれるものがある。

(注4)：https://myspace.com/

(注5)：http://www.whatthetrend.com

a） メンション

特に指定しない場合、ユーザの行う投稿はフォローしたユーザのタイムラインに流れてくるのみで、明確な読者を想定しない。しかし、投稿者は固有の ID を用いることで特定のユーザを指定し、そのユーザ宛の投稿であることを明示した投稿を発信することができる。このことを本稿ではメンションと呼ぶ。一般にはリプライと呼ぶこともある。メンションを受けたユーザは、本人に分かるように通知が届くようになっている。

b） リツイート

ユーザは、短文記事を投稿するほか、他ユーザの投稿を選択して自分をフォローしているユーザに対して選択した投稿を流すことができる。この機能をリツイートと呼ぶ。共感したり、批判したいと思った投稿や、速報などの有益なコンテンツを含む投稿に対しこのリツイートという操作を行うことで、フォローされているユーザたちのタイムラインにその投稿を表示させることができる。このリツイート構造は連鎖的になることも多く、情報拡散と呼ばれることもある。

3.1.2 インタラクショングラフ

前述したインタラクション関係は、ネットワークグラフ構造として表現することができる。つまり、ユーザをノードと、ユーザ間のインタラクションを有向エッジとみなすことができる。このようにして構成されるグラフをインタラクショングラフと呼ぶ。インタラクショングラフのエッジが張られる条件やエッジの重みについては問題設定によって異なっており、本研究のものに関しては 4 章で述べる。

3.2 スパムユーザとその分類

リンク誘導型スパムユーザたちは自分の誘導リンクを含む投稿をできるだけ多くのユーザに見せたいと考える。それを達成する目的で、ユーザたちの興味を惹く情報（テキストや画像）を無断転載し、フォロワーを増やそうと企んでいるものもいれば、URL の含まれる投稿を複数のアカウントで結託してリツイートしあうことで結託したアカウントたちをフォローしているユーザたちのタイムラインにその投稿を流すことを行っているものも存在している。

本説では筆者の判定によりスパムと判定されたアカウントについて、スパムの手法および投稿に含まれるリンク誘導先の観点から分類を行う。筆者によるスパムアカウント選出方法の詳細は 3.3 節で述べる。

3.2.1 スパム手法

本項ではリンク誘導型スパムの具体的な手法について述べる。リンク誘導型スパムたちは、アフィリエイトブログ、アダルトサイト、スマートフォン向けアプリケーション、アフィリエイトブログ、美容商品のオンラインショッピングサイトなどに対するリンクを投稿に含める。誘導リンクの投稿および拡散方法として、次のようなものがある。

- **単独型** 誘導リンクを自分で投稿しているもの。後述の支援型と結託している可能性が考えられる。
- **支援型** 誘導リンクの含まれる投稿をリツイートし、自分のフォロワーに見せているもの。後述の「興味を惹くコンテンツ」を投稿する中にこのような誘導リンクの含まれる投稿のリツイートを織り交ぜているアカウントも存在している。
- **相互型** 上記の単独型と支援型の両方を行っているもの。

また、スパムユーザにはフォロワーやリツイートを獲得する目的で食事や動物などの画像コンテンツや、まとめサイトやニュース速報へのリンクなどといった興味を惹くコンテンツを投稿したり、その投稿をリツイートしたりするものが存在している。

- **単独型** コンテンツを自分で投稿しているもの。他のユーザの投稿から転載されているものが多い。
- **支援型** そのコンテンツをリツイートしているもの。
- **相互型** 上記の単独型と支援型の両方を行っているもの。

3.2.2 スパムの挙動変化

今回の検出の対象となったアカウントの中にはあるタイミングを境に投稿パターンが変化するものがあった。具体的には、誘導リンクを全く投稿せず、画像を投稿していたアカウントが急に他ユーザの誘導リンク入りの投稿をリツイートをするようになるものである。これは一般ユーザに対しスパムであることを一時的に隠すことで怪しまれずにフォロワーを獲得することを目論んでいたものと推察される。

3.3 リンク誘導型スパムユーザ調査

本研究では、2014 年 7 月 1 日から 2014 年 7 月 7 日までの期間のスパムユーザを得ることを行った。そのための手順を述べる。

まず、当研究室で収集している Twitter の投稿データのうち、期間中に http という文字列を含む投稿を取得する。この期間に取得した投稿数は 1 億 4712 万 8138 投稿、そのうち http という文字列を含むものは 3895 万 9759 投稿であった。これらの投稿をもとに、以下の方法で候補となるアカウント群を得る。

- 期間中、URL 入のツイートをリツイートした回数が多いアカウント上位 600
- 期間中、URL 入のツイートがリツイートされた回数が多いアカウント上位 600
- 期間中、URL 入のツイートのリツイート回数が多いツイートを順に並べ、重複を取り除いて得られた 600 アカウント

ただし、投稿期間中に一定数以上の日本語文字が含まれていないアカウントおよび、期間中の投稿群に存在しないアカウントは除外した。この方法でユーザを選び出したのち、重複を取り除くことで 1507 アカウントを得て、残った全アカウント全員に対し、投稿に含まれるリンク先や投稿画像などをもとに筆者によりスパムであるかどうかを判定する。1507 アカウントのうち、117 アカウントのスパムがあった。表 1 に結果をまとめて示す。リンクについては、単独型がほとんど存在せず、支援、相互が 9 割以上を占めていた。

4 スパムユーザ検出手法

4.1 概　　要

3 章でも述べたように、リンク誘導型スパムの攻撃の目的は、投稿に含めた URL で自分の Web ページに誘導することである。その目的を達成するためにはできるだけ多くのユーザに対して目的の URL を見せる必要があり、そのためにインタラク

表 1: 誘導手法と誘導先に基づくスパムアカウントの分類。各カラムの括弧の外側の数字が個数で、括弧内の数字部分 (S.**) はそのうち 2015 年 2 月 4 日現在で凍結されているアカウントの個数を表す

誘導リンク ＼ コンテンツ	支援	相互	単独	合計
支援	34（S.12）	32（S.14）	5（S.1）	71（S.27）
相互	5（S. 0）	36（S.15）	1（S.0）	42（S.15）
単独	0（S. 0）	1（S. 0）	3（S.1）	4（S. 1）
合計	39（S.12）	69（S.29）	9（S.2）	117（S.43）

ションにおいて工夫を凝らしている。

そこで本研究ではインタラクショングラフの構造をもとにユーザごとに特徴量を得て、スパムであるか否かを判定する。また、インタラクショングラフの特徴のほか、投稿に含まれるリンクの個数や割合なども特徴量として有効であると考えられる。そこで本研究では次の 3 種類の特徴量を用いる。

- 基本特徴量
- メンショングラフ（インタラクションがメンションであるインタラクショングラフの本稿での呼称）の特徴量
- リツイートグラフ（インタラクションがリツイートであるインタラクショングラフの本稿での呼称）の特徴量

4.2 特 徴 量

4.2.1 基本特徴量

ユーザごとに表 2 に示す特徴量を算出する。

投稿中に含まれるメンションの割合や、URL を含む投稿の割合などはスパムアカウントとそうでないアカウントとで分布に差が見られる。各ユーザに対する投稿中のメンションの割合についての分布を図 1 に示す。赤く塗られているものがスパムと判定されたアカウントのほうの分布である。スパムアカウントのほうがメンションの割合が低いものが多く存在することがわかる。同様に各ユーザごとの投稿中の URL 率を図 2 に示す。スパムのほうは分布のピークが右に寄っている。

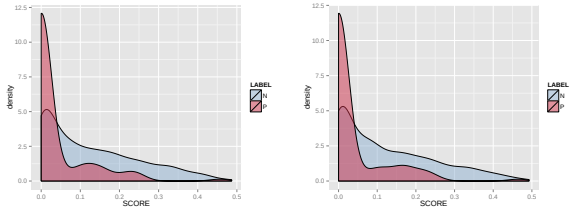
ユーザ ID とは、アカウントが新規に作られた順番にアカウントごとに割り振られる固有の番号であり、この番号が大きいほど新しいアカウントである。この番号を特徴量として導入することにより、同時期に作成された複数のスパムアカウントの検出精度向上が期待される。

4.2.2 インタラクショングラフの特徴量

本研究では次のようにしてインタラクショングラフを構成する。

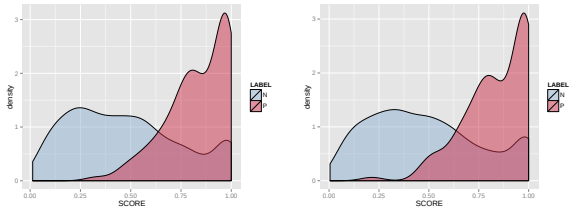
- 各ノードはユーザとする
- エッジは、期間中に一度でもインタラクションがあればそのユーザ間に張られる
- エッジのウェイトは、その期間中のインタラクションの回数とする
- エッジの向きを考慮する必要がある場合は、インタラクションを起こしたユーザから受けたユーザの方向をエッジの方向と定める

メンションとリツイートについてそれぞれインタラクショングラフを用意する。本稿では、インタラクションがメンション



(a) 2014 年 6 月 3 日から 2014 年 7 月 7 日まで　(b) 2014 年 7 月 1 日から 2014 年 7 月 7 日まで

図 1: 各サンプルユーザごとの、投稿中のメンションの割合の分布
（赤：スパム　青：非スパム）



(a) 2014 年 6 月 3 日から 2014 年 7 月 7 日まで　(b) 2014 年 7 月 1 日から 2014 年 7 月 7 日まで

図 2: 各サンプルユーザごとの、URL を含む投稿の割合の分布
（赤：スパム　青：非スパム）

表 2: 基本特徴量

#	特徴量	スケール
1	期間中の通常ツイート（総ツイート数からメンションとリツイートの回数を引いたもの）の回数	対数
2	期間中のメンションの回数	対数
3	期間中のリツイートの回数	対数
4	期間中のツイート総数に対する通常ツイートの比率	等倍
5	期間中のツイート総数に対するメンションの比率	等倍
6	期間中のツイート総数に対するリツイートの比率	等倍
7	期間中にメンションを受けた回数	対数
8	期間中にリツイートを受けた回数	対数
9	期間中の投稿のうち、URL を含むものの個数	対数
10	期間中の投稿のうち、URL を含むものの割合	等倍
11	ユーザ ID	線形

であるものをメンショングラフと呼び、リツイートであるものをリツイートグラフと呼ぶ。

インタラクショングラフに関する特徴量は表 3 の通りである。

a） 2-hop 特徴量

各ノードに対し、2-hop 先までに存在するノードの個数を特徴量として導入する。この特徴量のことを本稿では 2-hop 特徴量と呼ぶ。各インタラクショングラフごとの、スパムと非スパムの 2-hop 特徴量の密度分布を図 3、図 4、図 5、図 6 に示す。横軸が特徴量の大きさに対応し、縦軸が頻度に対応する。図 5b および図 5c から、今回のデータセット中には非スパムユーザのほうが 2-hop 特徴量が大きくなるものが多いことがわかる。これは今回のスパムユーザ候補に含まれるユーザで、多くの一般ユーザに投稿が拡散されたユーザの影響によるものとだと考えられる。

b） クラスタリング係数

クラスタリング係数 [13] とはネットワーク解析に用いられる指標であり、あるノード i について、次数（向きは無視する）を d_i とし、隣接ノード間に張られているエッジの本数を N としたとき

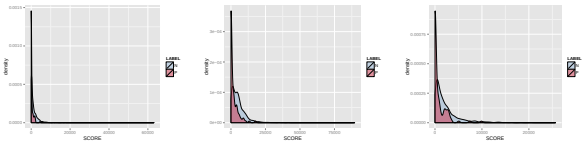
$$C_i = \frac{2N}{d_i(d_i - 1)}$$

で定められる。各インタラクショングラフごとの、スパムと非スパムのクラスタリング係数の密度分布を図 7、図 8 に示す。

4.3 時間変化に関する特徴量

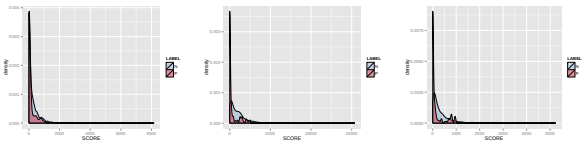
本実験では、時期に依存する特徴量を加えている。具体的には、次に挙げた期間ごとに、前節で述べた特徴量を算出する。インタラクショングラフも、次に挙げた期間に対して取得する。そして、1 週間刻みの特徴量については、5 期間分の平均、標準偏差を算出し 5 週間ぶんの特徴量として導入する。

図 9 に、リツイートグラフ上の各ユーザごとの 2-hop 特徴量について、横軸に 7 日間のものを、縦軸に 35 日間のものをプロットした散布図を示す。また、図 10 には、リツイートグラフ上の各ユーザごとの 2-hop 特徴量について、横軸に 5 期間の平均を、縦軸に 5 期間の標準偏差をプロットした散布図を示す。同様の方法でクラスタリング係数についてプロットした散



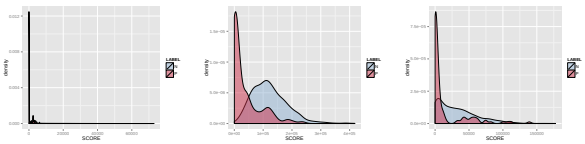
(a) 出方向　(b) 方向無し　(c) 入方向

図 3: 2-hop 特徴量（2014 年 6 月 3 日から 7 月 7 日までの メンション）



(a) 出方向　(b) 方向無し　(c) 入方向

図 4: 2-hop 特徴量（2014 年 7 月 1 日から 7 月 7 日までの メンション）

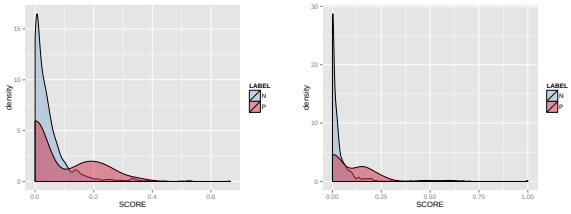


(a) 出方向　(b) 方向無し　(c) 入方向

図 5: 2-hop 特徴量（2014 年 6 月 3 日から 7 月 7 日までの リツイート）

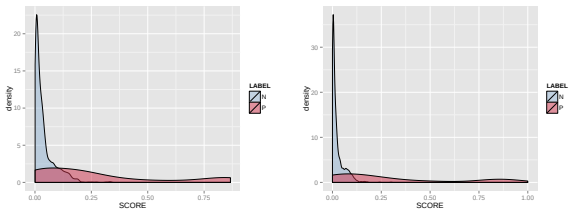
(a) 出方向　(b) 方向無し　(c) 入方向

図 6: 2-hop 特徴量（2014 年 7 月 1 日から 7 月 7 日までの リツイート）



(a) 35 日　(b) 7 日

図 7: メンショングラフ上における、各サンプルユーザごとのクラスタリング係数



(a) 35 日　(b) 7 日

図 8: リツイートグラフ上における、各サンプルユーザごとのクラスタリング係数

表 3: メンショングラフ、リツイートグラフの特徴量

#	特徴量	スケール
1	入次数	対数
2	出次数	対数
3	エッジの向きを無視した次数	対数
4	自分を含む隣接ノードとの間に張られているエッジの本数	対数
5	上記 4 のエッジのウェイト（インタラクション回数）の総和	対数
6	クラスタリング係数	等倍
7	2-hop 先までに含まれるノード数（入方向）	対数
8	2-hop 先までに含まれるノード数（出方向）	対数
9	2-hop 先までに含まれるノード数（方向無視）	対数

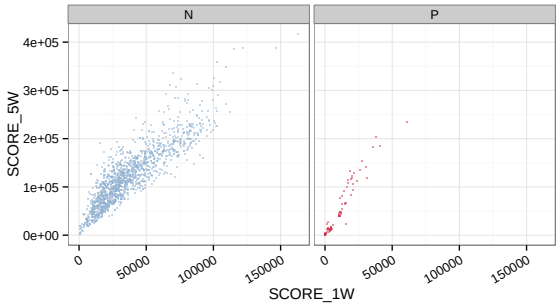


図 9: リツイートグラフにおける、各ユーザごとの 2-hop 特徴量　横軸：7 日間　縦軸：35 日間

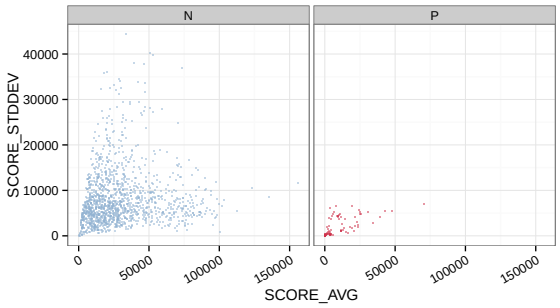


図 10: リツイートグラフにおける、各ユーザごとの 2-hop 特徴量　横軸：5 期間の平均　縦軸：5 期間の標準偏差

布図を図 11 および図 12 に示す。このことから、時間変化に関わる特徴量が分類に有用であると期待することができる。

本実験で用いるインタラクショングラフのノード数、エッジ数を表 4 に示す。

5　実験および結果の考察

5.1　概　　要

4 章で述べたように、特徴量として時間変化に関わるものが分類に有用であると考えられるため、そのような特徴量を考慮して実験を行う。

分類器は LIBSVM [7](注6) の C-SVC を使用し、カーネル関数として RBF カーネルを用いた。パラメータは $C : 2^{-2}, \ldots, 2^{14}$

(注6)：http://www.csie.ntu.edu.tw/~cjlin/libsvm/

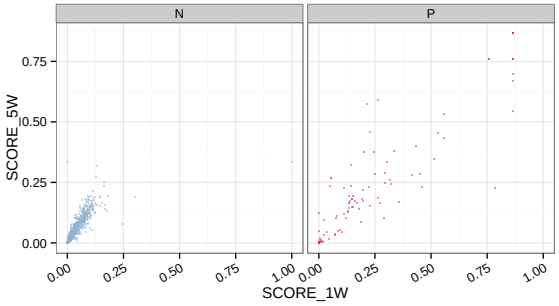


図 11: リツイートグラフにおける、各ユーザごとのクラスタリング係数　横軸：7 日間　縦軸：35 日間

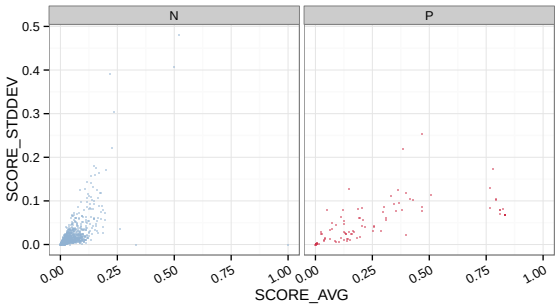


図 12: リツイートグラフにおける、各ユーザごとのクラスタリング係数　横軸：5 期間の平均　縦軸：5 期間の標準偏差

表 4: 本実験にて用いるインタラクショングラフのノード数、エッジ数

（上）メンショングラフ（下）リツイートグラフ

期間	ノード数	エッジ数
2014 年 6 月 3 日〜6 月 9 日	989,642	7,544,849
2014 年 6 月 10 日〜6 月 16 日	990,986	7,742,173
2014 年 6 月 17 日〜6 月 23 日	980,843	7,519,900
2014 年 6 月 24 日〜6 月 30 日	981,064	7,532,076
2014 年 7 月 1 日〜7 月 7 日	968,494	7,371,183
2014 年 6 月 3 日〜7 月 7 日	1,231,669	24,120,783

期間	ノード数	エッジ数
2014 年 6 月 3 日〜6 月 9 日	1,037,466	12,170,407
2014 年 6 月 10 日〜6 月 16 日	1,041,660	12,917,704
2014 年 6 月 17 日〜6 月 23 日	1,032,577	13,123,120
2014 年 6 月 24 日〜6 月 30 日	1,038,341	12,837,406
2014 年 7 月 1 日〜7 月 7 日	1,023,310	12,888,677
2014 年 6 月 3 日〜7 月 7 日	1,260,986	48,748,836

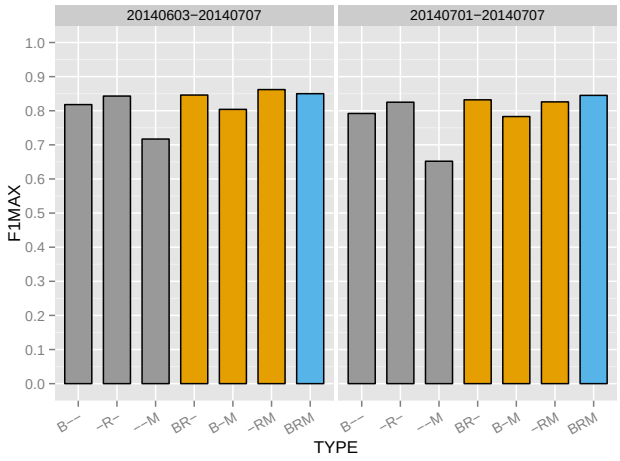


図 13: 特徴量の種類による比較実験結果

および $\gamma : 2^{-12}, \ldots, 2^{6}$ とし、パラメータごとに 10 分割交差検定を 20 回行い、20 回の F 値の平均が最も高かったものを採用した。

5.2 実験結果

まず、時間変化に関係のある特徴量を加えずに行った実験について結果を述べる。2014 年 7 月 1 日から 7 日までの 7 日間の特徴量および、2014 年 6 月 3 日から 7 月 7 日までの 35 日間の特徴量を用いた実験の結果を表 5 および図 13 に示す。図 13 については左枠が 2014 年 6 月 3 日から 2014 年 7 月 7 日までの特徴量を用いた場合の結果であり、右枠が 2014 年 7 月 1 日から 2014 年 7 月 7 日までの特徴量を用いた場合の結果である。枠内の 7 個のグラフの内訳は左から順に、基本特徴量のみを使用したもの、リツイートグラフ特徴量のみを使用したもの、メンショングラフ特徴量のみを使用したもの、メンショングラフ特徴量以外を用いたもの、リツイートグラフ特徴量以外を用いたもの、基本特徴量以外を用いたもの、全特徴量を用いたもの、である。7 日間の特徴量を用いた場合では全特徴量を用いた場合が最も F 値が良いが、35 日間の特徴量の場合だと基本特徴量を使わない場合が一番 F 値が良いという結果になっている。

そこで、先述の 2 種類の特徴量（7 日間の特徴量と 35 日間の特徴量）を統合し、さらに時間特徴（7 日間特徴量の 5 期間ぶんの、平均・標準偏差・変動係数）を加えた実験の結果を表 6 に示す。7 日間の特徴量としては、全特徴量を用いた。先述の実験で最も高い F 値を記録した組み合わせであるからである。35 日間の特徴量としては全特徴量のほか、先述の実験で最も高い F 値を出したリツイートグラフとメンショングラフとの組み合わせのみのものも用いた。表中の t が時間特徴を加えたことを表している。僅かではあるが、時間特徴の追加により結果が良くなっている。

5.3 考　　察

まず、特徴量の種類（基本特徴量、メンショングラフの特徴量、リツイートグラフの特徴量）を変えた実験の結果から、リ

表 5: 特徴量の種類による比較実験結果

期間	B	R	M	F 値	Prec.	Recall
7 月 1 日〜7 月 7 日	○	×	×	0.792	0.908	0.702
7 月 1 日〜7 月 7 日	×	○	×	0.825	0.908	0.756
7 月 1 日〜7 月 7 日	×	×	○	0.652	0.655	0.650
7 月 1 日〜7 月 7 日	○	○	×	0.832	0.903	0.771
7 月 1 日〜7 月 7 日	○	×	○	0.783	0.876	0.707
7 月 1 日〜7 月 7 日	×	○	○	0.826	0.930	0.743
7 月 1 日〜7 月 7 日	○	○	○	0.845	0.891	0.803
6 月 3 日〜7 月 7 日	○	×	×	0.818	0.847	0.792
6 月 3 日〜7 月 7 日	×	○	×	0.843	0.952	0.756
6 月 3 日〜7 月 7 日	×	×	○	0.717	0.761	0.677
6 月 3 日〜7 月 7 日	○	○	×	0.846	0.913	0.788
6 月 3 日〜7 月 7 日	○	×	○	0.804	0.878	0.741
6 月 3 日〜7 月 7 日	×	○	○	0.862	0.885	0.840
6 月 3 日〜7 月 7 日	○	○	○	0.850	0.901	0.804

表 6: 7 日間の特徴量と 35 日間の特徴量と時間特徴を統合させた場合の比較実験結果

7/1〜7/7	6/3〜7/7	F 値	Prec.	Recall
—	-RM	0.862	0.885	0.840
—	BRMt	0.857	0.947	0.783
—	-RMt	0.861	0.884	0.839
BRM	-RM	0.860	0.940	0.792
BRM	-RMt	0.866	0.963	0.787
BRM	BRMt	0.860	0.912	0.814

ツイートグラフの特徴量が有効に作用することがわかる。一方メンショングラフの特徴量のみを用いた場合が最も精度が低くなっているものの 65% および 72% 程度の F 値が得られている。

また、7 日間の特徴量と 35 日間の特徴量とを同時に導入した実験結果については、僅かながら精度向上がみられるものがあった。ただ実験結果から分かるように精度向上は小さく、精度が向上していないものもあるため非スパムユーザの特徴量の時間変化のばらつきが直接的な精度向上に寄与するとは言い切れない。

6 まとめと今後の課題

本研究では、意図しない Web ページへ一般ユーザを誘導するリンク誘導型スパム検出タスクに取り組んだ。まず、2015 年 7 月 1 日から 2015 年 7 月 7 日までの期間の投稿から “http” という文字列を含む投稿を約 3900 万集めた。その中からスパムユーザに見られる特徴を持った 1507 ユーザを選出し、人手でスパムアカウントであるかどうかを分類し、117 ユーザがスパムと判定された。スパムの検出は学習ベースの手法を用い、正例と誤例を予め与えて交差検定により適合率と再現率を算出した。用いる特徴量として、リンク誘導型スパムの宣伝行動目的から周辺のインタラクション構造が特異なものになることに着目して、インタラクショングラフと呼ばれる、アカウント間のインタラクション構造に関わる特徴量を用いたほか、ユーザごとの投稿数や URL 投稿回数などの基本的な特徴量も加味した。さらに本研究ではユーザごとの 1 ヶ月にわたる挙動に着目し、時間による違いがスパムユーザでないものに多く見られることから時間変化を考慮した特徴量を加えた。そして、実験セット中に含まれるスパムユーザを高い精度と再現率を保って検出することに成功した。

今後の課題としては、より短い期間で実現可能な検出方法について検討するほか、ユーザ単位ではなく投稿単位でのスパム検出方法について検討したい。

文　献

[1] Facebook reports forth quarter and full year 2014 results. `http://investor.fb.com/releasedetail.cfm?ReleaseID=893395` 2015 年 2 月リンク先確認.

[2] Twitter reports third quarter 2014 results. `https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=878170` 2015 年 2 月リンク先確認.

[3] Twitter ヘルプセンター — twitter ルール. `https://support.twitter.com/articles/253501-twitter` 2015 年 2 月リンク先確認.

[4] What is web 2.0? `http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html` 2015 年 2 月リンク先確認.

[5] Jonell Baltazar, Joey Costoya, and Ryan Flores. The real face of koobface: The largest web 2.0 botnet explained. `http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp_the-real-face-of-koobface.pdf` 2015 年 2 月リンク先確認, 2009.

[6] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, Vol. 6, p. 12, 2010.

[7] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, Vol. 2, No. 3, pp. 27:1–27:27, May 2011.

[8] Pei-Chi Chen, Hahn-Ming Lee, Hsiao-Rong Tyan, Jain-Shing Wu, and Te-En Wei. Detecting spam on twitter via message-passing based on retweet-relation. In Shin-Ming Cheng and Min-Yuh Day, editors, *Technologies and Applications of Artificial Intelligence*, Vol. 8916 of *Lecture Notes in Computer Science*, pp. 56–65. Springer International Publishing, 2014.

[9] Hongyu Gao, Yan Chen, Kathy Lee, Diana Palsetia, and Alok N Choudhary. Towards online spam filtering in social networks. In *NDSS*, 2012.

[10] McCord M. and Chuah M. Spam detection on twitter using traditional classifiers. In *Autonomic and Trusted Computing*, BBATC ’11, pp. 175–186. Springer, 2011.

[11] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC ’10, pp. 1–9, New York, NY, USA, 2010. ACM.

[12] Wang and Alex Hai. Don’t follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, SECRYPT ’10, pp. 1–10, July 2010.

[13] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’ networks. *nature*, Vol. 393, No. 6684, pp. 440–442, 1998.