

# Twitter のスパム検知機能となりすまし検知機能を強化する アプリケーション LookUpper の開発と評価

若井一樹<sup>†1</sup> 佐々木良一<sup>†2</sup>

Twitter におけるスパム行為となりすまし行為の検知手法を提案する。これらの検知手法はスパム行為となりすまし行為の様々な特徴から検知対象であるか判定する項目を複数個作成し、それら項目を数量化理論の適用によって最適な項目組み合わせを選定することによって検知するものである。またこれら手法を実装するとともに、検知結果をユーザにわかりやすく提示するよう Twitter の表示系を強化したアプリケーション LookUpper の開発と評価を行った。この結果、本検知手法ではスパム行為となりすまし行為どちらも 90%以上の的中率で検知することが可能であった。LookUpper の開発と評価について、本検知手法を実装し検知結果をわかりやすく表示する機能を開発し、被験者 10 人によってなされた LookUpper のユーザビリティに関する実験結果から全体的に高い評価を得るとともに、今後 LookUpper の改良を行っていくためのアイデアを導く種々のコメントが寄せられた。

## Development and Evaluation of the Application Program “LookUpper” to Improve the Spam Discovery Function and the Spoofing Discovery Function of Twitter

KAZUKI WAKAI<sup>†1</sup> RYOICHI SASAKI<sup>†2</sup>

We propose the spam discovery method and the spoofing discovery method for Twitter. These discovery methods are achieved by using items created from the various features of spam and spoofing select the best combination items by application of mathematical quantification theory. In addition, we also performed the development and evaluation results of the application program named LookUpper with methods described above and enhanced display system of Twitter to provide clarity to the user detection result. We could know that the predictive values for discovering spam and spoofing were more than 90% each. From the experiments with regards to the usability of LookUpper performed by 10 subjects, we could obtain high ratings overall with various comments leading to the ideas to improve the LookUpper in future.

### 1. はじめに

ソーシャルメディアが発達し世界中の人々とコミュニケーションが可能となった。なかでも Twitter は利用者を急激に増やしており、2013 年には月間アクティブユーザ数が全世界で 2 億人を突破した[1]。一方で Twitter を利用した迷惑行為も増えている。例としてスパム行為（以後、スパムと記す）やなりすまし行為（以後、なりすましと記す）が挙げられる。

スパムの被害についてただ宣伝広告を受け取るだけでなく、ユーザ自身が知らないうちにスパムに加担するケースが増えている。そのスパムの手口として、一般ユーザにスパムを発信するアプリケーションを気づかれないよう認証させてユーザのアカウントを乗っ取り、スパム被害を拡大させるツイートを投稿させるものもある[2]。このようなスパムの被害について Twitter 社はデータを公表しており、2010 年 1 月の時点では 1 日に約 5000 万回、1 秒あたり約

600 回ものツイートが発信されているが、その内の 2%はスパムによるツイートである[3][4]。つまり 5000 万ツイートの内約 100 万ツイートがスパムによるツイートであり、毎秒約 12 回もスパムによるツイートがされている。

なりすましの被害について政財界や芸能界の著名な人のなりすましが特に多い[5]。公人の発言は一般の人と比べ影響力が強いため、なりすましが発生すると被害が大きくなる危険性がある。この問題に対し Twitter 社は著名な人々に対して認証バッジを導入し Twitter のアカウントが本人であるとする指標を提示しているが[6]、著名な人でも認証バッジが無いアカウントは存在している。

Twitter 利用者はスパムによって必要としない情報が大量に送られ、なりすましによって誤った情報を手に入れてしまう危険性がある。しかしこれらの行為をするアカウントに対しフォロー解除やブロックなどの管理を行うとき、Twitter の表示系では操作が困難な場合がある。Twitter の表示系はフォローしているユーザのツイートを最新のものから順に表示するものが多く、アカウントの数が増えるとスパムやなりすましを行うアカウントを発見してもフォロー解除などの管理を行うことが困難である。最新のものからアカウントを表示させるのではなく特定のアカウントを一

<sup>†1</sup> 東京電機大学大学院  
Tokyo Denki University

<sup>†2</sup> 東京電機大学  
Tokyo Denki University

括で見やすく表示するような工夫が必要である。

そこで本論文では Twitter におけるスパム検知手法と、なりすましについては発生した時の被害が特に大きいと考えられる著名人のなかで、対象を政治家に絞ったなりすまし検知手法を考案し実際に検知が可能であることを確認し、その後各手法で得られた検知結果をわかりやすく Twitter ユーザに通知するアプリケーション LookUpper の開発と評価について報告する。

## 2. 関連研究

ソーシャルメディアが流行する以前からスパムやなりすましについて多くの研究が行われてきた。

スパムに関する研究について、長谷ら[7]はアフィリエイトに着目したスパムブログ評価方式に関する検討を行っており、主にアフィリエイトリンクアンカーを用いてスパムブログと非スパムブログを区別が可能か検討している。佐藤ら[8]はキーワードの時系列特性に着目したスパムブログを収集し、類型化、データセットの作成をしている。なりすましに関する研究について、石川ら[9]は特定のコミュニケーションサイトを題材に、投稿されたメッセージから著者を推定する手法について述べている。

ソーシャルメディア流行後は、Twitter を題材にした研究が多く行われている。榊ら[10]は Twitter のリスト機能の応用についての報告を行っており、Twitter だけに限っても多くの特徴があることがわかる。吉田ら[11]は Twitter におけるリンクを含むつぶやきの分析を行っており、ニュースなどよりも写真や動画など娯乐的サービスの URL が投稿されやすいと報告している。Java ら[12]や Pak ら[13]はマイクロブログ、主に Twitter を利用したユーザ特性や特定の言語の抽出に関する研究をしており、ユーザのコミュニティレベルに関連付けられる意図の分析や、曖昧な表現な語の分析などを行っている。

これら Twitter の研究には本論文で題材として取り上げているスパムに関する研究もあり、Almaatouq ら[14]は Twitter を題材にオンラインソーシャルネットワーク全体的な視点からスパムが関与する流れを調査している。中村ら[15]は対象を Twitter に絞ったスパムユーザフィルタの開発を行っており、Twitter 社にアカウント停止されていないスパムアカウントに対しても 94.7%の割合で正しく検出している。

本論文ではスパム検知手法となりすまし検知手法、どちらも対象の様々な特徴を複数個挙げ、それら特徴と合致するか総合的に判定するという手法で検知する。これはスパムなど行う側の手法が別の手法に切り替えても対応可能にするためである。関連研究で述べた従来手法と比べ本手法との違いは以下のとおりと考えている。

(1)従来手法では Twitter 社にアカウント停止されていないスパムアカウントを 94.7%の精度で検出しているが、従来手法に劣らず本手法では後述するように判別の中率 98.8%の精度を持つスパム判別手法を確立し、予測対象とした Twitter 社が見逃しているスパムアカウントを全て検出している。

(2)スパムなど行う側の変化に柔軟に対応することができる。すなわち、判定項目を複数作成することでスパムやなりすましの手法が変わり一部の判定項目ではスパムやなりすましであると判定されなくとも、残りの判定項目でスパムやなりすましであると判定が可能であると考えられる。

(3)判定結果により自動でスパムやなりすましを排除するのではなく結果を開発したアプリケーション LookUpper に反映し、スパムやなりすましの可能性があるとしてユーザに検知した該当アカウントを提示し最終的にフォローを外す等の処置をユーザに委ねられるようにしている。

(4)Twitter のなりすましについて、なりすましの実態の調査結果や、SNS 全体としてのなりすましに関する考察[16][17]は存在するが、対象を Twitter に絞った具体的ななりすまし検知手法に関する研究は報告者らの調査した範囲では存在せず、新規性があると考えている。

## 3. 検知手法について

各対象における本手法の詳細について、3.1 節ではスパムに対する検知手法について、3.2 節ではなりすまし検知手法について記す。

### 3.1 スパム検知手法

本論文ではスパムアカウントを、Twitter をコミュニケーションツールとして使用せず一般ユーザが必要としない宣伝広告ばかりを行うアカウントと定義する。本節ではこの定義におけるスパムアカウントの検知手法について記す。

#### 3.1.1 スパムの特徴の分析

Twitter 社はスパムに対しスパムアカウントと判断する基準を公開しており[18]、この基準を元に判定項目を作成した。しかしこの基準は主にフォロー数やツイート内容などに着目したものが多く、この基準だけで判定項目を作成すると、Twitter 社が見逃してしまうスパムアカウントは本手法でもスパム判定し損ねる可能性がある。より高い精度でスパム検知を行うためにはこれら以外の特徴を用いた基準が必要である。そこで別の特徴を用いた判定項目として、投稿元のクライアント名を使った判定項目、自己紹介文を使った判定項目を考案した。

投稿元のクライアント名を使った判定では、ユーザが使用するクライアント名を元に判定する。クライアントとは Twitter のサービスを利用し独自機能を搭載したクライ

ントソフトウェアの総称であり、主に「via ○○」の形で記される。クライアントの例として「Twitter Web Client」や「TweetDeck」、「Janetter」など様々な種類があり、スパムアカウント特有の傾向があると考え判定項目を作成した。

自己紹介文を使った判定ではアカウントのプロフィール欄のうち自己紹介文を元に判定する。自己紹介文ではユーザが自分の趣味や興味あるものなど自由に記載する。「○○に興味アリ」など文として記すもの、「/○/△/」など単語を斜線で区切って記すものなど様々な記入方法があり、スパムアカウント特有の傾向があると考え判定項目を作成した。

さらに作成した判定項目について、項目をフォロー系とツイート系に分け採点が高い場合にさらに加算する判定項目を作成した。各判定項目の詳細を表1に示す。

表1 スпам判定項目の詳細

Table 1 Details of the judgment items of spam.

判定項目	内容	内訳
判定項目1 (ツイート系)	同じ内容をなんども繰り返す場合	Twitter社の判断基準
判定項目2 (ツイート系)	つぶやきの内容が個人的なものではなく、主にリンクばかりである場合	Twitter社の判断基準
判定項目3 (フォロー系)	フォローしている数に対しフォロワーの数が極端に少ない場合	Twitter社の判断基準
判定項目4 (フォロー系)	相互フォロー割合	Twitter社の判断基準
判定項目5	投稿元のクライアント名を使った判定項目	新たに考案した判定項目
判定項目6	プロフィールの特徴を使った判定項目	新たに考案した判定項目
判定項目7	特定パターンで加算 (フォロー系)	新たに考案した判定項目
判定項目8	特定パターンで加算 (ツイート系)	新たに考案した判定項目

### 3.1.2 使用する判定項目の設定

表1の判定項目を用いて実際にスパム検知が可能か数量化理論Ⅱ類を用いて検証する。本検証では株式会社エスミ社のソフトウェア Excel 数量化理論 Ver3.0[19]を用いて、各変数について目的変数は「スパムである」または「スパムでない」の2群、説明変数のアイテム数は「判定項目1」から「判定項目8」までの計8個、カテゴリー数は全てその判定項目でスパムを「検知できる」または「検知できない」の2個と設定した。適用対象はスパムアカウント50件と一般アカウント50件、計100件のアカウントを用いる。

初めにスパムアカウント、一般アカウントそれぞれ40件ずつ計80件を用いてパラメータフィッティングを行い、その後残りの20件を用いて実際にスパムアカウントと一般アカウントを判別できるか予測する。予測するスパムアカウントについては明らかにスパム行為をしているが、

Twitter社にアカウント停止されていないものを用いる。これらアカウントは東京電機大学情報セキュリティ研究室の4名でスパムであるかどうか判断した。

また、予測の際パラメータが多いと過適合の問題が生じるため、各項目の組み合わせについて赤池情報量基準(以下、AICと記す)を求める。AICはモデルの当てはまりの良さを評価する指標であり、以下の式によって求められる最小AICの時のパラメータ数を選択することで、最適なモデルの選択が可能となる。

$$AIC = -2\ln L + 2k$$

Lは最大尤度、kはパラメータの数である。これをすべての項目の組み合わせについて実施しスパム判定項目の最適な組み合わせを選出する。その後最小AICの時のパラメータ数を選択することで、なりすまし判定項目の最適な組み合わせを選択する。結果を表2に示す。

表2 パラメータフィッティングの結果

Table 2 Results of parameter fitting.

説明変数	カテゴリー名	カテゴリースコア
判定項目1	検知できる	0.505
	検知できない	-0.414
判定項目2	検知できる	0.661
	検知できない	-0.661
判定項目6	検知できる	0.176
	検知できない	-0.167
判定項目8	検知できる	0.250
	検知できない	-0.205
AIC		1.673
判別の中率		98.8%
判別の中点		0.0000
判定項目1 (ツイート系)	同じ内容をなんども繰り返す場合	
判定項目2 (ツイート系)	つぶやきの内容が個人的なものではなく、主にリンクばかりである場合	
判定項目6	プロフィールの特徴を使った判定項目	
判定項目8	特定パターンで加算 (ツイート系)	

赤池情報量基準を用いてパラメータの推定を行った結果、判定項目の全ての組み合わせ255パターンの中で最も低いAICの値は1.673となり、その時の組み合わせは判定項目1,2,6,8で、判別の中率は98.8%となった。項目数が少なく判別の中率も高いので、判定項目1,2,6,8の組み合わせをスパム判定項目の最適な組み合わせとする。

続いて、表2の判定項目のカテゴリースコアを用いて残り20件のアカウントに対し検知可能か、判別の中点により予測する。予測する各アカウントのサンプルスコアが判別

的中点 0.0000 より大きければスパムである群, 小さければスパムでない群に判別される. 結果を表 3 に示す.

表 3 スпамアカウントと一般アカウントの  
 予測結果

Table 3 Prediction result of spam account and general account.

		予測結果	
		スパム である	スパム でない
使用 した ア カ ウ ン ト	スパム アカウント(10)	10	0
	一般 アカウント(10)	1	9

スパムアカウントと一般アカウントを正しく識別できるか予測したところ, スпамアカウントについて全てスパムであると検知できた. 一般アカウントについて 10 件中 9 件スパムでないと検知でき, 正解率は 95%であった.

### 3.2 なりすまし検知手法

本論文ではなりすましアカウントを, なりすましの対象である本人と同じ名前を名乗り, ①本人の顔写真をアイコンに設定している, ②本人の経歴を自己紹介文に記述している, ①または②どちらか一方でもあてはまるアカウントをなりすましアカウントと定義し, 実際になりすましアカウントであるかどうかは, 東京電機大学情報セキュリティ研究室の 4 名で合意を取りつつ判断した. 本物のアカウントについては, 自己紹介文で政治家であることを表記し, かつ政治家らの Twitter アカウントのまとめ[20][21]に表記があるものとした. 本節では上記の定義におけるなりすましアカウントの検知手法について記す.

#### 3.2.1 なりすましの特徴の分析

Twitter におけるなりすまし被害は増加しているが, 特に発言力の大きい政治家などのなりすましは被害が大きくなる可能性が高い. 2013 年 4 月の公職選挙法改正[22]によって Twitter が選挙活動にも取り入れられるようになり, 情報発信ツールとしてさらに活用される一方でなりすましによる被害によって誤った情報が流布される危険性がある. 1 章でも述べたとおり Twitter 社は著名人に対し認証バッジを導入し本人であるとする指標を提示しているが, 著名な人でも認証バッジが無いアカウントは存在している. そこで本論文では対象を政治家に絞り, なりすまし検知手法について Twitter の認証バッジに依存せずスパム検知方式と同様になりすましの様々な特徴を複数個挙げ, それら特徴と合致するか総合的に判定するという手法で検知する.

初めになりすましの特徴の分析から行う. 本物の政治家アカウント 30 件となりすましアカウント 15 件を用いて, どちらのアカウントも最新 200 件のツイートを取得, ツイートが 200 件に満たない場合は全ツイートを用いて調査した. 調査の結果以下の 4 点について本物のアカウントとなりすましアカウントで違いがあった.

#### (1) ツイートを行う時間帯

ツイートを行う時間帯について, どちらのアカウントも深夜未明にかけてツイート数が少なく, 昼時間帯 (11 時~13 時) にツイート数が多かった. しかし夕方の時間帯 (16 時~18 時) において本物のアカウントはツイート数が多いのに対し, なりすましアカウントはツイート数が少なかった. 政治家ではないが公共の Twitter アカウントを対象に上野ら[23]が自治体職員のツイートをを行う時間帯を調査しており, アカウントの運営者である自治体職員の活動する時間帯によってツイートされる時間帯が依存されると報告している. そこで政治家について活動時間帯を調査したところ[24][25][26][27], 夕方の時間帯で会合やミーティング, 事後処理を行っている場合が多く, 午前中に議会を終え夕方の来客対応や会合の合間にその日の活動をまとめ, フォロワーや支援者に報告する時間帯と重なったため夕方の時間帯でツイート数が多くなったのではないかと考えられる.

#### (2) スクリーンネームの規則性

スクリーンネームとは@記号から始まる英数字のみでしか設定できないアカウントの名前のことである. 日本語でも設定できるネームと異なり, スクリーンネームは他のアカウントと重複できない. このスクリーンネームの規則性について, 本物のアカウントは自身の姓名フルネームをローマ字表記している場合が多いのに対し, なりすましアカウントは姓または名に加えて意味のない数字や文字列をつなげて記している場合が多かった.

#### (3) 使用するクライアントの数

使用する PC 等のクライアントの個数について, 本物のアカウントを持つ政治家は議会や会合など様々な場所で活動しており, 常に同じクライアントからツイートされているとは考えにくい. 例えば同じアカウントであっても PC からのツイートと携帯端末からのツイートでクライアント名が異なるため, 使用するクライアントの個数は多いと考えた. 反対になりすましアカウントについて, 愉快犯としてなりすまし行為を行うものも多く存在しており, 政治家ほど複数のクライアントを用いてなりすまし行為を行うとは考えにくく, 使用するクライアントは少ないと考えた.

調査の結果本物のアカウントが使用するクライアントの数は最大 8 個と幅広く分布するのに対し, なりすましアカウントが使用するクライアントの数は最大でも 3 個しか使

用していなかった。さらに本物のクライアントについて、Facebook など他の SNS と連携したクライアントを使用しているものが多いことが判明した。またなりすましアカウントについて、ボットを用いてなりすましを行うアカウントが少数存在した。

#### (4) お気に入り、リツイートの数

お気に入りとはユーザがツイートを入ったとき後で読み返せるようにツイートを保存することが可能な機能のことである。リツイート（以後、RT と記す）とは他のユーザのツイートを自分のフォロワーに再投稿する機能のことである。なお RT には非公式 RT と呼ばれる他のアカウントのツイートを引用する形で自身のものとしてツイートするものがあるが、本論文では非公式 RT は対象外とする。お気に入り、RT は情報の拡散度合いを把握する指標として役立ち、著名人のツイートなどは多くのユーザから反響があると考え調査した。調査の結果本物のアカウントについてお気に入りされた数は最小でも 100 以上の反応、RT された数は最小でも 1 万以上の反応があり、最大ではお気に入り RT どちらも万単位の反応があった。一方なりすましアカウントについて、お気に入り RT どちらも最小ではまったく反応がなく最大でも 1000 以上の反応にとどまった。

以上の調査結果から、表 4 に示すとおりなりすまし判定項目を作成した。次節でこれら判定項目を用いて本物のアカウントとなりすましアカウントがそれぞれ識別可能であるか実験する。

表 4 なりすまし判定項目の詳細

Table 4 Details of the judgment items of spoofing

判定項目	内容	内訳
判定項目 1	昼時間帯 (11 時~13 時) にツイート多い	ツイートを行う時間帯
判定項目 2	夕時間帯 (16 時~18 時) にツイート少ない	ツイートを行う時間帯
判定項目 3	スクリーンネームに名前と苗字両方とも使用していない	スクリーンネームの規則性
判定項目 4	スクリーンネームに意味不明な数文字列を使用している	スクリーンネームの規則性
判定項目 5	via が 3 個以下である	使用するクライアント
判定項目 6	via が SNS 連携をしていない	使用するクライアント
判定項目 7	ボットを使用している	使用するクライアント
判定項目 8	お気に入りされた数が 100 以下	お気に入り、RT の数
判定項目 9	RT された数が 10000 以下	お気に入り、RT の数

#### 3.2.2 使用する判定項目の設定

スパム検知手法と同様に、表 5 の判定項目を用いて実際になりすまし検知が可能か数量化理論 II 類を用いて検証する。[19] を用いて、各変数について目的変数は「なりすましである」または「なりすましでない」の 2 群、説明変数のアイテム数は「判定項目 1」から「判定項目 9」までの計 9 個、カテゴリ数は全てその判定項目でなりすましを「検知できる」または「検知できない」の 2 個とした。適用対象はなりすましアカウント 50 件と本物のアカウント 50 件、計 100 件のアカウントを用いる。初めになりすましアカウント、一般アカウントそれぞれ 40 件ずつ計 80 件を用いてパラメータフィッティングを行い、その後残りの 20 件を用いて実際になりすましアカウントと本物のアカウントを判別できるか予測する。初めに予測するためのパラメータフィッティングを行う。その後最小 AIC の時のパラメータ数を選択することで、なりすまし判定項目の最適な組み合わせを選択する。結果を表 5 に示す。

表 5 パラメータフィッティングの結果と

AIC によるなりすまし判定項目の最適な組み合わせ

Table 5 Results of parameter fitting and Optimal combination of the judgment items of spoofing based on AIC.

説明変数	カテゴリ名	カテゴリースコア
判定項目 2	検知できる	0.238
	検知できない	-0.306
判定項目 4	検知できる	0.190
	検知できない	-0.181
判定項目 5	検知できる	0.115
	検知できない	-0.173
判定項目 6	検知できる	0.039
	検知できない	-0.135
判定項目 8	検知できる	0.536
	検知できない	-0.536
AIC		90.330
判別的中率		96.3%
判別的中点		-0.1250
判定項目 2	夕時間帯 (16 時~18 時) にツイート少ない	
判定項目 4	スクリーンネームに意味不明な数文字列を使用している	
判定項目 5	via が 3 個以下である	
判定項目 6	via が SNS 連携をしていない	
判定項目 8	お気に入りされた数が 100 以下	

赤池情報量基準を用いてパラメータの推定を行った結果、判定項目の全ての組み合わせ 511 パターンの中で最も低い AIC の値は 90.330 となり、その時の組み合わせは判定項目

2,4,5,6,8 で、判別率の中率は 96.3%となった。項目数が少なく判別率の中率も高いので、判定項目 2,4,5,6,8 の組み合わせをスパム判定項目の最適な組み合わせとする。

続いて、表 5 の判定項目のカテゴリースコアを用いて残り 20 件のアカウントに対し検知可能か、判別率の中心により予測する。予測する各アカウントのサンプルスコアが判別率の中心-0.1250 より大きければなりすましである群、小さければなりすましでない群に判別される。結果を表 6 に示す。

表 6 なりすましアカウントと本物のアカウントの予測結果

Table 6 Prediction result of spoofing account and genuine account.

		予測結果	
		なりすましである	なりすましでない
使用したアカウント	なりすましアカウント(10)	9	1
	本物のアカウント(10)	1	9

なりすましアカウントと本物のアカウントを正しく識別できるか予測したところ、なりすましアカウントについて 10 件中 9 件なりすましであると検知できた。本物のアカウントについて 10 件中 9 件なりすましでないと検知でき、正解率は 90%であった。

#### 4. 表示系の開発と評価

前章でスパム検知手法となりすまし検知手法について述べた。どちらの手法も様々な特徴を元に判定項目を作成、検証実験、数量化理論への適用を行った。95%の正解率でスパムアカウントを、90%の正解率でなりすましアカウントをそれぞれ検知が可能であることを実証した。しかしこれら検知機能を実装する際、検知結果の表示方法について問題がある。

Twitter を利用し続けていけばユーザ自身のフォローやフォロワーは多くなっていく。フォローやフォロワーが多い時、スパムアカウントやなりすましアカウントが紛れていると探し出すのが困難である。多くの対象の中からでもスパムアカウントやなりすましアカウントの発見が可能となる表示系が必要である。しかし Twitter クライアントの多くはツイートを表示するタイムラインの機能に特化しており、フォローやフォロワーなどの管理に優れているものは少ない。Twitter のインタフェースと差別化を図り、単にタイムラインを表示するのではなく自分のフォローやフォロワーを可視化することで、より Twitter のコミュニケーション

ン機能やフォローやフォロワーの管理機能を有効活用できるアプリケーションが必要である。

この問題を解決するにあたり、3 章で述べた検知手法を実装し Twitter の表示系を発展させたアプリケーションとして LookUpper を開発した。

#### 4.1 LookUpper の開発

LookUpper では 3 章で述べた検知手法を用いてスパムであるかまたはなりすましであるか判定し、結果をスコアで算出する。このスコアは点数が高いほどスパムまたはなりすましの可能性が高くなり、LookUpper では 60 点以上でスパムまたはなりすましであると判断している。しかしフィルタを設定し自動でスパムやなりすまし排除するのではなく、検知結果の提示だけ行い最終的なフォロー解除等の処置はユーザに委ねられている。

図 1 に示すように、自身を中心にしてタイムラインに現れるアカウントを対象に、相互フォローやフォローをスコアと同時に配置する。

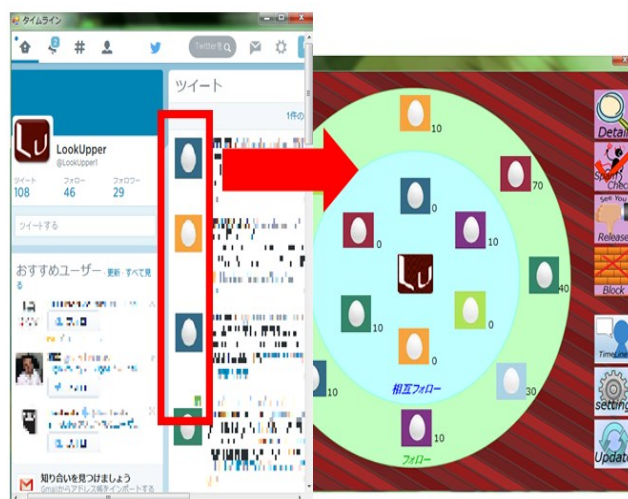


図 1 LookUpper のメイン画面

Figure 1 Main screen of LookUpper.

例えばユーザが Twitter を使用しているときタイムライン上にスパムの疑いがあるアカウントが現れたとする。その時 LookUpper を起動するとタイムラインに現れていたアカウントがスコアとともに LookUpper のメイン画面に表示される。スパムの疑いがあるアカウントのスコアを見て点数が高い場合、ユーザは対象のアカウントをメイン画面右側にあるメニュー欄へドラッグ&ドロップすることで、スパムのスコアに対する詳細メッセージを見ることができる。ユーザが詳細メッセージを見てスパムであると判断した時対象のアカウントをさらにメニュー欄へドラッグ&ドロップすることでフォローの解除やブロックなどの各種管理を行うことが可能となる。

既存の Twitter のアプリケーションに多い、単にツイートを表示させるだけのものではなく、わかりやすく表示することに重点を置いた。これにより LookUpper ではスパム検知と同時に Twitter を見て楽しむことを可能にした。

#### 4.2 LookUpper の評価

LookUpper のユーザビリティなどについて、東京電機大学情報セキュリティ研究室 10 名に 5 段階評価によるアンケートを行った。評価はあらかじめ LookUpper で使用する Twitter アカウントを用意しておき、そのアカウントのフォローするアカウントの中からスパムアカウントを発見するというシナリオで行った。評価結果を表 7 に示す。

表 7 LookUpper の評価  
 Table 7 Evaluation of the LookUpper.

設問	内容	平均値 (n=10)
LookUpper について		
①	LookUpper は使いやすかったか	4.7
②	LookUpper のデザイン、表示されるアカウント等の配置は見やすかったか	4.4
③	スコアの表示はスパム発見に有効か	4.4
④	フォロー解除など管理機能操作は使いやすかったか	4.8
スパム判定について		
①	スコア詳細画面はわかりやすかったか	3.9
②	スコア詳細画面で表示されるコメントは適切か	4.0
③	スコアに対する説明(60 点以上でスパムの可能性が高い)は明示的か	4.0
表示機能について		
①	表示されるアカウントの人数は適切か	4.2
②	自身のフォローや相互フォローは把握しやすかったか	4.4
③	スパムを点数化し可視化するという表示方法は、スパムを見つける上で有効か	4.8

LookUpper についての項目と表示機能についての項目は全体的に高い評価を得ることができた。しかしスパム判定に関しての項目は全体的に少し評価が下がり、特に詳細画面のわかりやすさについて低い評価となった。詳細画面では判定対象となったアカウントに対する各判定項目の傾向からスパムの可能性についてのコメントが表示される。スパムの可能性について、より詳細なコメントが欲しい、具体的にどのようなツイートでスパムの疑いが見られたのか知

りたい、などの意見が寄せられた。

また LookUpper に追加して欲しい機能について任意のアンケートを取ったところ、ツイートを吹き出しで表示する機能や、LookUpper ユーザ同士でスパムアカウント情報を共有する機能などが欲しいといった意見が寄せられた。

今後はスコア詳細画面のわかりやすさについて改良を行いつつ、アンケートで寄せられた追加して欲しい機能について実装していきたい。

#### 5. おわりに

本論文ではスパム検知手法となりすまし検知手法の提案、そしてそれら検知手法を実装し Twitter の表示系を発展したアプリケーション LookUpper の開発と評価を行った。

検知手法についてスパムとなりすましどちらが対象であっても高い精度で検知が可能であったが、いくつかのアカウントにおいて誤検知が発生した。スパム検知手法ではフォロワー数とフォロー数との関係について、なりすまし検知手法ではスクリーンネームの規則性についてそれぞれ誤検知しており、今後これらの傾向を分析し検知手法をより精度の高いものにする必要がある。

LookUpper の評価について全体的に高い評価を得た一方でさらに改良の余地のある部分や追加が望まれる機能が多数存在することが判明し、今後 LookUpper を改良する上でよい指標を得た。

#### 参考文献

- 1) Twitter を利用している人 | Twitter for Business, <https://business.twitter.com/ja/whos-twitter>
- 2) Twitter スпам大流行、手口と防御法～他人のアカウント悪用、著名人装い詐欺も - ライブドアニュース, <http://news.livedoor.com/article/detail/8540033/>
- 3) Twitter : Twitter Blog: Measuring Tweets, <http://blog.twitter.com/2010/02/measuring-tweets.html>
- 4) Twitter : Twitter Blog: State of Twitter Spam, <http://blog.twitter.com/2010/03/state-of-twitter-spam.html>
- 5) 政治家の Twitter なりすまし 最多は橋下徹大阪市長 | 地震予測検証・地震予知情報 / 防災情報【ハザードラボ】, [view-source:http://www.hazardlab.jp/know/topics/detail/1/9/1965.html](http://www.hazardlab.jp/know/topics/detail/1/9/1965.html)
- 6) 認証済みアカウントについて | Twitter Blogs, <https://blog.twitter.com/ja/2014/ren-zheng-ji-miakauntonituite>
- 7) 長谷巧, 山本匠, 原正憲, 山田明, 西垣正勝: アフィリエイトに着目したスパムブログ評価方式に関する検討, 情報処理学会研究報告コンピュータセキュリティ(CSEC), Vol.2009, No. 20(2009-CSEC-44), pp.97-102(2009).
- 8) 佐藤有記, 宇津呂武仁, 福原知宏, 河田容英, 村上嘉陽, 中川裕志, 神門典子: キーワードの時系列特性を利用したスパムブログの収集・類型化・データセット作成, DEWS2008 論文集(2008).
- 9) 石川尚季, 西村涼, 渡辺晴彦, 村田真樹, 岡田至弘: コミュニケーションサイトに投稿されたメッセージに対する著者の推定, 電子情報通信学会技術研究報告 NLC, Vol.109, pp.79-84(2009).
- 10) 榊剛史, 松尾豊: ソーシャルブックマークとしての Twitter リスト機能の応用, The 24th Annual Conference of the Japanese Society for Artificial Intelligence, Vol.2010, No.3B3-2, pp.1-3(2010),

- 11) 吉田光男, 乾孝司, 山本幹雄: リンクを含むつぶやきに着目した Twitter の分析, DEIM Forum 2010, 第 2 回データ工学と情報マネジメントに関するフォーラム, Vol.2010, No.5A-1, pp.1-3(2010),  
(<http://hdl.handle.net/2241/111107>)
- 12) Java, A., Song, X., Finin, T., and Tseng, B. : Why We Twitter: Understanding Microblogging Usage and Communities, In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. *ACM*, pp.56-65(2007).
- 13) Pak, A., and Paroubek, P. : Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives, In Proceedings of the 5th International Workshop on Semantic Evaluation, *Association for Computational Linguistics*, pp.436-439(2010).
- 14) Almaatouq, A., Alabdulkareem, A., Nouh, M., Shmueli, E., Alsaleh, M., Singh, V., Alarifi, A., Alfaris, A., Pentland, A. : Twitter: who gets caught? observed trends in social micro-blogging spam, in Proceedings of the 2014 ACM conference on Web science, 2014, pp. 33-41(2014).
- 15) 中村悠一, 山田剛一, 絹川博之: D-011 Twitter におけるスパムユーザフィルタの開発とその評価 (マイクロブログ, D 分野: データベース), 情報科学技術フォーラム講演論文集, Vol.11, No. 2, pp.99-100(2012).
- 16) 折田明子: ソーシャルメディアにおけるなりすまし問題に関する考察, 情報処理学会研究報告電子化知的財産・社会基盤(EIP), Vol.2009-EIP-44, No.4, pp.1-6(2009)
- 17) 松坂志: Twitter なりすまし問題と対策, 独立行政法人情報処理推進機構(IPA)
- 18) Twitter : Twitter ルール,  
<https://support.twitter.com/articles/253501-twitter>
- 19) 株式会社エスミ,  
<http://www.esumi.co.jp/>
- 20) 「政治家・議員」の Twitter アカウントまとめ一覧,  
<http://meyou.jp/group/category/politician/>
- 21) Twitter と政治(α) / ぼりったー(politter),  
<http://politter.com/>
- 22) 総務省 | (1) インターネット等を利用する方法による選挙運動の解禁,  
[http://www.soumu.go.jp/senkyo/senkyo\\_s/naruhodo/naruhodo10\\_2.html](http://www.soumu.go.jp/senkyo/senkyo_s/naruhodo/naruhodo10_2.html)
- 23) 上野亮, 飯島泰裕: 自治体公式 Twitter の利用実態及び発信情報に関する考察, 社会情報学会 (SSI) 学会大会研究発表論文集, Vol. 2012, pp. 253-256(2012).
- 24) 国会議員の 1 日って?,  
<http://www.nakane.jp/sense/2013082009391413.html>
- 25) 政治家の 1 日 | 政治家の仕事, なるには, 給料, 資格 | 職業情報サイト Career Garden,  
<http://careergarden.jp/seijika/ichinichi/>
- 26) 今日の山田太郎 ?国会議員の 1 日? | 参議院議員 山田太郎 公式 web サイト みんなの党,  
<http://taroyamada.jp/?p=1579>
- 27) 『I-CAREER』「新人若手議員の 1 日を追う」,  
[http://www.yoshidataisei.com/houdou\\_folder/n13/i\\_career.html](http://www.yoshidataisei.com/houdou_folder/n13/i_career.html)