SNS の投稿内容に含まれる興味に着目したアカウント到達可能性算出モデルの検討

吉國 綺乃[†,a]　渡辺 知恵美[††,b]　小林 一郎[†,c]

† お茶の水女子大学大学院人間文化創成科学研究科　†† 筑波大学システム情報系情報工学域

a) *yoshikuni.ayano@is.ocha.ac.jp*　b) *chiemi@cs.tsukuba.ac.jp*　c) *koba@is.ocha.ac.jp*

概要　近年，ソーシャルネットワーキングサービス（SNS）が普及し利用者も増加し続けている．それにともない，SNS の利用において自身が投稿した内容やプロファイルがどの程度のプライバシリスクになっているかを把握する必要性が増している．我々は SNS におけるプライバシリスクの提示指標として，アカウント到達可能性を定義している．アカウント到達可能性とは，攻撃者が利用者の既知のアカウントから別のアカウントを見つけ出す可能性を表す指標であり，アカウント到達可能性を求める手法はさまざま考えられる．本論文ではアカウント到達可能性を求める手法のひとつとして，利用者の ” 興味 ” に着目した算出モデルを考える．具体的には，SNS の友人の投稿内容を元に利用者の興味を推定し，アカウント到達可能性を算出する手法を考える．

キーワード　ソーシャルネットワーキングサービス，プライバシリスク

1　はじめに

近年ソーシャルネットワーキングサービス (以下 SNS とする) の普及により，世界中で利用者が増えている．主流である Facebook[1] のアクティブユーザは 12 億人を超え，日本国内だけでも 2013 年 9 月現在で 2100 万人を超えている [**?**] ．また主流 SNS の一つである Twitter[2] のアクティブユーザは月間 2 億 4100 万人を超えている [3] ．利用者は友人や知人，同じ趣味を持つ他者とのコミュニケーション，またオンラインのゲームなどをするために SNS を利用している．SNS ではプロフィールの公開，日記やショートメッセージの投稿，またチャットのやり取りなど，さまざまな方法でコミュニケーションをとることができる．しかしながら一方では SNS でのトラブルが原因となり，利用者の個人情報が取得されるという事例が発生している．その事例のひとつに炎上事件が例に挙げられる．きっかけはさまざまであり，いつ，だれが，どのように被害にあうかはわからない．SNS を利用している人は誰でも被害者になり得るのである．サイバーストーカーと呼ばれる攻撃者は，利用者が利用している SNS の情報をもとに個人情報を多く取得しようとする．これらの個人情報は自身が気が付かないうちに，投稿内容やプロファイルに自身で公開していることが多い．サイバーストーカーは利用者が公開している情報の組み合わせで，多くの情報を取得していく．

SNS の利用において個人情報が取得されることを防ぐためにも，自身が投稿した内容やプロファイルがどの程度のプライバシリスクになっているかを把握する必要性が増している．我々は SNS におけるプライバシリスクの提示指標として，アカウント到達可能性 [1] を定義している．本論文ではアカウント到達可能性を求める具体的な手法のひとつとして，利用者の興味に着目したアカウント到達可能性算出モデルを検討する．

本論文は，2 節でアカウント到達可能性の定義とモデルの検討を行い， 3 節では興味に着目したアカウント到達可能性算出モデルの検討を行っている．4 節は検証，5 節はまとめとなっている．

2　アカウント到達可能性算出モデルの検討

2.1　アカウント到達可能性

アカウント到達可能性 (Account Reachability) とは攻撃者が利用者の既知のアカウントから別のアカウントを見つけ出す可能性を表す．たとえば，ある利用者が二つの異なる SNS のアカウント s_1 ， s_2 をそれぞれ持っているとする．また攻撃者は利用者の SNS アカウントのうち， s_1 のみしか知らないとする．攻撃者は s_1 の情報をもとにして，まだ知らないアカウントである s_2 をさまざまな手法を通して見つけ出そうとする．ここで攻撃者は s_1 のプロファイルや投稿内容から s_1 のキーワードを抽出し，検索エンジンなどを用いて検索を行い，s_2 になりうるアカウントの候補を取得する手法を用いたとする．このとき，取得した候補アカウントそれぞれと s_1 から取得したキーワードをもとに s_1 との類似度をはかり， s_2 が s_1 のアカウントであると特定していく．この可能性がアカウント到達可能性である．

[1] http://www.facebook.com
[2] https://twitter.com/
[3] http://news.mynavi.jp/news/2014/05/16/324/

アカウント到達可能性

アカウント s_1 から別のアカウント s_2 を見つけ出す可能性は以下に表される．

$$AR(s_1 \rightarrow s_2) = max_{q \subseteq Q}\left(AR(s_1, s_2, q)\right)$$

$$Q = GenQueries(s_1.prof, s_1.msg).$$

$$AR(s_1 \rightarrow s_2, q) =$$

$$Match(s_2, Cand(q)) * \frac{Score(s_1, s_2)}{\sum_{c \in Cand(q)} Score(s_1, c)}$$

$$Match(s_2) = \begin{cases} 1\ if\ s_2 \in Cand(q) \\ 0\ else \end{cases}.$$

ここで $GenQueries\ (s_1.prof, s_1.msg)$ は s_1 のプロファイルや投稿内容から，s_2 のアカウントを見つけ出すためのクエリを生成する式である．Q は生成されたクエリの集合であり，$Cand(q)$ はクエリ q で得られた s_1 の別アカウントの候補アカウントの集合である．$Score(s_1, c)$ は s_1 と候補アカウント c との類似度を表す．

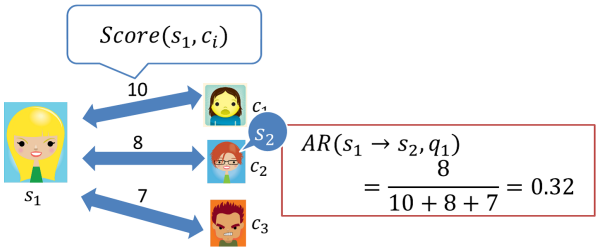

図 1 アカウント到達可能性算出イメージ

例をもとに，実際にアカウント到達可能性を求めてみる．攻撃者は s_1 のプロファイル情報をもとに，$Q = q_1, q_2$ のクエリを生成する．これらのクエリはそれぞれ，{"$keywords$"を"検索エンジン"を用いて検索する } といったクエリである．クエリ q_1 から c_1 ， c_2 ， c_3 の候補アカウントが取得でき，それぞれ s_1 との類似度を 10，8，7 とする． c_2 が s_2 であるとすると，$AR(s_1 \rightarrow s_2, q_1)$ は $8/(10+8+7) = 0.32$ となる．同様にクエリ q_2 から c_1 … c_6 の候補アカウントが取得できたとする．それぞれ s_1 との類似度が 30，2，2，1，1，0 であり， c_1 が s_2 であるとき， $AR(s_1 \rightarrow s_2, q_2)$ は $30/(30+2+2+1+1+0) = 0.83$ となる．このように，それぞれのクエリで $AR(s_1 \rightarrow s_2, q)$ を求め，最終的に $AR(s_1 \rightarrow s_2)$ を求める．この例の場合，アカウント到達可能性は 2 つの値のうち大きい値である 0.83 となる．

アカウント到達可能性を求めるために用いられるこれらの関数は，攻撃者が入手できるデータや利用できる技術に基づいて実装される．先行研究 [1] では技術知識を持たない攻撃者でもできる最もシンプルな方法として，$GenQuery(s_1.prof, s_1.msg)$ では Profile に含まれる名前，所属をキーワードに検索エンジンで検索する実装を，$Score(s_1, c)$ では検索エンジンのランキングの逆数をとる類似度計算を採用した．

2.2 アカウント到達可能性算出モデルの検討

先行研究ではプロファイルデータのみを用いたもっともシンプルな攻撃モデルを採用したが，データ分析技術を用いることで，利用者の投稿履歴からも特徴的なキーワードを抽出し攻撃に利用することができる．$GenQueries(s_1.prof, s_1.msg)$ と $Score(s_1, c)$ を求める代表的な手法を以下の表に示す．

$GenQueries(s_1.prof, s_1.msg)$	$Score(s_1, c)$
プロファイルから生成	検索エンジンでのランキング
投稿内容から居住地を推測	検索エンジンでのランキング
投稿内容から特徴語を抽出	著者推定を利用し類似度を算出

表 1 $GenQueries(s_1.prof, s_1.msg)$ と $Score(s_1, c)$ の代表例

表に示す手法以外にも，さまざまな手法が考えられる．たとえば，友人関係を利用した方法も考えられる．文献 [2] では，異なるソーシャルグラフ間で同一人物であるノードを推定する手法を提案している．同一人物のノードであるとわかっている 2 つのソーシャルグラフの構造をトレーニングデータとして学習を行い，未知のノード間の類似性の推定を行っていく．

先行研究や表 1 に示している手法において $Score$ の値は検索エンジンでの検索順位をもとに求めているが，より精度の高い類似度を測るためには候補アカウントの属性を抽出し，対象アカウントの属性との類似度を測ることが考えられる．代表 SNS である Twitter や Flickr のプロファイル記入欄は自由記述形式であり，機械的に利用者のプロファイル情報を抽出することは困難である．文献 [3] [4] では，自由記述形式のプロファイル欄や投稿内容，また友人関係から利用者の属性を抽出する技術を提案している．

また著者推定を用いて類似度を測る手法も提案されている [5] ．この手法は利用者が投稿した文章を元に，利用者の「癖」をテキストマイニングの技術を用いて見つけ類似度を測る手法である．

算出モデルにこれらの技術を実装することで，利用者

が攻撃される可能性のあるさまざまな攻撃をシミュレートすることができるようになる．

3 興味に着目したアカウント到達可能性算出モデルの検討

アカウント到達可能性は利用者のプロファイル情報が明示である SNS アカウントからプロファイル情報があまり明示されていない別の SNS アカウントを見つける可能性である．プロファイル情報があまり明示されていない SNS からプロファイル情報を推定することでアカウント到達可能性を求めることができる．本節では対象とする SNS を Twitter とし，情報が明示でない SNS アカウントから情報を抽出する手法を検討していく．

SNS において友人関係，フォロー関係は重要な利用者のプロファイル情報のひとつである．そこで以下の 2 つの仮説を考える．

仮説 1

Twitter は自身で自由にユーザをフォローできる SNS であるため，「興味のあるユーザをフォローする」ことが多いと考えられる．ここで「興味のあるユーザ」とは，自身の趣味や好きなこと，興味に関して投稿しているユーザとする．

仮説 2

仮説 1 が成り立つのならば，自身の投稿にはあまり興味のあることについて投稿していなくてもフォローしているユーザの投稿やプロファイルから利用者の興味を推定することができるのではないかと考えられる．

たとえば,「データベース」に興味のある利用者はデータベースに関連する投稿を多くするユーザをフォローすることで，自身のタイムラインを見るだけで自分の欲しい情報を得ることができる．このように興味のあることに関して多く投稿しているユーザをフォローすることでより多くの情報を得ることができるため，多くの利用者は興味のあるユーザをフォローしていると考えられる．

仮説を元に利用者の興味は以下の式であらわすことができる．

$$interest(u) = \{(w_0, freq_0), ..., (w_l, freq_l)\}$$

$$w = \{\text{ 興味のあることを表す名詞 }\}$$

$$freq = \{\text{ 興味の度合い }\}$$

ここで w や $freq$ を求める手法はさまざまある．文献 [6] や文献 [7] では自身の Tweets を元に興味を推定している．本論文では自身の Tweets から抽出するのではなく，フォローしているアカウントの Tweets を元に興味を抽出する．

利用者がフォローしているアカウント $follow(u) = \{f_0, ..., f_n\}$ それぞれに対して Tweets を取得する．全フォローユーザの Tweets から名詞を抽出し w とし，$freq$ を出現頻度とする．

仮説について検証を行っていく．

4 検証

フォローユーザの Tweets から利用者の興味が推定できるか検証を行う．フォローユーザの Tweets から抽出した興味と利用者の Twitter のプロファイル欄に記載されている情報がどれだけ一致するか検証を行う．

4.1 設定

被験者は Twitter アカウントを持つ 5 名である 2.

アカウント	フォローユーザ数
$u1$	90
$u2$	90
$u3$	147
$u4$	409
$u5$	348

表 2 各被験者のフォローユーザ数

このうち $u5$ はリストを作成しており，興味のあるユーザをまとめていると考えられるので興味の抽出にはリストに含まれるユーザを利用した．

それぞれフォローしているアカウントに対して 1000 tweets 取得し，形態素解析器 MeCab[4] を用いて形態素解析を行い，一般名詞と固有名詞のみ抽出した．得られた単語のうち出現頻度の高い上位 5 件を推定した興味とする．また各利用者のプロファイル欄から同様に一般名詞と固有名詞を抽出し，これを正解データとする．推定した興味と正解データから適合率を求める．

4.2 結果

アカウント	適合率
$u1$	0.4
$u2$	0.2
$u3$	0.2
$u4$	0.0
$u5$	0.0

表 3 結果：適合率

結果を表 3 に示す．適合率はどの被験者も低く，あま

[4] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

りいい結果とは言えない．各被験者の推定された興味を見てみると，プロファイルに書いている興味と関連の高い興味が取得されている．

そこで推定された興味と正解データの関連度を求める．関連度の算出には wordnet[5] を用いた．各正解データに対して推定された興味との関連度をすべての組合せで求め，最も高い値をその語の関連度とする．$u2$ の結果を 4 に示す．

4.3 結果

正解データ	関連度
i_1	1.0
i_2	0.167
i_3	0.167
i_4	0.125
i_5	0.125
i_6	0.125
i_7	0.111
i_8	0.111
i_9	0.111
i_10	0.091
i_11	0.083
i_12	0.0

表 4 結果：$u2$ の関連度

これを元に正解データと推定した興味の関連度を各関連度の平均値として求める．結果を表 5 に示す．

アカウント	関連度
$u1$	0.195
$u2$	0.185
$u3$	0.141
$u4$	0.055
$u5$	0.044

表 5 結果：関連度（平均値）

アカウント	関連度
$u1$	0.094
$u2$	0.115
$u3$	0.036
$u4$	0.044
$u5$	0.056

表 6 結果：関連度（入れ替え）

結果をみると関連度はそこまで高いとは言えない．ここで推定した興味を他の被験者の推定した興味と入れ替えて有意差が見られるか検証を行う．

結果を比べてみると一部被験者を除いて，自身のフォローユーザから抽出した興味の方が正解データと関連度が高いことがわかる．あまり有意差が見られないが，異なるユーザから推定された興味と差があることを用いることでアカウント到達可能性を求めることができると考えられる．

詳しく結果をみると関連度がきちんと算出されていない語の組合せが存在した．これは wordnet を利用する際，一度英語の wordnet を介して関連度を計算しているためであると考えられる．英語の wordnet にはない語が含まれているためその語に関しては関連度が求められない．また Twitter のような SNS では略語や独特の表現などがあるため関連度がきちんと算出されない問題がある．

今後はこれを解決するため wikipedia を用いた関連度算出手法 [8] を用いたいと考える．

5 まとめと今後の課題

アカウント到達可能性を求める新たな手法として，友人の投稿内容から推定した興味を用いる手法を検討した．本論文では検討した手法が有効であるか検証するため仮説を立て，仮説について検証を行った．

大きな有意差は見られないものの，自身がフォローしているアカウントから推定した興味と異なるアカウントから推定した興味は，自身の興味との関連度に差があることがわかった．今後はより精度の高い推定結果を得るため，新たな抽出の手法や，より精度の高い関連度の算出手法の検討を行い，アカウント到達可能性算出モデルを構築していく．

[5]http://nlpwww.nict.go.jp/wn-ja/

参考文献

[1] YOSHIKUNI, A., WATANABE, C. : Account Reachability : A Measure of Privacy Risk for Exposure of a User's Multiple SNS Accounts, Proceedings of the 15th International Conference on Information Integration and Web-based Applications & Services (iiWAS2013).

[2] Narayanan, A., Shmatikov, V. : De-anonymizing Social Networks, Proceedings of the 2009 30th IEEE Symposium on Security and Privacy.

[3] Rao, D., Yarowsky, D. and Shreevats, A., et.al : Classifying latent user attributes in twitter, SMUC '10.

[4] Mislove, A., Viswanath, B., and Gummadi K. P., et.al : You are who you know: inferring user profiles in online social networks, WSDM '10.

[5] Stamatatos, E. : A survey of modern authorship attribution methods, J. Am. Soc. Inf. Sci. Technol., 60(3):538-556, mar 2009.

[6] 大原啓詳，灘本明代 : Twitter 上のあるユーザの意外な情報抽出手法の提案，DEIM Forum 2014.

[7] 渡邉恵太，加藤昇平 : Twitter における語の関連性に着目したユーザ興味語抽出手法の提案，The 26th Annual Conference of the Japanese Society for Artificial Intelligence, 2012.

[8] 伊藤雅弘, 中山浩太郎, 原 隆浩, 西尾章治郎: Wikipedia のリンク共起性解析によるシソーラス辞書構築, 情報処理学会論文誌:データベース (TOD), Vol. 48, No. SIG 20 (TOD 36), pp. 39?49 (Dec. 2007).