マイクロブログ上の匿名ユーザの所属推定

内金 亮太郎† 井上 潮‡

†‡東京電機大学大学院 工学研究科情報通信工学専攻 〒120-8551 東京都足立区千住旭町 5

E-mail: †12kmc06@ms.dendai.ac.jp, ‡inoue@c.dendai.ac.jp

あらまし マイクロブログで有名な Twitter は匿名で利用されていることが多い. 利用者の所属が特定できれば, そのグループ全体の興味や意見を抽出可能になる. 本研究では, 特定の組織に属する利用者を抽出する方法として, 多数のアカウントをクラスタリング化する手法を提案する. Twitter は学生の利用率が高いので同じ学校の知り合い同士で利用していることが多いと仮定する. プロフィール等の情報から学校が判明しているアカウントを収集し, そのアカウント同士のつながりから, どのようなグループが構成されているかを分析し, 学校毎にクラスタリング化することで, そのアカウントとつながりの深いアカウントを同じ学校の学生として推定する.

キーワード Twitter, クラスタリング, プロファイリング

1. はじめに

SNS を分析する事が, 注目を浴びている. SNS を分析することで自社の製品やサービス等がどのような評価を得ているのか, どのくらい評判になっているのか, 現在の流行等を測る事ができるとされている. そのためのサービスも増えてきており, 例えば, 評 Ban[1]というサービスは SNS を分析し, その評価結果を提供している.

しかしながら, 現在の評価は SNS の全体の総意見としての分析を行なっているのが主流であり, SNS の中でユーザを特定の層(学生や, 地方等の括り)に分けた詳しい分析を行なっているものの数は少ない. それは SNS の特性に依るものが大きい.

SNS を大きく分けると 2 つに分類される. 匿名性が低いものと高いものの 2 つである.前者は Facebook に代表されるサービスの利用時には実名で登録を必要とされる事によって個人の特定が容易である. また, 所属している組織等の本人の情報も付随していることが多く, その人の情報が分かることも多い. それに対して, 後者はユーザの SNS の新規利用の敷居を下げるため, 仮名での登録が可能であり, 本人の所属の情報なども求められない. Twitter に代表される SNS の多くはこちらに分類される.

表 1. SNS の特性の違い

	Twitter	Facebook
日本のユーザ数	2000 万人	1000 万人
発言数	多い	少ない
匿名性	高い	低い

表 1 に示すように日本では匿名性が高い SNS が好まれる傾向にあり, 匿名性が低い SNS より多く利用されている. また, 一人あたりの自分の意見を発信している回数も多い. そのため, 多くの意見を集めやすい匿名性の高い SNS を分析することが多い. しかしながら, 匿名という特性上, ユーザの所属等の本人情報が得られず, 所属組織毎に分けた分析などの詳しい分析をすることが非常に難しくなっている.

本研究では匿名性が高い SNS として多くの人に利用されている Twitter を分析する. Twitter は 20 歳代前半以下の人に多く利用されるため, 多くのユーザが学校に所属している. また, Twitter の用途として身近の知り合いと友人関係になっていると仮定する. この過程に基づいて, 本研究では特定の学校に所属するユーザをアカウントのつながりから推定する手法について検討する.

2. 関連研究

白木ら[2]は、Twitter の「なう」というキーワードに注目し、発言者の状況の推定を行なっている。また、グェンミンヘンら[3]は, 事前に用意した動詞データから, ユーザがどのような行動をとり, どのような活動をしているのかを分析する研究している. これらの研究の特徴はユーザの発言から分析を行なっている点である.

この手法を用いて特定の大学の学生を探す場合を考える. その大学のみ行われている動作がもしあるならばこの手法は有用であるといえる. 例えば「XX 時に○○大学行きのバスに乗っている」という行動から, ○○大学の学生であると, 推定できるならばこの手法は有用である. しかし, 多くの場合, このような情報を知ることは難しく, その大学に精通している人しか使うことはできない. そのため, ある大学を対象として, その大学の学生を見つけようとする場合, その大学に詳しい人の協力が不可欠になり汎用性が高いとはいえない. また, あまり発言をせず, 他の人の発言を見ているだけの人も存在する. そのようなアカウントを抽出することも難しい.

3. 提案手法

3.1 Twitter のモデルと用語

Twitter でよく用いられている用語について以下に説明する（図 1）.

【ツイート】

Twitter 上のユーザの発言のこと.

【タイムライン(TL)】

Twitter に，自分がフォローしている人たちの発言のリストのこと.

【フォロー】

相手のツイートを見るために登録すること.

【フォロワー】

あるユーザをフォローしている人のこと．例えば図 1 のように B, C, D が A をフォローしていると A にとって B, C, D はフォロワーとなる

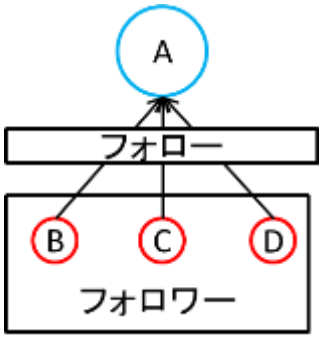

図 1 フォロー，フォロワーの関係図

また，Twitter のアカウントは大きく分けて以下のように 3 つに分類される.

(1) 【個人アカウント】
エンドユーザーが個人のプライベートな事柄を発言するアカウント.

(2) 【公式アカウント】
会社や学校などの組織の代表として広報を目的としたパブリックな事柄を発言するアカウント.

(3) 【Bot アカウント】
世界の名言を発言するなど，特定の目的をもったアカウント．アカウントを操作するプログラムがあり，自動で発言することが多い.

3.2 所属推定方法

学生は所属する学校の公式アカウントをフォローしている可能性が高いので，大学の公式 Twitter アカウントからフォロワーを集め，集まったフォロワーの更にフォロワーを分析することでその大学の学生のアカウントを探す.

分析の流れを図 2 に示す.

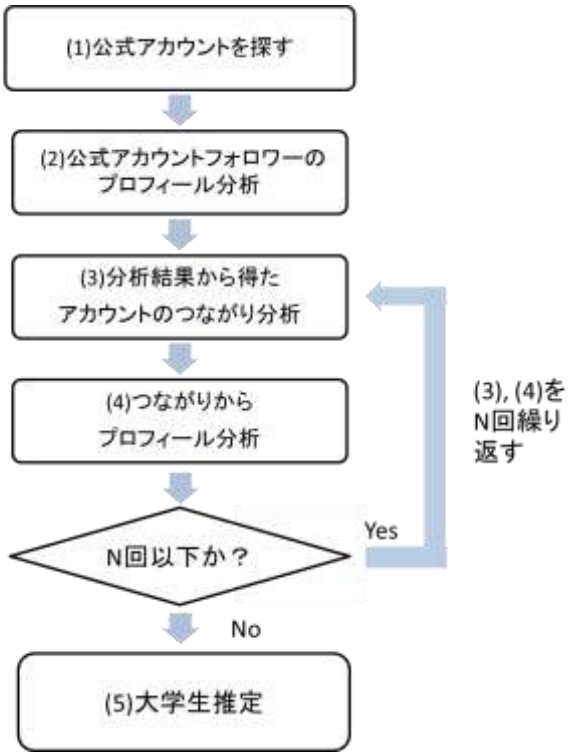

図 2 本システムの流れ

3.2.1 公式アカウント探索

公式アカウントが存在するならば，そのアカウントは個人アカウントより有名であるので，多くの場合 Google や Twitter で大学名をキーワードとして公式アカウントを検索することができる.

3.2.2 公式アカウントのフォロワーの分析

公式アカウントのフォロワーをプロフィールから大学の名称や”学生”といったキーワードを含むかどうかで以下の 3 種類に分類する(図 3).

(1) 【特定の大学の学生】
特定大学の学生であることがほぼ確定であるアカウント．この集合を A と表す.

(2) 【学生候補】
特定大学の学生であるかはわからないが，少なくとも大学生であると思われるアカウント．この集合を B と表す.

(3) 【その他】
(1), (2)以外のアカウント．この集合を C と表す.

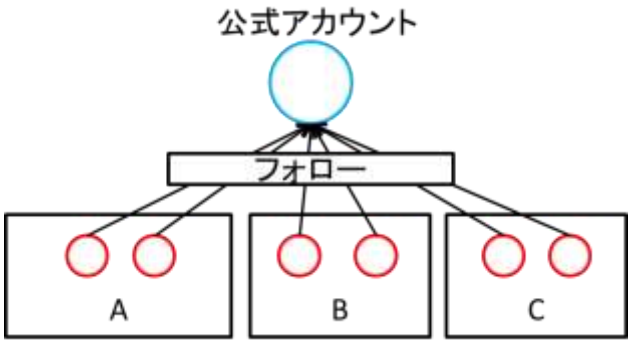

図 3 フォロワーの大学生分類

分類の手順を図 4 に示す．また，これ以降はフォロワー数が極端に高いアカウントは Bot の可能性が高いため排除する．

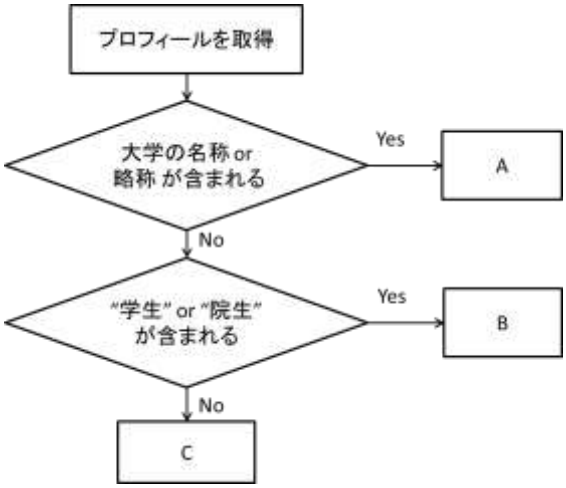

図 4 大学生分類の手順

3.2.3 分析結果から得たアカウントの分析

特定の大学生をフォローする人はその人と同じ大学生である可能性が高いので，特定の大学生及び学生候補のアカウントのフォロワーを分析することで新たに大学生のアカウント見つける．

具体的には特定の大学生と学生候補のアカウントを最低 2 人フォローしているアカウントを探す．また，関係の深いアカウントを探すために最低 3 人, 4 人と増やして，さらに関係の深いアカウントを探す．詳細な手順を以下に示す．

(1)集合 A の分析

集合 A のアカウントをフォローしているアカウントを，図 5 のように，フォローしているアカウント数によって分類する．

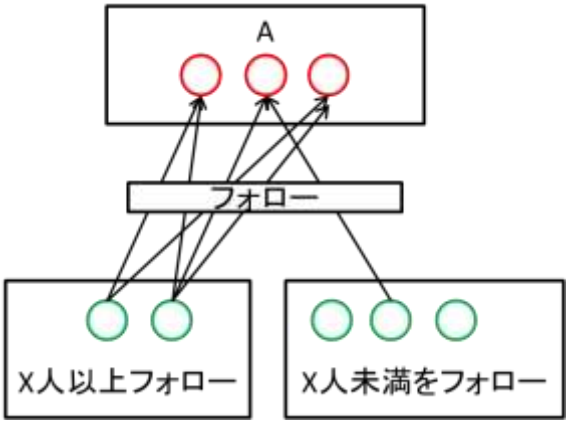

図 5 集合 A のサーチ手順

(2) 集合 B の分析

B 組のアカウントをフォローしているアカウントを探し出す．図 6 のようアカウントを分ける．

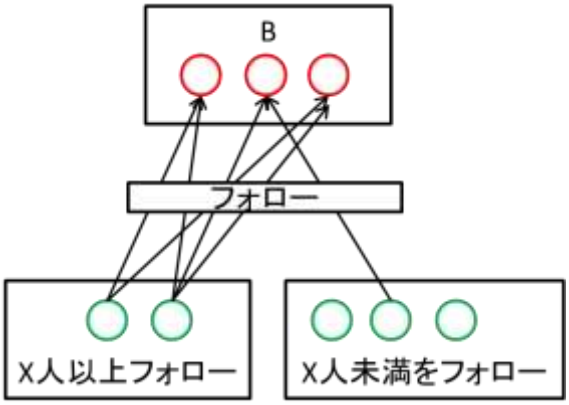

図 6 集合 B のサーチ手順

(3) 集合 A＋B の分析

集合 A, B を区別せずどちらをフォローしてもフォロー数にカウントして分類分けを行う.図 7 のようアカウントを分ける．

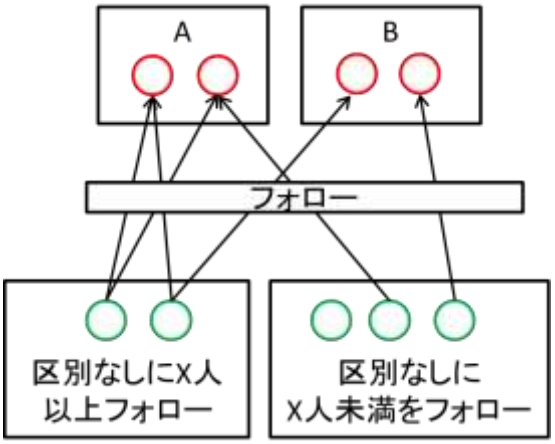

図 7 集合 A+B のサーチ手順

3.2.4 つながりからのプロフィール分析

前述の 3 種類のサーチの手順で得たアカウントのプロフィーから図 4 の手順で分類を行う．最終的に集合 A のアカウントは特定大学の学生と判断する．

3.2.5 大学生推定

B の集合はプロフィールだけでは特定大学の学生とまでは推定ができないが，公式アカウントとフォローの関係が近いので特定大学の学生の可能性が高い。B の集合の内，図 8 の手順で A の集合にフォローされている数を求めることで特定の大学の学生の可能性が高いかどうかを推定する。

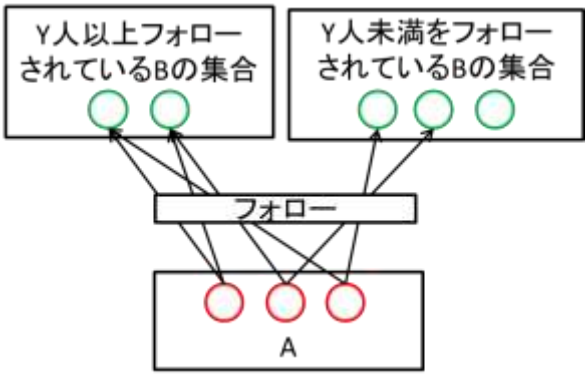

図 8 フォロワー分析

4. 評価

4.1 評価方法

提案手法の有用性を示すためにケーススタディとして東京電機大学の学生のアカウント推定を行い，評価した．

今回は図 9 のシステムを実装し，Twitter のデータを取得は Java のライブラリの Twitter4J を用いた．2012/12/3~2013/1/7 までの期間に取得した合計 104,488 件の Twitter アカウントを分類対象とした．

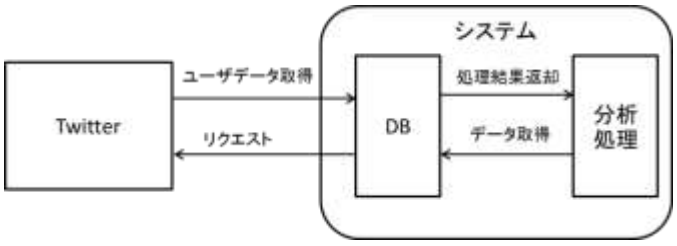

図 9 システムの構成図

4.2 評価結果

東京電機大学の公式アカウントのフォロワー数は 2369 件となり，これらのアカウントを分類した結果は表 2 のようになった．

表 2 東京電機大学公式アカウントのフォロワーの分析結果

分類	アカウント数	比率[%]
A	322	13.5
B	221	9.3
C	1824	77.0

次に公式アカウントのフォロワーを分類した結果を表 3, 4, 5 に示す．なお，フォロー数 X は 2~6 まで変化させた．

表 3 集合 A の分類結果

フォロー数 (X)	アカウント数			
	総数	A	B	C
≧2	889	207	80	607
≧3	536	151	47	338
≧4	335	110	29	196
≧5	251	84	18	149
≧6	188	64	13	111

表 4 集合 B の分類結果

フォロー数 (X)	アカウント数			
	総数	A	B	C
≧2	208	14	24	170
≧3	80	10	3	67
≧4	35	4	1	25
≧5	20	0	4	16
≧6	9	0	1	8

表 5 集合 A＋B の分類結果

フォロー数 (X)	アカウント数			
	総数	A	B	C
≧2	1315	228	115	972
≧3	778	174	73	531
≧4	519	110	29	390
≧5	362	95	30	237
≧6	269	77	19	173

公式アカウントのフォロワーの B の集合である 221 人分のアカウントを大学生推定をフォロワー数を 1~10 まで変化させた結果を表 6 に示す。

表 6 集合 B の大学生推定の結果

フォロワー数	アカウント数
≧1	144
≧2	73
≧3	54
≧4	44
≧5	36
≧6	30
≧7	26
≧8	22
≧9	19
≧10	16

4.3 考察

4.3.1 東京電機大学学生のアカウント数の推定

前節の分析結果の妥当性を検証するための比較データとして，Twitter 利用率と東京電機大学の学生数よりアカウント数を求めた．東京電機大学の学生の Twitter の利用率は一般的な利用率と等しいと仮定すると，表 7 のように，2805 人となった．

表 7 東京電機大学のユーザ数

東京電機大学の学生数[3]	10,668 人
Twitter の一般的な利用率[4]	26.3%
東京電機大学の学生の Twitter の利用数	2805 人

4.3.2 検出した大学生のアカウント数について

表 2 から公式アカウントをフォローしているアカウントの中で 322 人の集合 A のアカウントを取得し，また表 3 から新たに 889 人の集合 A のアカウントを取得することができ，合計 1221 人のアカウントを取得することができた．これは全体の利用者数 2805 人の 44%に相当する．集合 B のアカウントの中にも東京電機大学の学生が相当数いると思われるので，提案手法によってかなりの割合のアカウントを取得できたと考えられる．

4.3.3 大学生のアカウントのフォロワーは同じ大学の大学生か

表 3 と表 4 を比較すると東京電機大学の学生のアカウントをフォローしている人は東京電機大学の学生の学生である可能性が高いことも分かる．このことから判明した学生のアカウントをフォローしているアカウントをさらに探しだすことで連鎖的に同じ大学の学生のアカウントを探し出すことができる．表 5 から学生候補のアカウントを含めた分析を行うと，アカウント数総数は増えるが，東京電機大学の大学生のアカウントが含む割合が下がる(フォロー数 2 以上の場合の時，集合 A の分類結果では全体の 23%が東京電機大学の大学生に対し，集合 A＋B の分類結果は 17%になった)．よって，集合 A＋B で分類する方法は集合 A の分類する方法に比べ，有用性が低いと思われる．また，アカウントをフォローしている数が多いほど，抽出できるユーザは少なくなっている．しかし，東京電機大学の学生の学生をフォローしている数が多いほど，そのアカウントは東京電機大学の学生の学生である確率は高くなっている．

4.3.4 学生候補のアカウント数が少ない理由

同じく表 3 と表 4 を比較すると学生候補のアカウント数が東京電機大学の学生アカウントと比べて数は少ない．これは抽出するキーワードが”学生”と”院生”だけでは学生であるアカウントを探し出すことは難しいためであると思われる．多くのアカウントが大学の名称を入れるか，もしくは学生というキーワードを用いていないかのどちらかであると思われる．

4.3.5 学生候補のアカウントが特定の大学生かどうか

表 6 より 65%以上のアカウントが 1 人以上 A の集合にフォローされていることが分かる．また 2 以上フォローされているアカウントも 33%存在し，複数人にフォローされているアカウントが多いことも分かる。

よって学生候補のアカウントの中に多くの東京電機大学の大学生が含まれていると思われる．

5. おわりに

本稿では大学の公式アカウントを起点としてフォロワーを連鎖的に分析することによりその大学の学生のアカウントを抽出する方法を提案した．提案手法の有効性を確認するため，東京電機大学をモデルケースとして実験した結果，約半数のアカウントを抽出することができた．しかし，課題点が残る．プロフィールから大学生判定を行ったが，判断基準がキーワードが含まれているかどうかだけである点であるため， 本当に大学生かどうかの保証がない．また，略称が他の大学と重複することがあり，必ずしも特定の大学のアカウントを探しだすことはできるとは限らない．他の手法と組み合わせることによって精度を高めていく必要がある．

参 考 文 献

[1] 評 Ban
http://www.hyohban.jp/

[2] 白木敦夫, 矢野幹樹, 酒井佑太, 小澤俊介, 杉木健二, 松原茂樹, 河口信夫”モバイルアプリケーション推薦のための Twitter 発言者の状況の推定”DICOMO シンポジウム, pp251-257, 2010

[3] グェンミンテイ, 川村隆浩, 田原康之”Twitter からの人間行動属性の自動抽出”電子情報通信学会技術研究報告. AI, 人工知能と知識処理 110(105), pp19-23, 2010

[4] 東京電機大学の学生数
http://www.dendai.ac.jp/gakuji/disclosure/201207-2-3-4.pdf

[5] インターネット個人利用動向調査 2012
http://japan.internet.com/wmnews/20120615/7.html