Twitterにおける特定分野に「濃い」アカウントの発見手法

田沼　勇輝† 　鈴木　政巳† 　小林　亜樹†

† 工学院大学工学部　〒163–8677 東京都新宿区西新宿 1–24–2
† 工学院大学大学院研究科　〒163–8677 東京都新宿区西新宿 1–24–2
E-mail: †{c507055,cm09022}@ns.kogakuin.ac.jp, ††aki@cc.kogakuin.ac.jp

あらまし　Twitter は発言が 140 文字と制限されているため容易にツイートを投稿することが可能なので，リアルタイムな情報伝搬，収集や特定の話題について議論するツールとして期待が高まっている．また，Twitter は相互承認機能が無いため，興味のある分野についてツイートするアカウントを気軽にフォローすることができる点でも人気がある．Twitter ユーザが増えるということはそれに比例してツイートの数も増え続けている．多数のツイートの中から有用なアカウントか否かを判断するのは困難であると考えられる．そこで，本研究では特定分野に関する情報をツイートするアカウントを Twitter 機能の 1 つであるリプライに着目し，発見する手法を提案する．

キーワード　マイクロブログ，Twitter，アカウント発見

Finding ‘Otaku’ Account in Twitter

Yuki TANUMA†, Masami SUZUKI†, and Aki KOBAYASHI†

† Faculty of Engineering, Kogakuin University　Nishishinjuku 1–24–2, Shinjuku-ku, Tokyo, 163–8677 Japan
† Graduate School of Engineering, Kogakuin University　Nishishinjuku 1–24–2, Shinjuku-ku, Tokyo, 163–8677 Japan
E-mail: †{c507055,cm09022}@ns.kogakuin.ac.jp, ††aki@cc.kogakuin.ac.jp

Key words　Microblog，Twitter，Finding Account

1. はじめに

近年，140 文字以内で自身の雑記や近況などを書いて投稿できる Web サービス Twitter が普及してきている．日本でもユニークユーザ数が 2010 年 4 月に 988 万人 [1] を超え，さらに増え続けている．Twitter は文字数が制限されている分，容易に書き込むことが可能なので，リアルタイムな情報の伝搬，収集のツールとして期待が高まっている．ユーザの増加に比例してツイートの量も増加している．Twitter では興味のあるアカウントをフォローすることで相手のツイートを見ることができ，有益な情報を得ることができるが，Twitter を始めたばかりのユーザが興味のある分野のツイートを投稿するアカウントを発見する為にはリアルタイム検索で特定分野の用語を検索し，発見したアカウントのツイートを読み判断するか，Twitter 上の機能を使い，おすすめユーザをカテゴリから探し，そのアカウントのフォロー関係から，こちらも実際のツイートを見て判断する必要があり，非常に手間がかかる困難な作業だと考えられる．そこで，本研究では特定分野に関する情報をツイートする可能性のあるアカウントを発見する手法として，Twitter 機能の一つであるリプライに着目した方法を提案する．

2. 提案手法

2.1 概　　要

特定分野に関する情報をツイートするアカウントを発見し，そのアカウントをフォローすることで，ユーザにとって有益な情報をツイートから得られると考えられる．アカウント発見に関して，本研究では 2 つの手法を組み合わせ，特定分野に濃いアカウントを発見する．1 つは分野に関する語を含む割合の高いアカウント（core アカウント）を発見する手法であり，その後，2 つ目の手法として，発見した core アカウントのリプライ関係に着目し，同様に分野に関する語をリプライ内で含む割合を用いる．2 つの手法を統合することで 1 つ目の TwitterAPI の status/User::timeline[3](注1)の形態素解析の結果だけでは発見できない，特定分野に濃いアカウントを発見することができる．特定分野に濃いアカウントの定義を図 1 に示す．2 つの手法を組み合わせることで，図 1 中の塗色部分の特定分野に濃いアカウントの発見ができる．

(注1)：1 アカウントのツイートのみで構成されるタイムラインを取得する API

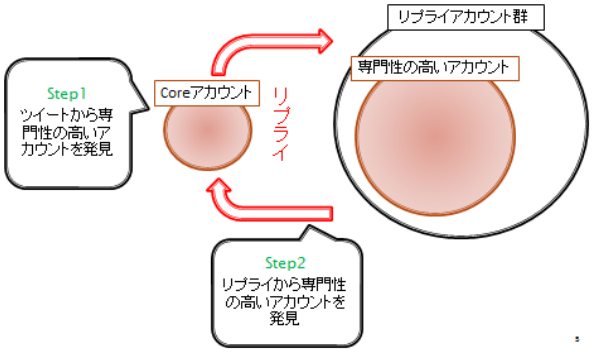

図 1　特定分野に濃いアカウントの定義

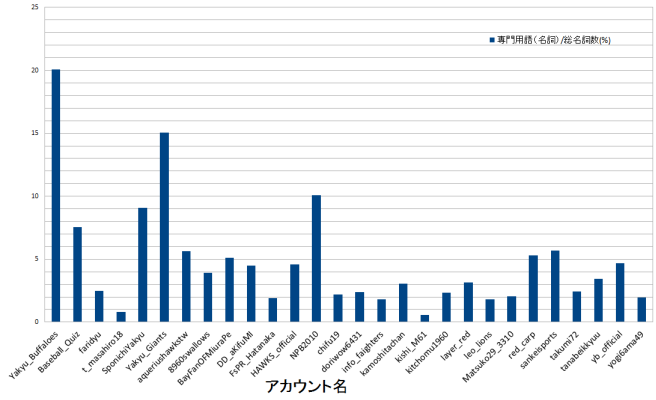

図 2　野球に関する情報をツイートするアカウントの User::timeline の解析結果

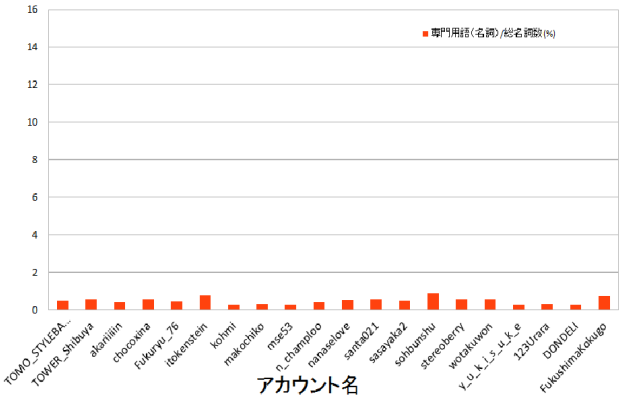

図 3　一般アカウントの User::timeline の解析結果

3.　core となる特定分野に関する情報をツイートするアカウントの発見方法

3.1　事前準備

当該分野の情報をツイートするアカウント（以下 core アカウントとする）を発見するための以下の準備をする．

- 任意の特定分野に関する専門用語辞書（単語集）を作成する．
- core アカウント発見のために，TwitterAPI を利用し，status/Public::timeline(注2) を取得する．

TwitterAPI を利用し，取得した Public::timeline を任意の当該分野専門用語（名詞）集と照らし合わせ，ツイートの内容と作成した専門用語が一致したアカウントを core アカウント候補とし，User::timeline を取得する．取得した core アカウント候補の User::timeline を形態素解析にかけ，ツイートに含まれる当該分野専門用語数（名詞）とツイート全体を構成する名詞数（以下総名詞とする）との割合 (以下専門用語率) を算出することで特定分野に関するツイートを行うアカウントを発見する．

3.2　予備実験 1

core アカウント判断のためのしきい値設定を目的とし，野球分野において予備実験を行う．任意の野球分野に関する専門用語辞書（単語集）230 語を準備し，事前にツイート内容やプロフィール，背景の画像など主観で判断した野球分野に関連するアカウント約 20 アカウントの User::timeline の解析結果と適当に Twitter の先頭画面のリアルタイムに流れているタイムラインの中から適当に選択した一般アカウントの約 20 アカウントの User::timeline の解析結果を比較し，core アカウント判断のためのしきい値を設定する．野球に関する情報をツイートするアカウントと一般アカウントの解析結果をそれぞれ図 2，図 3 に示す．

図 2，図 3 の結果より，しきい値を 1 %と定め，User::timeline 内の専門用語率が 1 %を上回るアカウントを core アカウントとする．しきい値を 1 %と設定することで事前に野球に関するアカウントとして分類した正解集合のうち 90 %の確率で core アカウントを抽出することができる．

(注2)：Twitter ユーザの中でランダムに２０件ツイートを取得する API

4.　リプライを用いたアカウントの発見方法

4.1　リプライの定義

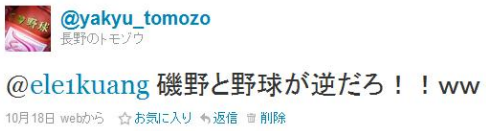

図 4　リプライの例

リプライとは「@アカウント名」を先頭に含むツイートを指し，相手に対する返信の事である．この例を図 4 に示す．本研究では Twitter のリプライのフォーマットで示している，@アカウント名の後に半角スペースを含んでいるツイートだけをリプライツイートとして扱うこととする．

4.2　リプライを用いたアカウントの発見

1 つ目の手法で core アカウントを発見し，次に図 5 のように，core アカウントとリプライを行っているアカウントの中で，リプライツイートの総数が多いターゲットアカウントとのリプライのやり取りに着目をする．リプライを用いてアカウントを発見するための手段を以下に示す．

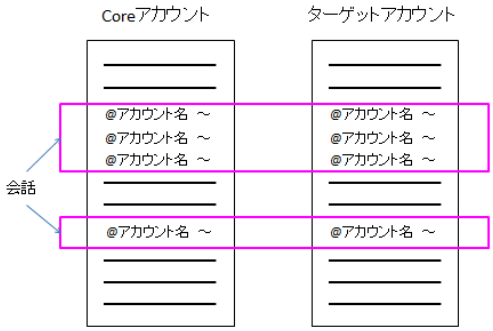

図 5　core アカウントとターゲットアカウント

（1）　リプライを抽出する前にリツイートの処理が必要である．リツイートとは自分のフォロワーに見てもらいたい「他者のツイート」を，改めて投稿することである．このリツイートにも「@アカウント名」が含まれているため，リプライと区別する必要がある．公式 API のリツイートのフォーマットではツイートの先頭に「RT @アカウント名」となっている．そこで User::timeline から先頭が RT で始まるツイートを除去する．

（2）　リプライだけを抽出するために User::timeline からリツイートのみを除去したタイムラインから「@アカウント名」を先頭に含むツイートを先頭マッチで抽出する．core アカウントからターゲットアカウントに対しての全リプライを抽出し，自然言語処理を行い，当該分野の専門用語率を算出する．

（3）（2）と同様に，ターゲットアカウント側からのリプライにも着目し，ターゲットアカウント側でも User::timeline から core アカウントに対しての全リプライを抽出し，自然言語処理を行い，当該分野の専門用語率を算出する．

（4）　core アカウントとターゲットアカウントのそれぞれの、当該分野の専門用語率の結果から、判断に必要な条件を定義する．

（1）から（4）までの流れを図 6 に示す．本稿では図 5 で示したリプライを全体でひと塊りとみなし，core アカウントからターゲットアカウントへのリプライとターゲットアカウントから core アカウントへのリプライの結果を用いて，それぞれの当該分野の専門用語率の値を散布図に表わし，有用なアカウントの発見を行う．

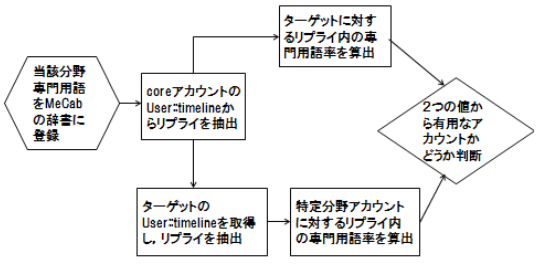

図 6　提案手法の流れイメージ

4.3　予備実験 2

有用なアカウント判断の為の基準を決める事を目的とし，野球分野において予備実験を行った．事前準備として 1 つ目の core アカウントの発見手法に用いた任意の野球専門用語 230 語（単語集）を利用した．なお，本研究では簡易的な用語集で多くのアカウントを発見することをメリットとしているので専門用語集の名詞数を 200 語前後とする．次に図 2 に記されている野球に関する情報をツイートする core アカウント（以下野球アカウントとする）のうちリプライツイートの存在する 19 アカウントの User::timeline を取得し，各リプライを抽出した．各野球アカウントとそれぞれリプライ関係のあるアカウントの定義とし，野球アカウントからの，専門用語の含まれるリプライが 5 ツイート以上のアカウントに限定した．限定した理由としてリプライ数が少なく，ツイートの中身の専門用語数と総名詞数の割合の異常に高いイレギュラーなアカウントを取り除く為である．その結果，野球アカウントのリプライから 194 のターゲットアカウントにたどり着くことが出来た．リプライを自然言語処理するために本稿ではオープンソース形態素解析エンジン MeCab を用いた．MeCab の辞書に，作成した野球専門用語を追加し，core となる野球アカウントからターゲットアカウントに対してのリプライを形態素解析にかけ，算出した専門用語率と，同様にターゲットアカウントから野球アカウントに対してのリプライを解析し，算出した専門用語率を図 7 の散布図に示す．横軸が野球アカウントからターゲットアカウントに対するリプライ内に含まれる専門用語率を表わし，縦軸がターゲットアカウントから野球アカウントに対するリプライに含まれる専門用語率を表わしている．

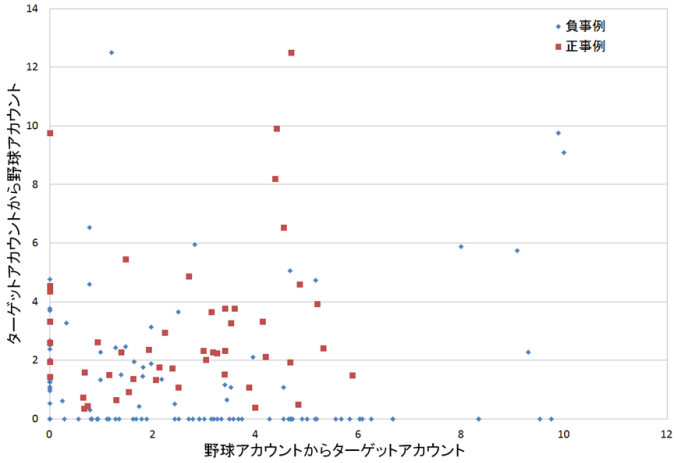

図 7　双方向の専門用語数/総名詞数の解析結果

本研究では非公開アカウントや何らかの理由で存在しないアカウントのリプライの値を 0 とする．正事例として表記されているものはターゲットアカウントの User::timeline のツイートを見て，野球分野に濃いアカウントとして主観で判断したものである．図 7 のデータの内訳は表 1 になる．表中では野球アカウントを B，ターゲットアカウントを O と記している．

表 1 に示されている，共に 0 以外のそれぞれのデータを散布図の 0 に近い点とそうでない点の実際のリプライツイートを見

表 1　図 1 のデータ内訳

	アカウント数
双方向共に 0	51
B から O が 0 ではないかつ O から B が 0	51
O から B が 0 ではないかつ B から O が 0	22
双方向共に 0 ではない	70
総数	194
主観評価による正事例	49

て特徴を述べる．リプライツイートに含まれる専門用語率が共に 0 の場合は無条件で有用なアカウントではないと判断する．

- B から O が 0 ではないかつ O から B が 0

実際のツイートを見たところ，2 %を超える辺りから野球専門用語も増えツイート内で野球の話題についてリプライをしていると判断出来た．しかしターゲットアカウントからのリプライをみると，存在はするが単純な挨拶や特定分野以外の内容が多い傾向がある．本稿ではターゲットアカウントの中から有用なアカウントを発見することを目的とするので，この条件に当てはまるアカウントは有用でないと判断する．

- O から B が 0 ではないかつ B から O が 0

こちらもリプライの内容を見たところ，1 %を超える辺りから野球の専門用語が増えてきている傾向があり，野球の話題についてリプライのやり取りがあると判断できた．ただこちらの場合はリプライ数も少ない為判断が難しく，野球アカウントからのリプライ内にも野球用語が出ている場合もあるが，散布図の結果に反映されていないのは作成した野球分野の専門用語外の野球語が出ているからだと考えられる．この範囲では野球について core となるアカウントへの質問をしているケースが多いとみられる．

- 共に 0 ではない

リプライの内容を見たところ B から O が 1.5 %,O から B が 1 %を超える辺りから野球選手の名前が出てくるほど野球について濃い会話をしていると判断できた. 共に 0 %から大きく離れた位置にあるアカウントは 2 種類に分かれる．1 種類目は野球アカウントの方の野球専門用語にその用語を用いて反応しているアカウントである．2 種類目はリプライ数が少なく，総名詞数も少ないツイート中で専門用語を使っているアカウントである．

4.4　考　　察

野球分野において，野球アカウントとターゲットアカウント間でのリプライのやり取りについて，専門用語辞書（単語集）を作成し，専門用語数と総名詞数の割合を図 7 で示した．正事例として有用なアカウントかどうかの判断基準はツイート内で野球選手名を具体的に挙げ，一般では知らないような情報をつぶやくアカウントとなっている．この結果から，ターゲットアカウントからのリプライに専門用語が含まれていないアカウントは，有用でないと判断できる．野球アカウントとターゲットアカウントの双方に専門用語が多く含まれている結果を示し，正事例に含まれていない範囲にあるターゲットアカウントの特徴として，ツイート内で野球に詳しくないと明言していたり，野球アカウントからの専門用語の含まれるツイートを引用してリプライを行っている傾向がある．この範囲にいあるターゲットアカウントは野球の知識を得たい，または興味がある可能性があり，野球アカウントとのやり取りを重ねることで有用なアカウントになると考えられる．有用なアカウントとそうでないアカウントが混在している範囲にあるターゲットアカウントの違いは全ツイートに対しての野球ツイートの多さの印象が強い．また野球以外のツイートが複数存在することから濃いアカウントではないと判断した．この結果から有用なアカウントを発見する為には，しきい値処理を加えることが有効であると考えられる．

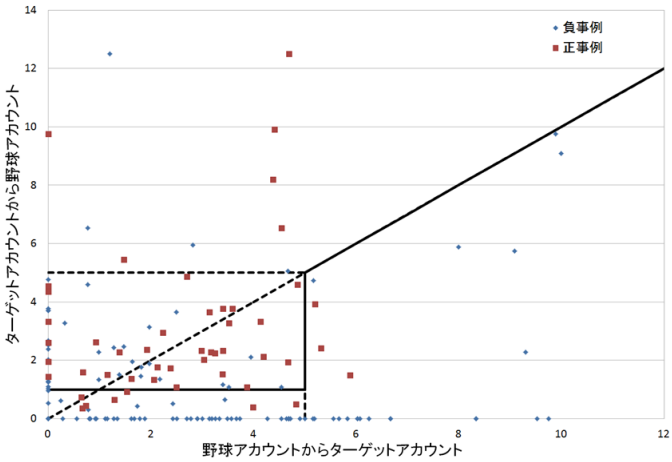

図 8　有用アカウント発見のためのしきい値

4.5　有用なアカウント発見のための判断基準

図 7 より，双方向に 5 %を超える範囲では正事例と負事例の値が明確に別れていることから，有用なアカウントとそうでないアカウントを明確に判断することが出来る．共に 5 %以内の範囲にある値に関しては実際のリプライツイートを見た特徴にも述べたように，1 %を超える辺りから野球の専門用語が増えてきている傾向があり，野球の話題についてリプライのやり取りがあると判断でき，また正事例の値も増えてきている．以上の結果から，双方向に 5 %以内の範囲にある値で，ターゲットアカウントから野球アカウントに対してのリプライに含まれる専門用語数と総名詞数の割合が 1 %以上の判断基準を与えることで有用なアカウントを発見できると考えられる．図 8 の実線より上の部分が，この条件の範囲である．表 2 に図 8 で示した範囲の判断基準で区切った性能指標として，正事例と総数の割合を示す．この判別基準での open data による検証は今後の課題となる．

表 2　性 能 指 標

	総数	正事例	精度
全体	194	49	0.25
フィルタ適用	79	43	0.54

表 2 より，全体の状態と判断基準（フィルタ）適用後を比べると，正事例の値が 49 から 43 とあまり変化が無い中で，精度

をみると 0.25 から 0.54 となり，2 倍以上向上していることがわかる．この結果から設定した判断基準の有効性が証明された．

5. 評　　価

野球分野において提案手法の有効性を検証する．参考値として本研究での 1 つ目の core アカウントの発見手法の予備実験において，主観で判断した野球についてツイートを投稿しているアカウント集合より，一般アカウントとの比較でツイート内に含まれる専門用語率の値が設定したしきい値以上を示したアカウントと 2 つ目のリプライを用いたアカウントの発見手法として，判断基準を用いた手法を組み合わせ，発見したアカウントの結果を表 3 に示す．2 つの手法とも共通して任意の野球専門用語辞書 230（単語集）を利用した．野球専門用語 230 語は野球のルールブックに記載されている語と主観で選んだ．評価実験 2 の部分でも記したが，本研究では，簡易的な語で，より多くの濃いアカウントを発見することをメリットとしている．しかし，野球分野を用い，評価を行うため，野球の用語をルールブックに記載されている語を利用した．

表 3　有　効　性

	提案手法が判別したアカウント数	発見できたアカウント数	精度
2 つの提案手法を組み合わせた結果	105	69	0.65

1 つ目の core アカウントの発見手法において，予備実験の一般アカウントと主観で判断した野球分野に関連するアカウントの User::timeline の比較結果により，しきい値を 1 %と決め，それ以上の専門用語率の値を示したアカウントを core アカウントとした．表 3 の発見できたアカウントの内訳は，1 つ目の core アカウントの発見手法で発見した，User::timeline 内の専門用語率が 1 %のしきい値を超える 26 アカウントと 2 つ目のリプライを用いたアカウントの発見手法に，4.1 のしきい値処理を加え発見した 43 アカウントを足した結果となる．リプライを用いた手法で発見したアカウントでは，しきい値処理を加えた方が有効であると判断したが，図 8 より双方向のリプライツイートに含まれる専門用語率が 5 %以内であり，ターゲットアカウントから野球アカウントに対してのリプライツイートに含まれる専門用語率が 1 %以上の有用なアカウントとそうでないアカウントが混在する範囲を示した値のターゲットアカウントに関しては，リプライツイートだけの判別基準に加え TwitterAPI の User::timeline を用い，全ツイート内に含まれる専門用語率の判断を加えた方が，より有用なアカウントを発見できると考えられる．そのような処理は今後の課題となる．また，本研究ではリプライに着目したアカウントの発見手法に重点を置いたため，1 の Public::timeline を用いた core アカウントの発見を行っていない．それを踏まえた全体の評価実験についても今後の課題となる．

6. おわりに

本稿では Twitte 上で特定分野に関するツイートをする可能性のあるアカウントを Twitter 機能の一つであるリプライに着目し，発見する手法を提案し，野球分野を用いて評価実験を行った. 今後の課題は，SVM など機械学習による自動的な判別の導入である．

文　　献

[1] 篠原 修司，:日本人つぶやきすぎ！ 日本における Twitter のユニークユーザー数、ついに mixi を抜く,http://news.livedoor.com/article/detail/4808127/, (2010).

[2] Twitter API Wiki,http://apiwiki.twitter.com, (2010).

[3] 辻村 浩，TwitterAPI プログラミング, 辻村 浩, 株式会社ワークスコーポレーション，pp.120-129，(2010).

[4] 田中淳史，田島敬史，： twitter のツイートに関する分類手法の提案 ”, DEIM Forum 2010 A5-4, (2010).

[5] 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 A5-1, (2010).

[6] 風間一洋，今田美幸，柏木啓一郎: Twitter の情報伝搬ネットワークの分析 ”, Artifical Intelligence 2010 1F2-OS8-4, (2010).