Twitterにおける興味トピックの変遷に基づくアンフォロー予測

中地　祥剛†　　田島　敬史††

† 京都大学工学部情報学科　〒606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒606–8501 京都府京都市左京区吉田本町
E-mail: †nakaji@dl.kuis.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　Twitterというソーシャルネットワーク上では日々フォロー，アンフォローといったユーザー関係グラフの更新が頻繁に行われている．基本的に一般のユーザーのフォロー，アンフォローについては自由意志で行われるべきであるが，プロモーションアカウントのように，フォローされていることに価値があるアカウントというのも存在する．今回はそのようなアカウントの運用のために，現在の各フォロワーについてアンフォローのされやすさを測るための指標について分析を行った．フォローという行為の予測については先行研究が幾つか存在するが，アンフォローの予測については先行研究は少ない．本論文では，あるユーザーとそのフォロワーについてLDAアルゴリズムを用いたつぶやきのトピック抽出による興味対象の分析を一定期間ごとに行うことによって興味範囲のすれ違いを検出し，それを利用したアンフォロー予測のための指標の提案を行う．

キーワード　マイクロブログ,LDA, トピック抽出

1. はじめに

今日の多くのインターネットユーザーにとっては，ニュースサイトなどのwebサイトからの情報だけでなく，様々なソーシャルネットワークサービス（以下SNS）上で繋がっているユーザーからの情報が重要になっている．

Twitterはその中でも特に人気のあるSNSである．他のSNSと比較した際のTwitterの顕著な特徴として，ユーザー同士の関係を表すグラフが有向グラフである（すなわち，ユーザ間の関係が必ずしも双方向ではない）ことがあげられる．他の有名なSNSとしてはFacebookがあるが，そのユーザーグラフは無向グラフ（すなわち，ユーザ間の関係が必ず双方向）であり，申請相手に承認されない限りユーザーがつながることは出来ない．それとは違い，Twitterは対象の許可がなくとも一方的に関係を構築し，また解除することができる．その為，ユーザーは日々自由に色々な人間を選んでフォロー，アンフォロー（フォローを解除することをそう呼称する）することが出来，結果として色濃くユーザーの指向がフォローユーザーに出ると考えられる．

当然，ユーザーはユーザーをフォローすることもあれば，そのフォローを解除することもある．ここでアンフォローとはフォローと比較してかなりコストの高い行為であると考えられる．何故なら一般にフォローした人間をアンフォローするのは容易だが，アンフォローした人間を再び探し出し，フォローすることは高コストな行動であるからである．そのため，アンフォローにはフォローより明確な理由が存在すると考えられる．

反面，情報を発信する側から見たときには，どれだけの人間からフォローされているか，アンフォローされていないか，ということを考えることは重要である．Twitterが情報発信の場としての価値を獲得しつつある現在では，プロモーションの場としての価値もまた高まってきており，そのようなアカウントにとってはどのようにすればフォロワーを獲得できるのかということや，獲得したフォロワーにアンフォローされないにはどのような振る舞いをすれば良いかということは十分に考慮の対象になる．

そこで，本論文ではそのようなアカウントの運営を支援するための技術として，プロモーションを目的とするアカウントの各フォロワーのうち，アンフォローする可能性が高いものを推定する手法を提案する．

上述のように本論文で提案する手法はプロモーションを目的とするアカウントに適用することを想定しているため，本論文では，アンフォローの中でも特につぶやきの内容について興味があるという目的で成立しているフォローについて扱う．論文[?]では，Twitterにおけるフォロー関係を「個人に対する興味の有無」,「ツイートの内容に対する興味の有無」「相互コミュニケーションの有無」という三つの軸によって分類している．この内，個人に対する興味を持つフォロー関係は，友人同士であるか有名人とそのファンである場合が多い．友人同士の場合は，実世界でも直接交流があることが多く，これらのフォロワーのアンフォローについては，Twitttterを離れた現実世界のつながりや，それに付随する出来事について考慮する必要があるため，Twitter上でのアクションから類推することは難しい．また，芸能人とそのファンの場合も，Twitter上の情報だけからアンフォローを予測することは難しい．そのため，今回は情報発信を行っているユーザと，その「ツイートの内容に対する興味」によりフォローをしているフォロワーとの間のアンフォローについて考える．

本論文で提案する手法ではあるユーザーとそのユーザーがフォローしている全ユーザーの一年間のツイートを一定期間ごとに分割した後，LDA手法を用いてトピック抽出して分析することで，興味トピックのすれ違いにより生じるアンフォローを予測する．手法の概要は以下の通りである．分析対象

となるユーザーを U として，そのユーザーが一年前の時点でフォローしているユーザーを Ufi,$Ufii$……とする．次に U と Ufi,$Ufii$……と分析対象ユーザーの過去一年分のツイートを取得して，およそ三か月ごとの四つのシーズンに分ける，その三か月ごとの全員のツイートを LDA にかけてトピックを抽出し，各ユーザの各シーズンについて，そのユーザが各トピックについてツイートする確率を求める．そして，各シーズンにおけるユーザー U の全トピック上の確率の分布と各 Ufi の全トピック上の確率の分布とを比較することで，各 Ufi がアンフォローしそうかどうかを推定する．

提案手法の性能を評価するため，現在 500 人程度のユーザーをフォローしており，直近一か月内でつぶやきが存在するユーザをランダムに 5 人選び，収集した．これらのユーザーについて，一年前の時点の Twitter の日本語圏ユーザーのフォロー，フォロワー関係を表現したグラフを用い，一年前の時点でフォローしているアカウントをすべて収集，現在のフォロイーと比較し，この一年でアンフォローされたアカウントを収集した．また，その一年前のフォロイーと調査対象のユーザーの一年間のツイートをリツイートを含めすべて収集し, 形態素解析を用いて名詞と動詞のみにした後およそ三ヶ月ずつの四期間に分割しその期間内のつぶやきについて LDA アルゴリズムを用いてトピック抽出を行った．

2. 関 連 研 究

Twitter 上のユーザーのつながりなどについて言及した研究については以下のようなものがある．小出ら [1] はユーザーのフォローが一体どのような目的で行われているのかの特徴をつかむために，他のインターネット上の交流メディアであるブログサイトの読書関係とレビューサイトのお気に入り関係を Twitter のフォローネットワークの構造と比較した．その結果，ブログやレビューサイトでは比較的小規模な高コリンクグループ (コリンクとは，ノードの同士が相互につながっている割合を表す) が得られたのに対して，Twitter のフォローネットワークでは大規模な高コリンクグループと小規模な低コリンクグループが混在していることを示した．低コリンクグループにはフォロワーとはつながろうという意思はない実世界での著名人のアカウントが多く見られ，高コリンクグループは Twitter 上でのユーザー同士の交流においての情報発信や対話などを通じて繋がっていたユーザーのグループであると分析している．

フォロー関係を分類するような関係には田中ら [2] の研究がある．一口に Twitter のユーザーと言っても各々異なる利用目的があり，あるユーザーがユーザーをフォローする際の理由もまた様々であるという考えに基づき，まずユーザーの Twitter の利用目的を大きく情報収集とコミュニケーションの 2 つのタイプに分け，さらにフォローの意図をユーザー指向，内容指向，相互性の 3 つの軸に分類できるのではないかという仮説を立て，分類している．ただし，この研究についてはユーザーの申告をもとに分類を行っているので，自動分類の実現については今後の課題としている．このように，Twitter 上のユーザーグラフの構造自体についても様々な研究がある．これにはやはり Twitter がコミュニケーションツールと情報発信ツールという両方の側面を合わせもつメディアであつことからユーザーの利用目的やグラフのエッジについても従来の web ページのリンクのつながりにはなかった意図が含まれるようになったということだと考えられる．従来の Web ページのリンク関係と Twitter のユーザーグラフ関係の類似性に着目して，ユーザーの価値を計測する研究もある．

続いて本論文の関連研究として，Twitter 上のアンフォローについての先行研究及び，ツイートからユーザーの興味範囲を分析する論文を紹介する．

まず Twitter 上のアンフォローについて言及した論文には以下の様なものがある．Kwak ら [?] は 120 万人の韓国語ユーザーを 51 日間観測し，アンフォローに関連するユーザーの情報として関係の相互性（互いにフォローをしているかどうか），フォローしてからの経過期間，フォローした人間の情報発信性，フォロー，フォロワーの重複性などがフォロー関係を強固さを測る際の指標になることを発見したとした．また同論文で 22 人の韓国人ユーザーに直接インタビューを行い，興味のない話題をつぶやくようになったりそのユーザーの個人的な話題をつぶやくようになった場合にアンフォローをするという報告をしている．この論文はアンフォローについて体系的な研究を行った最初の論文である．この論文のインタビューの結果からも，主にアンフォローはフォロー対象のつぶやきのトピックの比率によって行われることがわかる．

また，Kawk らはよりアンフォローにフォーカスした研究 [?] も行っており，これは先の論文より更にアンフォローに関係する要素について詳しく調べている．ユーザーの構造的な要素，例えばフォローしている人間とフォローされている人間の割合や，自分をフォローしている人間をフォローし返しているかなどや，ユーザーの行動，例えばお気に入り登録やリツイートをしたり，同じハッシュタグについてつぶやいたりしているかということを変数に，どの要素がアンフォローに寄与するかを調べたものになっている．このような定量的に調べられる変数を考慮した上で定性的に思えるような興味範囲の異なりについても本論文の手法を用いて，定量的に扱うことができれば，よりアンフォローの予測やそれ以外の分析についても有用となると考えられる．

加えて，トピック抽出やその周辺技術をを利用した Twitter に対する研究についても紹介する．マイクロブログのトピック同定については実に多くの研究が存在する．奥村 [3] は「マイクロブログマイニングの現在」というタイトルで現在のマイクロブログマイニング研究についての人気のテーマの列挙，その各テーマについて使用されている手法をまとめたサーベイ論文である．その中のトピック判定を扱う章では LDA を用いてユーザーをトピックの確率分布として表現し，KL 情報量やコサイン類似度などを用いて類似度を計算する手法を挙げている．これは本論文で用いたやり方でもある．他にも，ユーザー自身が自らのツイートについて話題を表すタグ付けをすることができる Twitter のハッシュタグ機能を教師情報として利用する Labeled LDA という LDA の拡張手法なども紹介されてい

る．LDA は一連の単語の集まりを一つの文章として扱い，その文章集合にたいして統計的な処理を行うことでトピックを抽出する手法であるが，1tweet を 1 つの文章とするか，1user の tweet 全体を 1 つの文章して扱うかという二つのやり方が存在するとしている．

LDA とよく比較されるテキストマイニングの手法として，tf-idf 法が挙げられる．これは，トピック抽出というよりはある文章群を入力としてその文章群を象徴するような特徴語を抽出するための手法である．この手法を用いて Twitter を分析した論文もある．Hannon ら [?] は Twitter や Facebook のリアルタイム性に着目し，効率的なフォロー推薦のためのユーザープロファイリングのための手法として tf-idf 法を利用している．Hong [4] らは tf-idf 法や LDA 法などの手法などを用いた幾つかの推薦手法が，実際にどれほどフォロイー推薦の役に立つのかということを比較している．現在のマイクロブログの文章に対しての分析法としてはやはり，この tf-idf と LDA の二手法が主流であり，その上で細かなアレンジを加え，より精度を出そうと試みている研究が見られる．

Sriram [5] らもツイートをトピックを用いて分類する研究を発表しているが，Twitter のように短く散文的な文章を分類するのに Bag-of-words というドキュメントに単語が含まれているかのみを考える手法が有効かどうかを検証するものであった．Zhao ら [?] は Twitter が一つの投稿は 140 文字以内という制限があるために非常に短い文章が集まっているという特徴に対応する Twitter に特化した Twitter-LDA 法を提案している．実際，一回の投稿に 140 文字制限があることもまた Twitter の大きな特徴であり，このことを特徴を加味することによってよりトピック抽出の精度を上げる研究もこのようになされている．

LDA を用いたツイート内容のトピック抽出については Marco ら [6] の研究が詳しい．Marco らは機械学習の技術を用いて，ユーザーを「政治的所属」「民族」「特定の事業に対する相性」という全く異なる三つのトピックについてユーザー分類を試みるという研究を発表しており，この中で幾つかの LDA の発展的手法の成果について比較している．

北田ら [7] はコサイン類似度に基づいたトピック同士の類似度を利用して関連するトピック同士を「トピック系列」という単位でまとめ，これを一定期間のツイートに適用した後時間的変遷を可視化することで話題の変化についての理解を助けるという実験を行っている．実際に東日本大震災時につぶやかれた約 3 億 6 千万ツイートのトピック遷移をこの手法を用いて可視化したところ，「避難」「爆発」「自衛隊」などの現実の発生した出来事を表すようなトピックの特徴語が時系列に沿って検出されたことを報告している．

Twitter におけるトピック抽出やユーザー，ツイートの分類の手法はそれ単体でも大きなテーマではあるが．今回はあくまで一つの手法として利用するにとどめ，トピック抽出がユーザーの興味範囲を反映するかの判定に用いることにする．

3. 実験の概要

今回の論文では予測手法を以下のように分類する．まず LDA を用いた興味のずれをアンフォロー値と定義し，それをどのように定義するか，どのように扱うかということを説明する．

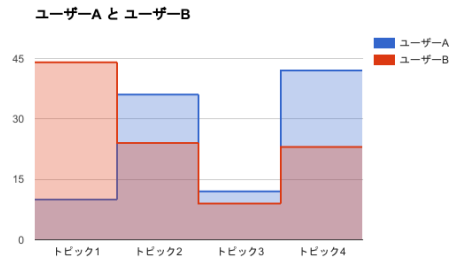

図 1　トピックの全体的な傾向差に着目

- アンフォロー値が時間とともに蓄積されていくモデル
- アンフォロー値のシーズンごとの変化率を考えるタイプ

これはアンフォロー値についての 2 つの考え方を表現したモデルである. 前者はトピックの不一致によるアンフォロー値が時間とともに蓄積されていき, ある閾値を超えたところでアンフォローが起こると捉えているモデルになっている.

それに対して後者はアンフォロー値の変化率がアンフォローの大きな要因となっていると考えるモデルである. ある時期を堺に一気に興味対象の異なるツイートが増えた際にアンフォローを決定するのではないかと考えるモデルになっている.

これらは, アンフォロー値の扱いについてのモデルである. これに加え, アンフォロー値自体をどのように定義するかについて今回以下のような手法を考えた.

3.1 全体の注目トピック傾向を考えるモデル

これは注目しているトピックの全体の傾向が似通っているかどうかがアンフォローを決定していると考えて, それを定量化することによりアンフォロー値とする考え方である (図 1). 具体的には以下のような値を使用した.

3.1.1 同トピックについて順位の差分の絶対値の和

これはたとえば, ある期間について, ユーザー A のつぶやきをトピック数 5 で分類した後, そのトピックを出現確率の降順でソートした配列が [4,1,2,5,3], そのユーザーがフォローしているユーザー B について [1,2,3,4,5] という配列が得られたとき, トピック 1 について $|2-1|=1$, トピック 2 について $|3-2|=1$, トピック 3 について $|5-3|=2$, トピック 4 について $|1-4|=3$, トピック 5 について $|4-5|=1$ となり, それらを合計した 1+1+2+3+1 = 8 という値がスコアになる, という手法である. これはトピックの重みを考えない, 傾向だけを比較する手法であると言える. つまり，ユーザーがあるユーザーのつぶやきを見るときはつぶやきのトピックそのものの比率ではなく，そのユーザーが主にどのようなトピックを話すようになったかという大まかな印象によってアンフォローを決定するのではないか仮設に基づくモデルである．以下手法 A と呼ぶ (図 2).

3.1.2 トピックの分布についての着目ユーザーからそのフォローしているユーザーについてのカルバック・ライブラー情報量

カルバック・ライブラー情報量 (以下 KL 情報量) とは 2 つの

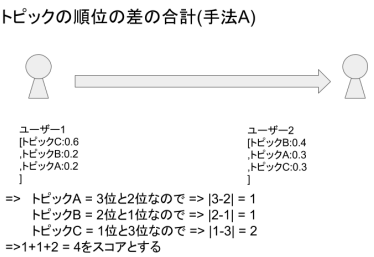

図 2　手　法　A

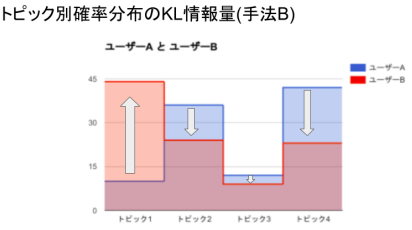

図 3　手　法　B

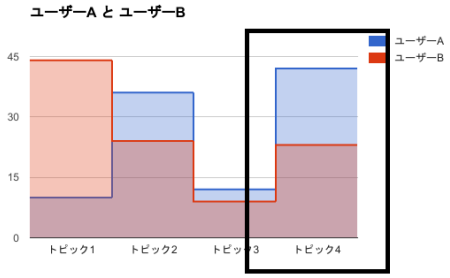

図 4　一部のトピックに着目

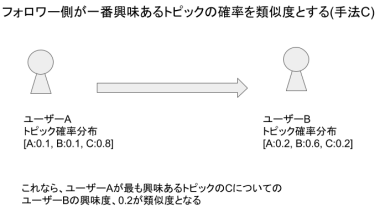

図 5　手　法　C

離散的な確率分布についての差異を表現する値である. 距離に近い概念であるが, 距離の公理を満たしていないので情報量と呼ばれている.

この値は P と Q が離散確率分布である場合,P から Q の KL 情報量は,

$$D_{KL}(P||Q) = \sum_i P(i)log(P(i)/Q(i))$$

として定義されている. これを今回のケースに適用すると, ユーザー A のトピックのスコアを正規化した配列について, トピック n の確率の値を A_n とすると,

各フォロワー $f \in Follows_A$ について,

$$D_{KL}(A||f) = \sum_i A(i)log(A(i)/f(i))$$

と表現できる.

順位の差を計測する前述の手法 A との違いを説明する．手法 A はトピックの「順位」を重視する手法で，各トピックの量について重みは付いていない．一方今挙げた手法はトピックの比率の差を計測している．以下手法 B と呼ぶ (図 3).

3.2　特定のトピックに着目し, そのトピックへの寄与度を考えるモデル

前述のモデルが興味トピック全体の傾向の差を考えるモデルだったとすれば, 以下で挙げる評価方は, 一般的にユーザーは一つのトピックのためにユーザーをフォローしているという考えに基づくモデルである 4.

3.2.1　各トピックの確率分布の値の積の最大値を類似値とする

これはすなわち, フォローしている人間について, 一番共通してよくつぶやいているトピックがそのユーザーをフォローしている理由であるトピックであるという考えに基づいて, そのトピックについてどれだけつぶやいているかということの積を類似度を表すスコアになるというモデルである. たとえば,A というユーザーについて, [トピック 1 が 0.2, トピック 2 が 0.3, トピック 3 が 0.5],A がフォローしている B というユーザーについて [トピック 1 が 0.1, トピック 2 が 0.3, トピック 3 が 0.6] という割合になっていた場合に, それぞれの同トピックの確率の積である 0.2*0.3=0.06,0.3*0.3=0.09,0.5*0.6=0.3 の内, 最大である 0.3 が類似度になるという考え方である. この場合この値は「アンフォローのされにくさ」となるので, アンフォロー値とするには-1 をかけることにする. 以下手法 C と呼ぶ (図 5).

3.2.2　対象ユーザーの最も興味あるトピックへの寄与度を類似度とする

これはある意味最も単純なモデルである. すなわち, 基本的にユーザーは一つのトピックのためについてユーザーをフォローしているという考えに基づいて, 単にそのトピックの確率値の合計をスコアとする方法である. 実際，トピックごとに幾つかのアカウントを使い分け，一つのアカウントの中ではそのアカウントのトピックに関する情報しか発信しないという使用法をしているユーザーも一部いることがわかっている．そのようなユーザーの場合，その対象トピックのつぶやきの比率こそがもっとも重要な情報であり，それ以外のトピックについての情報についてはノイズになりうる可能性もある．以下手法 D と呼ぶ (図 6).

今回の研究では最初のアンフォロー値の扱い方の 2 パターン * アンフォロー値の定義 4 パターンの敬 8 パターンについて結果のスコアを計測した後, その値でソートしたフォローユーザーのリストについて, アンフォロー数を縦軸とした AUC のスコアを用いて評価を行う. 例えば,500 人フォローしている中で一年後に 50 人のユーザーをアンフォローしていた場合, 無作為にアンフォローを予測して並べた場合およそ 10 人に一人一

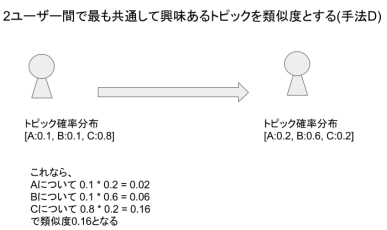

図 6　手　法　D

表 1　アンフォロー値が時間とともに蓄積されていくモデル

ユーザー	フォロー数	アンフォローの数	基準スコア	手法 A	手法 B	手法 C	手法 D
ユーザー 1	134	39	2613	2824	2280	2559	1640
ユーザー 2	117	17	994.5	1180	1227	905	579
ユーザー 3	152	9	684	415	506	1036	680
ユーザー 4	417	3	625.5	500	458	799	363
ユーザー 5	414	5	1035	784	948	826	1272

表 2　アンフォロー値の変化率を見るモデル

ユーザー	フォロー数	アンフォローの数	基準スコア	手法 A	手法 B	手法 C	手法 D
ユーザー 1	120	38	2280	2870	1929	2287	1529
ユーザー 2	101	17	858.5	1189	1126	946	630
ユーザー 3	141	6	423	351	251	600	466
ユーザー 4	377	3	565.5	586	571	786	451
ユーザー 5	391	4	782	786	832	753	867

定の割合でアンフォローが出現すると考え, 基準となるスコアは 50 * 500 /2 = 12500 となる. これよりも予測が優れていた場合, アンフォローが初期に固まるはずなので, スコアは大きくなるはずである.

以上 2 つの計測モデルと 4 つのアンフォロー値定義手法の計 2 * 4 = 8 パターンにおいて検証を行った.

4. 実 験 結 果

以下が実験結果である. まず, 最初のデータとして, 実際につぶやきを計測できたフォローユーザーの数とアンフォローの数を列挙する. これは必ずしも一年前の時点でフォローしているアカウント, 一年間でアンフォローしたアカウントの全てにならないことに注意したい. というのも, 実際にはフォローされつつもこの一年間でまったくつぶやきを行っていなかったり，調査を開始した時点でアカウント情報を非公開にしているアカウントや, 消去されたアカウントについてはツイートを収集できないため, 分析対象に含めることが出来ないからである. 非公開アカウントとのつながりは多くの場合，現実の人間関係の延長であることが多いと考えられるので，今回の予測の対象とした情報源的な価値を見出しているフォローについてのアンフォローに対しては大きく影響が無いと考える．また，変化率を考える手法については，一部連続した期間をまたいでつぶやいていないユーザーについてはデータを計測できなかったので（たとえば一定期間の間つぶやいていなかったりなど）さらに調査対象ユーザーが少なくなっている.

結果として，各ユーザーについて最もスコアが出た手法は以下のようになった.

表 3　各モデルについて最もスコアが出た手法

ユーザー	蓄積モデル	変化率モデル
ユーザー 1	手法 B	手法 B
ユーザー 2	手法 A	手法 B
ユーザー 3	手法 C	手法 C
ユーザー 4	手法 C	手法 C
ユーザー 5	手法 D	手法 D

表 4　最もスコアが出たモデル，手法の組み合わせ

ユーザー	最もスコアが出たモデル	手法	基準スコアとの比率
ユーザー 1	変化率モデル	手法 B	1.25877
ユーザー 2	変化率モデル	手法 B	1.38380
ユーザー 3	蓄積モデル	手法 C	1.514619
ユーザー 4	変化率モデル	手法 C	1.38992
ユーザー 5	蓄積モデル	手法 D	1.22899

5. 考　　察

まず，二つのモデルについては，モデルを変えても最もスコアが出る手法は 4/5 のケースで変わらなかったことから，アンフォロー予測において支配的なのはアンフォロー値をいかに扱うかということよりアンフォロー値の定義ということがわかった．手法間の比較では手法 B と手法 C が同数で最もスコアが出る手法となった．この二つの手法は二つのユーザー間のフォローに関与する興味トピックの差を全体の傾向で捉えるか，共通して最も興味のあるトピックのみで考えるというという手法であった．これは最初の手法の章で述べた通り，ユーザーのアカウントの利用法によってどちらの手法が有効かが変わってくると考えられる．すなわち，ユーザー C と D については単一のトピックのために Twitter を利用しているようなユーザーであり，逆にユーザー A,B については幾つかのトピック数についての銃砲を発信，収集するために Twitter を利用していると考えられる．また，多くの場合において手法 A と B のスコアはそこまで異なるものではないが，手法 C と D のスコアについては手法 C のスコアが大幅に大きくなっていることがわかる．このことから，基本的に単一のトピックによってアンフォロー予測する際は，最も共通して興味を持っているトピックの興味度を見ることが有効だとわかった.

精度が基準スコアと比較して伸びてないケースについても考える．まず，手法 D についてはほとんどのケースにおいて基準スコアを下回っている．つまり多くのユーザーは，自らがもっとも多く発信しているトピックが必ずしも他者に求めているトピックではないということを示している．今回唯一手法 D が最も高いスコアを出したユーザー E については逆に自分が発信しているトピックについて言及が減ったユーザーについて敏感に反応しアンフォローするユーザーであったと考えられる．また同ユーザー E について，手法 C のスコアがあまり高く出ていないことから，このユーザー E が支持しているトピックについてはあまりフォローしている人間は言及していなかったということがいえる．手法 D についてはこのようにマイナートピックについて言及するユーザーについて敏感にフォロー，アンフォ

ローを行うユーザーについてピンポイントで有効な手法だったと考えることができる．

6. 結　　論

本論文では Twitter におけるアンフォロー予測について，2 ユーザー間の興味対象のすれ違いを LDA を用いたトピック抽出をした後，一定期間のつぶやきのトピックの内訳の確率を集計することで表現し，その値を用いてアンフォローが予測できるのではないかという仮設に基づき，その値を用いたアンフォロー値の幾つかの定義および，その集計法について幾つかのモデルを考案し，その組み合わせを一年間の集計したツイートについて実験し，AUC スコアを計測することによって手法の有効性を比較した．その結果，集計法については各期間の変化率を見るモデルが，手法についてはトピックの順位の差分を集計する手法と掛け合わせて最も確率が高くなるトピックの確率を集計する手法が有効だとわかった．

今後の課題について述べる．まず，本研究のデータ数においては十分だったとは言えないので，もっと多くのユーザーについて調査することが課題である．また，今回アンフォローを情報源の意味を持つフォローに対して起こるものと考えて実験したが，勿論それ以外にもアンフォローの要因はいくつか考えうる．他のアンフォロー要因についても推測し，それぞれにおいて評価，予測できる手法を考案し，今回の実験と組み合わせることにより，より精度の高いアンフォロー予測が可能になると考えられる．

また，実験手法についてもいくつか課題は残った．今回の大元の目的はトピック抽出を用いてアンフォローを予測できるかという仮説の検証であったために，トピック抽出に使用する手法については単純に LDA を用いたが，より高度な精度を求めるなら，アンフォロー値の計算方法を今回高いスコアを出した手法 A や手法 C に固定して，トピック抽出の手法を比較するという実験も必要である．トピック数についても議論の余地がある．

また，今回は RT を含むツイートのみをトピック抽出の対象としたが，Twitter には他にも好感を覚えたツイートをブックマークすることができる「お気に入り」機能が存在することを考えて，お気に入りされたツイートについても分析対象に加えるなどの工夫も考えられる．また RT やお気に入りについては，ユーザー間での使用感が分かれるリアクションであると考えられるため，これらの行動についてユーザーごとに適切な重みをつけることができれば，よりユーザーの興味を高精度で抽出できるのではないかと考える．

謝　　辞

本研究は JSPS 科研費 16K12430 の助成を受けたものです

文　　献

[1] 小出明弘, 斉藤和巳, 風間一洋, 鳥海不二夫: ネットワーク分析による Twitter ユーザーのフォロー形成に関する一考察, 情報処理学会論文誌数理モデル化と応用（TOM） pp 164-173 (2013).

[2] 田中淳史, 田島敬史: Twitter のフォロー関係のユーザの意図に基づく分類, *DEIM Forum 2010* F5-1(2010).

[3] 奥村学: マイクロブログマイニングの現在, 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション pp 19-24(2012).

[4] Hong, . and Davison, B. D.(eds.): *Empirical Study of Topic Modeling in Twitter*,*SOMA '10 Proceedings of the First Workshop on Social Media Analytics.* pp 80-88 (2010).

[5] Sriram, B.(ed.): *SHORT TEXT CLASSIFICATION IN TWITTER TO IMPROVE INFORMATION FILTERING.*,*SIGIR '10 Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* pp 841-842 (2010).

[6] Pennacchiotti, M. and Popescu, A.-M.(eds.): *A Machine Learning Approach to Twitter User Classificaion.*,*the Fifth International AAAI Conference on Weblogs and Social Media* pp 281-288 (2011).

[7] 北田剛士, 風間一洋, 榊剛史, 鳥海不二夫, 栗原聡, 篠田孝祐, 野田五十樹, 斉藤和己: Twitter のトピック変遷の可視化法の提案 (2015), *DEIM Forum 2015* E2-6.