トピックモデルを用いたTwitterフォロー情報からのユーザ嗜好の推測手法の提案

WANG Yu† 前田 亮‡

†立命館大学情報理工学研究科 〒525-8577 滋賀県草津市野路東1-1-1
‡立命館大学情報理工学部 〒525-8577 滋賀県草津市野路東1-1-1
E-mail: †gr0320fv@ed.ritsumei.ac.jp, ‡amaeda@is.ritsumei.ac.jp

あらまし 近年、Twitterに関する研究が盛んに行われているが、ほとんどの研究はユーザが投稿した内容やユーザの行為に注目している。しかしながら、ユーザがフォローしているアカウントこそが、ユーザが最も興味を持っている内容である。本研究では、対象のTwitterユーザがフォローしているアカウントから着手して、それぞれのツイート内容をLDA手法を用いてクラスタリングする。そして、対象ユーザが登録した「いいね」やリツイートなどの行為データを使って、クラスタリングの結果との類似度を計算し、重み付けを行う。最後に、重み付けの結果をもとに、重みの大きさがユーザの興味の度合いを示していると推測し、タイムラインを並び替える。

キーワード Twitter , LDA、フォロー情報

1. はじめに

近年、Twitter、Facebook、Instagramなどのソーシャルネットワーキングサービス（SNS）が想像以上のスピードで人々の生活に巨大な影響を与えている[1]。2016年9月までの全世界でのFacebookの月間アクティブユーザは16億5000万人、Twitterは3億2000万人に上る[2]ということである。年齢や職業を問わず、多くの人が日常生活の中で長い時間SNSを利用している。SNSが世界規模でブームになるとともに、次々と膨大なデータが生まれている。サービスの質を向上させ、より良いユーザ体験を提供するため、このような大量のデータをどのように利用するかについて、現在、様々な研究が行われている。

また、2017の2月、Twitterはタイムラインの仕様を変更した。ユーザの行動履歴によって、重要な新着ツイートがタイムラインの最上部に表示されるようになった。それは、全体としては従来の純粋な時系列表示が守られている。今後、ユーザの好みに合わせて内容を表示することが一層重要になるとみられる。

Twitterを使っているユーザは、自分が関心や興味を持っているアカウントをフォローする傾向がある。つまり、誰をフォローしているかという情報から、一人のユーザの嗜好を大体推測できる。さらに、ユーザが単に興味あるアカウントをフォローするだけではなく、そのアカウントが今何をしているかあるいは何の情報があるかを最も知りたいと考えられる。換言すれば、一人のユーザがSNSで最も見ているのは、自分がフォローしているアカウントが何を投稿するかである。だからこそ、それを利用して、何らかの発見が得られるのではないかと考えている。

そこで、本研究ではTwitterユーザのフォロー情報をもとに、Latent Dirichlet Allocation (LDA)[3]トピックモデルを利用して、ユーザの嗜好を発見することを目標とする。

また、LDAモデルを用いた研究は主に、自動的なラベル付けあるいはトピック発見のために行われている。数学の視点から、LDAモデル自体を研究し、改善する研究が多い。しかしながら、発見されたトピックをどのように利用して、ユーザへの推薦やサービスの向上に繋げるかという点に注目した研究は少ない。

本論文は次のように構成される。第2章では、関連研究を紹介する。第3章では、提案手法について詳しく説明する。どのようにLDAモデルを使うか、そしてどのように重付けを行うかについて説明する。第4章では、評価実験を行い、提案した手法を検証する。

2. 関連研究

これまでのSNSについての研究は、主に二つの方向性に分けることができる。一つは、センチメント分析というテキストマイニング技術などを用いてSNS上の書き込みからユーザの感情を分析することである。Bifetら[4]は、Massive Online Analysisというデータストリーム分析フレームワークを用いて、Twitterのユーザの感情をリアルタイムに検出できると述べている。もう一つは、機械学習などの技術を使い、ユーザモデリングでペルソナを作り、ユーザの好みや性格などを推測することである。どちらもユーザにタイムラインの内容やアカウントの推薦を目指す研究である。

しかしながら、タイムラインをSNSにおける不可欠な機能として、その内容の優先度に注目している研究はごくわずかである。Takamuraら[5]は内容の緊急度を基準として、ツイートをいくつかの種類に分け、タイムラインに表示される優先度を検討した。例えば、「雨だ」というような有効時間が短い内容をタイムライン

に優先的に表示する必要があるということであり、「ハワイが綺麗だ」など有効時間があまりない内容は、タイムラインの下部に表示して構わないと考えられる。

斎藤ら[6]はソーシャルブックマーク（SBM）サービスを利用して、Twitterユーザの興味語を抽出するという研究を行った。SBMのタグをもとに構築した類似語を用いて、ユーザの嗜好や属性を抽出し、自身にとって相応しい内容を推薦する。

Zhaoら[7]はトピックモデルを用いて、新聞記事とTwitterからのトピックの抽出結果を比較した。抽出したトピックを三つの種類に分類し、新聞記事とTwitterから得たトピック結果を比較する。その結果、Zhaoらは、ツイート文は短いがLDAモデルも有効であるという結論を得た。

佐々木ら[8]はTwitter上で、オンライン学習可能なトピックモデルを提案した。このオンライントピックモデルは時刻ごとにパープレキンティを計算し、時間経過とともにトピックの変化を考慮している。この手法はLDAモデルの改善自体に注目しているが、得られた結果とユーザのTwitter使用履歴の比較は行っていない。

古賀ら[9]はトピックモデルを用いて、Twitter上のユーザ推薦手法を提案した。古賀らが提案した手法は対象ユーザのフォローしているアカウント名、リツイート内容、リストなどを文書データとして、潜在トピックを抽出する。この手法は確かにユーザ推薦を目指しているが、使っているデータが対象ユーザの嗜好を正確に体現できるかという疑問がある。つまり、データに対して、異なる重みを付けた方が良いのではないかと考えている。

3. 提案手法

3.1 定義

本論文中で使用する用語について定義する。

フォロー：誰かをフォローするとは、次のことを意味している。そのアカウントのツイートが配信されるようになる。そのアカウントのツイートが、ユーザのホームタイムラインに表示される。そのアカウントはユーザにデイレクトメッセージを送信できる[10]。

トピック：ここでは単に新聞記事やニュースの話題、あるいは話題になる事件を指すだけではなく、一定の分野を代表する名詞を指す。World Cup 2018、Barack Obamaあるいは Ritsumeikan Universityなど、これら全てをトピックとみなすことができる。

LDAモデル：大量の文書データから、いくつかの簡潔に文章を記述できる内容を選び出すことができるモデルである[]。簡単に言うと、入力された文章データから、いくつかのトピックを選出して、選出されたトピックごとにいくつかのキーワードで表すことである。

3.2 提案手法の概要

図1に提案手法の流れを示す。本手法は主に四つの段階に分けられる。第一段階は、Twitter API [11]を使って、目標ユーザがフォローしているアカウントのリストを取得する。そして、そのフォローリストの中にあるアカウントのツイート文を取得する。第二段階は、取得した全てのツイート内容を一つの文書とし、LDAモデルの入力として、幾つかの候補トピックを抽出する。次の第三段階は、目標ユーザのTwitterでの様々な行為を利用して、抽出した候補トピックリストに重みを付ける。最後の段階は、重み付けの結果をもとに、抽出したトピックリストを並び替えて、目標ユーザが最も興味を持っているトピックを発見する。

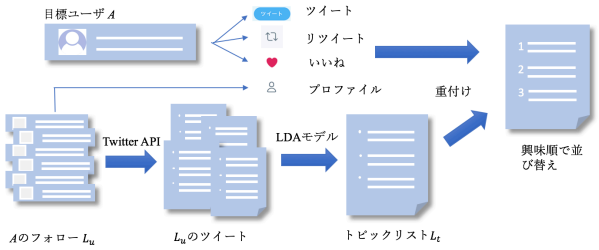

図1　提案手法の流れ

3.3 データ抽出

まずは目標ユーザ A がフォローしているアカウントをすべて取得し、フォローリスト Lu を作る。次は、フォローリスト Lu にあるアカウントに対して、アカウントごと500件のツイート文を取得して、一つの文書にまとめる。もし、対象ユーザが500個のアカウントをフォローしている場合は、利用されるデータ抽出手法を使うと、総計250,000件のツイート文を取得し、500個の文書を作るということである。もし、ユーザのツイート文が500件未満の場合、取得できる数のツイート文を取得する。

一つのツイート文が140字以内なので、LDAモデルはツイート文の処理に適切ではないという意見が少なくない。ここで利用しているのは、一つのユーザの500件のツイート文を一つの文書として扱うAuthor-topic model[12]というデータ処理手法である。

3.4 前処理

自然言語処理分野で、前処理は極めて重要な部分だと思われる。前処理をうまくできなければ、実験結果にも大きく影響して、結局、実験が失敗する場合が少なくない。本節では、実験を行うための前処理について説明する。

3.4.1 ツイート文分析

ここで、一つのツイート文を例として説明する。図2の赤い枠内はツイート文であって、本手法では主に

この部分を使う。次は紫枠内の絵文字を処理する。実際の実験をするとき、絵文字は自動的にアスキーコー

図２　ツイート文の例

ドに変換できる。しかも、その部分の頻度は非常に低いので、LDA モデルを使う際、実験結果にあまり影響がない。したがって、絵文字の部分をそのまま保持しても良いと考える。最後は、黄色枠のトレンドの部分と緑色のメッセージの部分である。多くの研究ではこの二つの要素は単語として直接に使えないので、削除した方が良いとしているが、この二つはユーザの関心が表せる部分であると考えている。トレンドはユーザが現時点でどのような話題について発言するかを表す。メッセージは、現時点でどのアカウントに対する関心を持つかがわかる。この二つはユーザの嗜好推測、特に今後のリアルタイム推測を行う時、極めて重要な内容だと考えている。

3.4.2　文書の正規化

本研究は、英語の文書を対象として実験を行うため、英語以外の文書を事前に削除しなければいけない。実験では、正規表現を用いて、数字や英語の単語以外の内容を認識し、削除する。これで、文書に混ざっている日本語や中国語などの英語以外の内容が削除でき、残された内容は次の段階の入力データになる。

3.4.3　ストップワード辞書

Bag of Words を使って、文書を分割する前に、ストップワードを削除する必要がある。なぜなら、この部分の出現頻度があまりに高いので、LDA モデルのキーワード抽出の精度が下がる可能性が高い。ストップワード辞書はすでに多く作られているが、基本的には論文や新聞記事など整った文書に対する辞書である。しかしながら、SNS 上では、日常の言葉の方が多く、ユーザが勝手に作った言葉も多い。既存の辞書は SNS の文に対して適応できない場合が多いため、改めて SNS 用のストップワード辞書を作る必要がある。現在、SNS ストップワード辞書として 362 個単語を収録している。

作成されたストップワード辞書を用いて、Bag of Words を利用し、ストップワードを削除して文書を単語ごとに分ける。

3.4.4　ステミング

自然言語処理、特に英語文の処理において、ストップワードの削除とともに、ステミングはもう一つの重要なステップである。ステミングとは、主に英語文を処理する時、語幹が同じである単語を一つの単語と認識するという処理である。例えば、maps と mapped の二つの単語を同じく、map という語幹として認識できるようにする過程である。本実験でのステミング処理は、Python の National Language ToolKit (NLTK)[13]というライブラリを利用して行う。

3.5　LDA モデル

Scikit-learn ライブラリ[14]の LatentDirichletAllocation という関数を利用して、事前に作成した文書を LDA モデルの入力として、トピック抽出を行う。この段階で、事前に設定した個数のトピックを抽出する。トピックごとに対して、設定した個数のキーワードを選び出す。ここで、いくつのトピックとキーワードを選び出すと最適な結果が得られるかは最も重要な課題である。この点について、後の実験を通じて検証する。LDA モデルを適用することで、候補トピックリストを作ることができる。この時、候補トピックリスト中の全てのトピックの重みを同じく 0 にする。

3.6　重み付け

ユーザが候補トピックリストの中でどのトピックに最も興味を持っているかを決めるため、重み付けをする必要がある。目標ユーザ A が日常行っている、Twitter での行為を使って、重み付けする。まずは、Twitter の「いいね」というユーザの行為を使う。「いいね」と登録したツイート文はユーザの関心や興味を最も表現できる内容だと考えられる。したがって、Twitter API を使って、目標ユーザ A の「いいね」の内容を全て取得し、前の段階で作ったトピックリスト中のトピックごとに、それぞれのコサイン類似度を計算する。それぞれの類似度の結果はそれぞれトピックの重みとして使うことができると考えられる。具体的には下の式のように表現できる。

$$W_{T_i} = \sum_{n=1} \mathbb{1} \cos(T_i, l_n)$$

ここで、*Ti* はトピックリスト中の *i* 番目のトピックを指している。*ln* は目標ユーザ A が登録した *n* 番目の「いいね」の内容である。つまり、この式は *i* 番目のトピックに対して、目標ユーザ全部の「いいね」の内容とコサイン類似度を計算し合計して、*i* 番目のトピックの重みを得るという計算を行う。

また、目標ユーザ A が他のアカウントに対してリツイートする内容と自分がツイートする内容も重要だと考えている。重みの値だけを変化させ、「いいね」と同じように重み付けをすれば良いと考えている。実験では、対象ユーザが興味を持っている可能性に従って、「いいね」の内容を 1、リツイート内容を 0.8、ツイート内容を 0.5 に重みを設定する。トピックごとに対して、それぞれ重みを計算した上で、次の段階に移行する。

3.7　トピックランキング

重み付けの結果、全てのトピックに重みが付いているため、その結果をもとに、ランキングをする。ここで、単に重みの大小でランキングして、上位のトピックはユーザが興味を持っているトピックであると推測する。しかしながら、下位のトピックは全く関心を持っていないというわけではなく、単に上位のものより興味が少ないと推測している。

4. 評価実験

4.1　実験概要

本実験をスムーズに進めるため、いくつの要求を満足するユーザを選択する。目標ユーザは必ずアクティブユーザである。つまり、目標ユーザは Twitter で、十分な使用履歴を残していることとする。この使用履歴とは、投稿することや「いいね」を登録すること、あるいは他のアカウントのツイート文に対しリツイートすることなどのユーザ行為を指す。そうではなければ、取得できるデータ量が足りないため、コールドスタート問題が起きて、実験の精度が下がる可能性が高い。本研究では、ユーザのフォロー情報をもとに研究を行うので、目標ユーザは大量のアカウントをフォローすることが望ましい。また、現在は英語のツイートを対象として、文の分割や LDA モデルの利用をしている。したがって、実験の対象はすべて英語の内容を用いる。

上で説明した要求に従って、本研究の実験データを作成した。Twitter で新しいアカウントを作って、ランダムで 200 名の英語のアカウントをフォローする。そして、同じくランダムでフォローしているアカウントのツイートを「いいね」を多数登録する。これで、本実験データの作成は完了である。次に実験の前処理の段階について説明する。

4.2　実験

LDA モデルは少なくとも二つの変数を設定する必要がある。一つ目はいくつのトピックを抽出するかである。もう一つは、各トピックに対して、いくつのキーワードを選びだすかである。ここで、モデルの精度を上げるため、実験を行った。

4.2.1　抽出するトピック数

まずは、本提案手法で、いくつのトピックを抽出すれば一番良い結果を得られるかを検討する。LDA モデルを評価するのは非常に困難な課題であるため、現在、基準となる評価手法は存在しない。ここで Chang ら[15]が提案した Word Intrusion という手法で、LDA モデルのトピック数を決定する。図 2 に示すように、事前にトピック数を 10 に設定した場合に、最も良い精度が得られた。したがって、本実験では、トピック数を 10 に決定した。

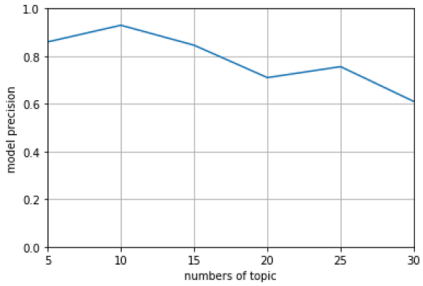

図 2　トピック数の決定

4.2.2　抽出するキーワード数

次は、各トピックに対して、いくつのキーワードを抽出すべきかという問題である。同じデータセットを利用して、10 個と 20 個のキーワードを抽出した。表 1 は 10 個のキーワードを抽出した結果、表 2 は 20 個の結果である。二つの表から見ると、キーワードの個数に関わらず、どちらもこの目標ユーザの嗜好がバスケットボールであるということが推測できる。しかしながら、最も注目すべき点は、下の Weight が 0 の部分である。10 個のキーワードの場合、下の三つの項目の重みが 0 で、20 個の場合、0 は一つの項目だけである。また、最も上の方、いわゆる提案手法のアルゴリズムで推測したユーザの嗜好の部分、すなわち表 1 の場合、上位 2 件の重みが近い。それに対して、表 2 の場合、上位 2 件の重みの差が大きい。また、中間のトピックの重みを見ると、表 2 の部分が表 1 より幅広いので、並び替えの精度は高いと考えられる。したがって、20 個のキーワードを抽出するのが最も良い結果が得られ

表 1　10 個のキーワード

Topic	Weight
#Topic 0 : game nba team win review season 10 player great tonight	0.189
#Topic 5 : night album tonight video live song tour music wait sal	0.132
#Topic 8 : china travel chinese world internet year 2017 visit says high	0.105
#Topic 4 : trump president says people year jerusalem state sexual tax al e	0.099
#Topic 1 : data ai la en el science learn learning big google	0.095
#Topic 2 : space earth live coffee open join video launch 2016 available	0.088
#Topic 9 : home christmas best make gift perfect save 2017 2018 look	0.048
#Topic 3 : great tonight like don make best ll ready happy got	0.000
#Topic 6 : wine gt weekend 10 2017 wines food chicago chef dinner	0.000
#Topic 7 : google dm like glad look great drive team steps happy	0.000

ると考えられる。

表２　２０個のキーワード

Topic	Weight
#Topic 2 : game team nba win season tonight great big best ball don year world watch play look player league games come	0.231
#Topic 6 : tonight like 2017 video night live coming happy album tomorrow music watch year amazing week tour 11 ve best wait	0.126
#Topic 7 : china chinese says year people city president world police state trump report york south old 000 media russia capital 2017	0.069
#Topic 8 : wine travel weekend wines 10 gt coffee best chef dinner food tasting night open holiday delicious tonight morning taste list	0.067
#Topic 9 : data ai google science app learn gt space steps using big learning latest earth use open tips read watch business	0.066
#Topic 0 : christmas home 2017 gift night best great shooting perfect 2018 magazine makes save sale look price buy london deal maket	0.065
#Topic 1 : trump alabama senate jones roy election tax says race president sexual house year special say win north seat korea star	0.065
#Topic 3 : chicago great join 2016 happy photo hope food wait work sunday excited event walk 30 team second looking sending honored	0.064
#Topic 4 : la el en que people world los women american years al make lo work las man real history need change	0.035
#Topic 5 : like dm glad look happy ll great check team ve drive experience feel sharing email family info account questions free	0.000

4.3　実験結果の評価

今回の実験では、新しい Twitter のアカウントを作成して、ランダムに様々なアカウントをフォローしたり、「いいね」を登録したりした。そして、最後の並び替えの結果は表 2 に示したように、抽出された 10 個のトピックの中で、バスケットボールの順位が最も高かった。そして、今回作成された Twitter アカウントのフォローリストを実際に見ると、やはりバスケットボールに関するアカウントを一番多くフォローしている。つまり、この対象ユーザはバスケットボールに関するトピックが最も好きだという本研究の推測と一致している。

現在、トピックモデルの抽出結果を有効に評価する方法がまだないので、アンケートの形で 20 人の英語の Twitter ユーザを実験対象として、実験結果を評価した。その結果、20 人のうち、19 人が本提案手法で推測されたトピック内容に確かに興味を持っていると答えた。しかしながら、19 人のうち 6 人は、本提案手法で推測された嗜好が一番好きなものではないと答えた。つまり、本提案手法は対象 Twitter ユーザの嗜好を発見することができるが、対象ユーザの好みの順番は正確ではない場合が少なくない。

5. 考察

本稿で紹介した実験は、3 章で提案した手法で、対象 Twitter ユーザの嗜好を発見し、推測するものである。そして、手動で対象ユーザのフォローリストや行動履歴をチェックし、推測の結果が正確かどうかを判断した。さらに、提案手法で発見した嗜好ランキングと被験者の実際の好み順番が一致するかどうかを比較した。前節の評価実験で述べたように、最終の結果は必ずしも良いとは言えないので、まだ改善すべきところが多くあると考えている。

本研究では、対象ユーザがフォローしているユーザ情報を利用するが、フォローされたユーザの間の関連性はまだ考えていない状態である。実際の Twitter の使用経験によると、Twitter ユーザは、興味を持っている分野の中で複数のアカウントをフォローする傾向がある。例えば、ある対象ユーザが Jack Wilshere と Mesut Özil の二人のアカウントをフォローしている場合、この二人ともサッカー選手である。さらに、同じアーセナルというサッカーチームのチームメートである。このような情報を先に抽出できれば、対象ユーザがサッカーのトピックが好きだと推測するだけではなく、より正確にアーセナルというサッカーチームのトピックが好きだと推測できると考えている。すなわち、フォローリスト中のアカウントの間の関係を有効に利用すれば、推測の精度を向上できると考えている。

6. おわりに

本論文では、Twitter ユーザのフォロー情報をもとに、LDA トピックモデルを利用して、ユーザの嗜好を発見する手法を提案した。対象ユーザが Twitter でフォローしているアカウントの投稿内容は、対象ユーザが最も興味を持っている内容と考えられる。それを利用して、対象ユーザの行為と合わせて、対象ユーザの嗜好ランキングを作る。このランキングにより、対象ユーザが何に最も興味を持つかを発見する。

現在、本研究はまだ多く改善の余地がある。前節に述べたフォローされたユーザの間の関連以外に、今後の課題はまだいくつか残っている。実験中の LDA モデルから抽出したキーワードリストの中に、ノイズが少し残っている。それについて、作成した SNS 辞書がまだ十分ではないと考えているので、これは今後の課題の一つとして、追加し続けていく予定である。また、現在は英語にしか対応できないため、今後は提案した手法を日本語にも対応できるようにすることを考えている。

参　考　文　献

[1] Kwak, Haewoon, et al. “What is Twitter, a social network or a news media?”, Proceedings of the 19th international conference on World wide web. ACM, 2010.

[2] ソーシャルメディアラボ, “11 のソーシャルメディア最新動向データ”,

https://gaiax-socialmedialab.jp, 2016

[3] Blei, David M., NG, Andrew Y., JORDAN, Michael I., “Latent dirichlet allocation”, Journal of machine Learning research, pp.993-1022, 2003

[4] Albert Bifet, et al. “Detecting Sentiment Change in Twitter Streaming Data”, Workshop and Conference Proceedings 17, pp.5-11, 2011

[5] H Takamura, K Tajima, “Tweet Classification Based on Their Lifetime Duration”, Proc. of ACM CIKM, pp.2367-2370, 2012

[6] 齋藤準樹, 湯川高志, “ソーシャルブックマークを基にした Twitter ユーザの興味語抽出・推薦手法の提案と評価”, 情報処理学会研究報告：情報基礎とアクセス技術, pp.1-8, 2011

[7] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X., “Comparing twitter and traditional media using topic models”, In European Conference on Information Retrieval, Berlin, Heidelberg, pp.338-349, 2011

[8] 佐々木謙太朗, 吉川大弘, 古橋武, “Twitter におけるユーザの興味と話題の時間発展を考慮したオンライン学習可能なトピックモデルの提案”, 情報処理学会論文誌数理モデル化と応用, pp.53-60, 2014

[9] 古賀裕之, 谷口忠大, “潜在トピックに着目した Twitter 上のユーザ推薦システムの構築”, ヒューマンインタフェースシンポジウム, pp.867-872, 2010

[10] https://help.twitter.com/ja/using-twitter/following-faqs

[11] https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/overview

[12] Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T, “Probabilistic author-topic models for information discovery”, In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 306-315, 2004

[13] http://www.nltk.org/

[14] http://scikit-learn.org/stable/

[15] Chang, Jonathan, et al. “Reading tea leaves: How humans interpret topic models”, In Advances in neural information processing systems, pp. 288-296., 2009