DEIM Forum 2021 I31-2

SNSにおけるフォローユーザツイートを用いた受動的ユーザの興味推定手法の開発

張 琪 † 羽山 徹彩†

† 長岡技術科学大学工学研究科 情報・経営システム工学専攻 〒 940–8603 新潟県長岡市上富岡 1603-1
E-mail: †t-hayama@kjs.nagaokaut.ac.jp

あらまし ソーシャルネットワーク上のコンテンツに基づくユーザモデリングはユーザごとの嗜好に応じた情報推薦を行うために研究開発されてきた．従来研究の多くは Twitter で積極的にツイートする能動的ユーザを対象として，そのツイートを分析し，興味を推定してきた．その一方で，ツイートせずに，情報収集のみを目的とした受動的ユーザも，一定数以上いるものの，それら受動的ユーザを対象とした興味推定については手掛かり不足のため，ほとんど研究開発されてこなかった．そこで，本研究ではフォローしているユーザのツイートに基づく能動的ユーザの興味抽出手法を開発し，それを転用した受動的ユーザの興味推定手法の開発を目的とし，実施した．提案手法ではまずフォローしているユーザのツイートから能動的ユーザの興味トピックを抽出するために，フォローしている一部のユーザの頻出トピックの抽出支援，ユーザごとの適切なトピック抽出数の設定，およびトピックに無関係なツイートの除去を適用し，そして，それらを受動的ユーザの興味推定のために転用した．評価実験ではまず能動的ユーザの興味抽出手法に対し，データセットをもとに単純なトピック抽出手法と比較することで，有効性を確認した．次に，受動的ユーザの興味推定手法に対し，12 ユーザの興味トピック評価付きデータをもとに単純なトピック抽出手法と比較することで，有効性を確認した．さらに受動的ユーザを対象とした提案手法が能動的ユーザを対象とした手法と同等な精度であり，その転用が有効であることが示唆された．

キーワード ユーザ興味推定，SNS データ分析，プロファイル生成，トピック抽出，ユーザモデリング

1 序 論

ソーシャルネットワークサービス（SNS）は，2000 年頃から登場して以来，急速に普及してきた．現在，多くのユーザが Twitter などの SNS で様々なタイプの情報（医療情報やニュースなど）を閲覧したり，収集したりしている [11]．そのため，このような SNS 上のユーザの振舞いは，それぞれのユーザの興味が影響されており，ユーザの興味に適合したコンテンツや情報の推薦を提供するうえで，重要な役割を果たす可能性が高い．そのため，Twitter の対象となるユーザが投稿したツイート文から，ユーザの興味プロファイルを推定する研究が数多くなされてきた．そのほとんどの研究が，Twitter 上で積極的にコンテンツを生成する能動的ユーザのユーザモデリングに焦点があてられてきた．しかしながら，ソーシャルネットワーク上において，コンテンツを生成せずに情報を閲覧し，収集するだけの受動的なユーザは増加しており，44%のユーザが一度もツイートを投稿したことがないなど大きな割合を占めてきた [10]．

これまで，受動的ユーザの興味推定を目的とし，いくつかの研究がなされてきた．例えば，フォローしているユーザの名前と Wikipedia のエンティティを関連付けし，Wikipedia のエンティティのカテゴリ構造を利用した，ユーザの興味の導出手法が開発されてきた [2]．しかしながら，そのようなエンティティを利用した方法では，Wikipedia のエンティティに含まれるユーザの名前に対して，興味導出の精度を高めることができるものの，有名人の Twitter アカウントにしか対処することができないため，フォローしているユーザの大部分の情報を活用することができない．その他の方法として，フォローしている Twitter アカウントの自己紹介文を利用する方法が開発されてきた [10]．このような自己紹介文には 160 文字の制限で，職業や興味が直接的に記述されていることもあり，興味プロファイル作成のために有用な手掛かりといえる．しかしながら，フォローしている多数のユーザの自己紹介文のなかから，対象ユーザの興味を示す記述を抽出することは大変難しく，また Twitter アカウントの自己紹介文では更新遅れや記述不足なども多く，その活用範囲には限界があるといえる．その一方で，フォローしているユーザのツイートを利用して，受動的ユーザの興味を導出することが考えられる．一般的に，ツイートは更新頻度が高く，ユーザごとの興味が含まれており，能動的ユーザの興味推定にも効果的に利用されてきた．しかしながら，受動的ユーザを対象とした場合には興味を含んだ内容を抽出するための手掛かりがほとんどなく，興味以外の内容も含んだ膨大な情報量となるため，適切な抽出が非常に困難となる．

そこで，本研究では Twitter を対象に，能動的ユーザの興味抽出手法を転用した，フォローしているユーザのツイートから受動的ユーザの興味推定手法の開発を目的とし，実施する．提案手法ではまずフォローしているユーザのツイートから能動的ユーザの興味トピックを抽出するために，フォローしている一

部のユーザの頻出トピックの抽出支援，ユーザごとの適切なトピック抽出数の設定，およびトピックに無関係なツイートの除去を適用し，そして，それらを受動的ユーザの興味推定のために転用した．評価実験ではまず能動的ユーザの興味抽出手法に対し，データセットをもとに単純なトピック抽出手法と比較することで，有効性を確認した．次に，受動的ユーザの興味推定手法に対し，12 ユーザの興味トピック評価付きデータをもとに単純なトピック抽出手法と比較することで，有効性を確認した．さらに受動的ユーザを対象とした提案手法が能動的ユーザを対象とした手法と同等な精度であり，その転用が有効であることが示唆された．

2 先行研究

積極的にツイートする能動的ユーザの興味プロファイルを推定する研究の多くは，そのユーザが投稿したツイートの分析に基づいている [1]．例えば，Siehndel ら [12] は，ユーザのツイート文から，Wikipedia のエンティティに基づいた興味プロファイルを生成している．Kapanipathi ら [6] は，Wikipedia のエンティティとそのカテゴリーを利用し，ユーザのツイートに出現する Wikipedia のエンティティを多く含むカテゴリーを活性化ノードとして抽出することで，そのユーザの重み付き階層型興味グラフを生成している．Piao ら [8] と Orlandi ら [7] は，Wikipedia のカテゴリーを利用する代わりに, DBpedia を利用してユーザの興味プロファイルの生成を行っている．DBpedia はエンティティに関する背景知識を提供しており，エンティティのカテゴリーだけでなく，様々なプロパティに関連するエンティティも含まれている．Piao ら [9] は，Twitter 中のリンクを推薦するシステムのために，DBpedia の構造的に異なるカテゴリや関連エンティティを探索することで，ユーザモデリングの質を向上させている．本研究ではツイートをしない受動的ユーザを対象としているため，これら方法をそのまま適用することが難しい．

さらに，能動的ユーザを対象にフォロー関係を利用した方法も研究されてきた．例えば，Faralli ら [4] は，Wikipedia のエンティティにリンクされているフォローしている名前を利用し，ユーザ推薦のためにユーザ興味プロファイルの推定方法を開発してきた．この方法ではフォローしているユーザのプロファイルを活用することで，フォローしているユーザのツイートを分析するよりも，より精度の高いユーザ興味プロファイルが構築できることを示してきた．しかしながら，フォローしているユーザのなかで，Wikipedia のエンティティにリンクできるのは平均で 12.7%にすぎないことも示唆している．受動的ユーザのユーザ興味プロファイルの推定に，フォロー関係を利用できるものの，単にフォローしているユーザ名だけを活用することは難しいといえる．また本研究ではフォローしているユーザのツイートを分析し，トピックを絞り込むことで，ユーザ興味プロファイルの精度を高める点で，この研究とは方法が異なる．

3 受動的ユーザのモデリングへの設計指針

フォローしているユーザのツイートに基づく受動的ユーザの興味推定手法を開発するために，まずその前提条件と，そのなかでの新たな検討事項について述べる．さらに，受動的ユーザをモデリングするための設計指針について挙げる．

3.1 前提条件

本研究ではユーザモデリングとして，生起確率を伴う単語集合 (Bag-of-Words) でトピックを表現するトピックモデルを用いる．つまり，対象ユーザがフォローしている全てのユーザのツイートを入力として，対象ユーザが興味のあるトピックを表現した単語集合のリストを出力する．その具体的な処理手順は，以下の通りである．

手順 1. 対象ユーザの Twitter ID を入力する．
手順 2. そのユーザがフォローしているユーザの Twitter ID と，そのフォローしているユーザの直近一定期間のツイートを収集する．
手順 3. その全てのツイートをトピック解析し，対象ユーザが興味のあるトピックを抽出する．
手順 4. 対象ユーザの興味トピックのリストを出力する．

手順 3 ではフォローしているユーザの全てのツイートに対してトピック解析し，トピックのリストを生成する．そのようなトピックのリストには対象ユーザと無関係なトピックも含まれるため，対象ユーザが興味があると推定されるトピックだけを抽出する必要がある．しかしながら，ツイートを投稿しない受動的ユーザに対しては興味を推定できる手掛かりがほとんどないため，正確に抽出することが困難である．そこで，本研究では，まず能動的ユーザを対象とし，フォローしているユーザのツイートから興味があるトピックの抽出手法を開発する．そして，その手法を転用することで受動的ユーザの興味を推定する．

3.2 能動的ユーザを対象とした興味トピック抽出手法の検討

能動的ユーザを対象に，そのユーザのツイートに含まれるトピックが，フォローしているユーザのツイートに含まれるトピックから，どのように高精度に抽出できるかを検討する．通常，フォローしているユーザのツイートには，対象ユーザの興味以外のトピックも多数含まれる．そのため，フォローしている全ユーザのツイートをトピック解析後に，対象ユーザのツイートに含まれるようなトピックの選定が必要となる．

対象ユーザがフォローしているユーザが多く扱っているトピックは，そのユーザにとっても興味がある可能性が高い．そのため，能動的ユーザの興味トピック抽出では，フォローしているユーザの全ツイートをトピック解析した後に，高頻出なトピックを選定することが考えられる．しかしながら，コミュニティ限定的や特殊なトピックに対しては抽出できない可能性がある．そのようなトピックはフォローしている一部のユーザのツイートにしか含まれていることもある．特に，それらユーザのツイート数が少ない場合には，単にツイートを解析後のト

ピック出現頻度だけで，その種のトピックを抽出することが困難である．

対象ユーザの興味のあるトピックの抽出数についても考慮が必要である．一般的に，対象ユーザのツイート数が多いほど，より多くのトピックが含まれる可能性が高い．その一方で，フォローしているユーザのツイート数が多い場合には対象ユーザが興味のあるトピックに比べ，より多くのトピックが生起される可能性が高い．つまり，対象ユーザの興味のあるトピックの抽出数は，すべてのユーザ一律に定めることが難しい．

また，装飾文字が多く含まれたり，トピックを直接含まないツイートがトピック解析を阻害する．アスキーアートなどツイートを装飾するための無意味な文字列や，簡単な応答などトピックが特定できない文は，単語の共起性を扱うトピック解析を阻害し，無関係な単語が生起され易くなる．特に，アスキーアートは文字数制限のあるツイートに対し，大部分を占めるため，それを含んだツイートはトピック解析に不要といえる．

以上をまとめると，能動的ユーザを対象とした興味トピックの抽出に対し，以下の課題が挙げられる．

- フォローしている一部のユーザだけのトピックは全ユーザを対象としたツイート解析結果のトピック出現頻度だけで，抽出が困難である．
- 対象ユーザの適切なトピック抽出数をユーザ一律に定めることが難しい．
- 装飾文字が大部分を占めたり，トピックが直接含まれないツイートが，トピック解析を阻害する．

3.3 設計指針

本研究では能動的ユーザの興味トピック抽出手法を転用した，フォローしているユーザのツイートに基づいた受動的ユーザの興味推定手法を開発する．そのために，3.2 節の課題を解決するための設計指針を以下に挙げる．

- フォローしている一部のユーザがツイート数が少なくても，そのユーザ間で高頻度なトピックも抽出可能にする．
- フォローしているユーザの数やツイート数などのトピック解析のためのデータ量に応じて，適切なトピック抽出数の設定を可能にする．
- 文の長さに対し，内容表現の割合が少ないツイートはトピック解析の対象から除外できるようにする．

4 フォローユーザのツイートから受動的ユーザの興味推定手法の実装

3.3 節の設計指針をもとに，フォローしているユーザのツイートから受動的ユーザの興味を推定する手法を開発した．その処理手順について，図 1 に示す．

手順 2 では手順 1 で入力した Twitter ID と Twitter API[1] を使用し，Twitter サーバから “対象ユーザがフォローしているユーザリスト”，“それらユーザの直近一定期間のツイート”，および “対象ユーザの直近一定期間のツイート”のデータを収集

1 : https://developer.twitter.com/ja/docs

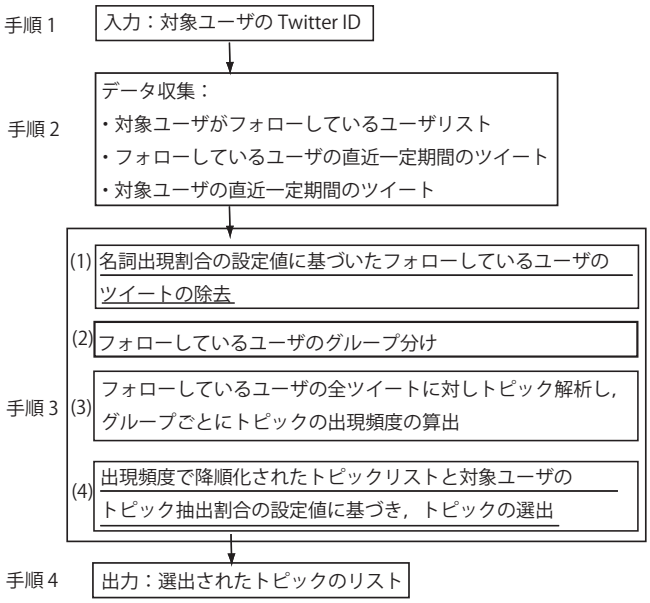

図 1 フォローしているユーザのツイートから受動的ユーザの興味を推定する手法の処理手順．下線部分は能動的ユーザの興味抽出に基づく設定値を使用する処理モジュール．

し，端末のデータベースに登録する．手順 3 では，(1) 名詞出現割合の設定値に基づいたユーザのツイートの除去，(2) フォローしているユーザのグループ分け，(3) フォローユーザの全ツイートに対しトピック解析し，グループごとトピックの出現頻度を算出，そして，(4) トピックの出現順位の降順から，対象ユーザのトピック抽出割合の設定値内に含まれるトピックの選出，の処理順序から成る．手順 4 では，それら選出されたトピックをリストにして，出力する．

実装された提案手法では，手順 3 の (2) フォローしているユーザのグループ分けでは，フォローしているユーザの数と平均ツイート数を特徴として，グループ数を自動的に決定する教師なし分類手法 gmeans [5] が適用されている．また手順 3 の (3) 全ツイートに対するトピック解析では，教師なし学習アルゴリズムのトピックモデル LDA(Latent Dirichlet Allocation) [3] に対し，そのトピック数を推定する HDP(Hierarchical Dirichlet Process) を適用した HDP-LDA [13] が用いられている．

手順 3 では，(1) の名詞出現割合の設定値，および (4) の対象ユーザのトピック抽出割合の設定値に対し，能動的ユーザに関するデータセット（5 章に記述）を利用し，能動的ユーザの興味抽出に最適な値が決められる．手順 3 の (1) では，各ツイートに含まれる単語品詞数に対し，名詞数が設定値以下の割合ならば，そのツイートを削除する．そのために，各ツイートに含まれる名詞出現割合を変化させながら興味トピックの抽出精度を調査し，精度が最高の名詞数割合を設定値として用いる．また手順 3 の (4) の対象ユーザのトピック抽出割合では，ユーザごとに異なる設定値を割り当てられる．そのために，フォローしているユーザの数と平均ツイート数をもとに gmeans でグループ分けし，グループごとに最高のトピック抽出精度となるトピック抽出割合を設定値とする．

提案手法の特徴を以下にまとめる．

- フォローしているユーザのツイート量に応じ，最適なトピック抽出割合が割り当てることで，対象ユーザごとに適切なトピック数が決まる．
- フォローしているユーザをツイート量が同じ程度のグループごとにトピック出現頻度を算出し，高頻度のトピックを選出することで，フォローしているユーザがツイート数が少ない場合も，そのなかの頻出トピックも抽出され易くなる．
- 名詞以外の単語が多く占めるツイートをトピック解析の対象としないことで，トピックに無関係な共起語を除いた適切なトピックモデルが構築される．
- 能動的ユーザを対象とするデータセットで最適な設定値が決められるため，能動的ユーザの興味抽出の精度が高まる．
- 対象ユーザに関する情報を直接利用していないため，受動的ユーザを対象とした興味推定への転用にも有効である．

5 データセット

提案手法の設定値を決めるために，ユーザの興味や挙動の偏りがないような能動的ユーザに関する Twitter データセットを作成した．本データセットには，ユーザの Twitter ID と，そのユーザのツイート，およびフォロワー数とフォローしているユーザ数とともに，そのユーザがフォローしているユーザの Twitter ID リストと，それらユーザのツイートも含まれる．

本データセットの作成ではまず Twitter API を使用して，ニュースサイト (@YahooNewsTopics) のフォロワーから約 150 万のユーザの Twitter ID を収集し，無作為に 2000 ユーザを抽出した．それらユーザから，アクセス権限の制限のない 541 ユーザと直近 3ヵ月内にツイートがない 54 ユーザ，およびそのフォローしているユーザのツイートが十分に収集できない 154 ユーザが除かれ，1251 ユーザが選出された．

次に，その選出された 1251 ユーザに対し，直近 3ヵ月のツイート，フォローしているユーザの数と Twitter ID リスト，およびフォロワー数が収集された．さらに，そのフォローしているユーザに対しても，直近 3ヵ月のツイートが収集された．

本データセットに含まれる収集した Twitter データに含まれるユーザの直近 3ヵ月に含まれるツイート数の分布，およびユーザがフォローしているユーザ数の分布について，それぞれ図 2，および図 3 に示す．また，本データセットに含まれるユーザ数，そのユーザのツイート数の平均と標準偏差，フォロワー数の平均と標準偏差，および，フォローしているユーザ数の平均と標準偏差について，表 1 にまとめる．

表 1 本データセットのプロパティ(ユーザ数:1251)

	平均	標準偏差
ツイート数	724.33	753.75
フォロワー数	265.56	316.14
フォローしているユーザ数	392.86	357.74
フォローしているユーザのツイート数	181.89	77.41

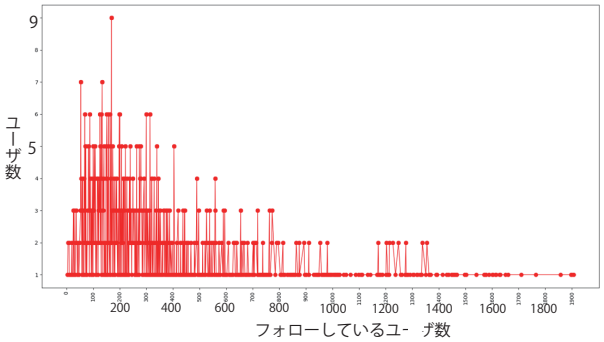

図 2 データセットに含まれるユーザがフォローしているユーザ数の分布

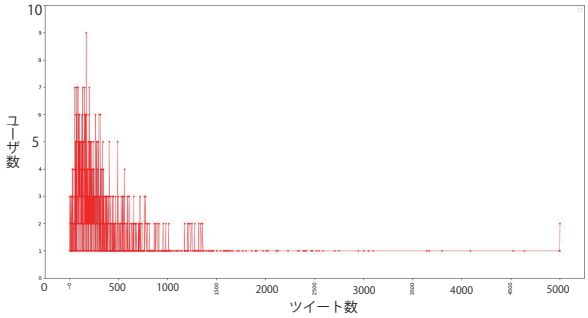

図 3 データセットに含まれるユーザの直近 3ヵ月に含まれるツイート数の分布

6 評　　価

6.1 概　　要

受動的ユーザの興味推定に関する提案手法の有効性を検証するために，実データに基づくトピックに対する興味評価付きデータを作成し，単純なトピック抽出手法と比較することで，提案手法の有効性を検証した．

実データに基づくトピックに対する興味評価付きデータの作成には，まず 12 ユーザの Twitter ID を入力値としたトピック解析結果からトピックリストをそれぞれ抽出した．次に，ユーザごとに，トピックリストから 20 分割位置にあるトピック 20 個を抽出する系統サンプリングを行い，図 4 に示すように，トピックごとに単語クラウドと 5 リッカード尺度をともなった興味調査アンケートを作成した．ユーザごとに，各自の Twitter ID から生成されたアンケートに回答し，その結果を興味評価尽きデータとして評価に用いた．本実験に参加したユーザの Twitter データの概要として，フォローしているユーザの数と平均ツイート数，およびユーザのトピック抽出割合ついて，表 2 に示す．

提案手法の設定値決めと評価の算出のために，式 (1)-(3) に示すような，適合率，再現率，および F 値の精度が用いられた．

表 2　本実験に参加したユーザの Twitter データの統計

参加ユーザ数	トピック抽出割合				フォローしている平均ユーザ数				フォローしているユーザの平均ツイート数			
	平均	標準偏差	最大	最小	平均	標準偏差	最大	最小	平均	標準偏差	最大	最小
12	0.47	0.10	0.95	0.20	183.58	38.91	491.00	13.00	232.60	24.46	420.45	77.21

図 4　トピックに対する興味調査アンケートの例

$$\text{適合率} = \frac{Num(Topics(user) \cap ExTopics(followees))}{Num(Topics(user))} \quad (1)$$

$$\text{再現率} = \frac{Num(Topics(user) \cap ExTopics(followees))}{Num(ExTopics(followees))} \quad (2)$$

$$F\text{ 値} = \frac{2 * \text{適合率} * \text{再現率}}{\text{適合率} + \text{再現率}} \quad (3)$$

ここで，$user$，$followees$，$Topics(A)$，$ExTopics(A)$ および $Num(Topics(A))$ はそれぞれ，対象ユーザ，対象ユーザがフォローしている全ユーザ，A が投稿したツイートに含まれる全トピック，A が投稿したツイートに含まれる全トピックから提案手法/比較手法による抽出後のトピック，および A が投稿したツイートに含まれるトピック数を表している．適合率では対象ユーザのツイートに含まれるトピック数に対し，対象ユーザがフォローしているユーザのツイートに提案手法/比較手法を適用した結果に含まれる対象ユーザのトピックとの一致数の割合を表している．再現率では対象ユーザがフォローしているユーザのツイートに提案手法/比較手法を適用した結果に含まれるトピック数に対し，そのなかで対象ユーザのツイートに含まれるトピックとの一致数の割合を表している．F 値は，適合率と再現率の両方を考慮した評価値となる．

比較手法では対象ユーザのフォロワーしているユーザのツイートを HDP-LDA [13] でトピック解析し，その結果のトピックを出現頻度の降順に並べ，設定割合で抽出した結果を用いた．その設定割合には，フォローしているユーザのツイートをトピック解析し，F 値が最高であったトピック抽出割合 0.3 およびそれを適用しないトピック抽出割合 1.0 の 2 種が用いられた．

6.2　提案手法の設定値

提案手法の設定値を決めるために，5 章で述べたデータセットを利用して，“対象ユーザごとのトピック抽出割合”，および“ツイート除去のための名詞出現割合”による興味トピック抽出精度の結果について，それぞれ 6.2.1 節および 6.2.2 節で述べる．さらに，それら設定値を用いた能動的ユーザの興味トピック抽出の結果について，6.2.3 節で述べる．

6.2.1　対象ユーザごとのトピック抽出割合

データセットに含まれる 1251 ユーザをグループ分けした結果，図 5 に示すように 53 グループに分類された．次に，その 53 グループに対し，能動的ユーザの興味トピック抽出の精度が高いトピック抽出割合の設定値について調査を行った．トピック抽出割合に対する各グループのトピック抽出精度の結果について，図 6 に示す．

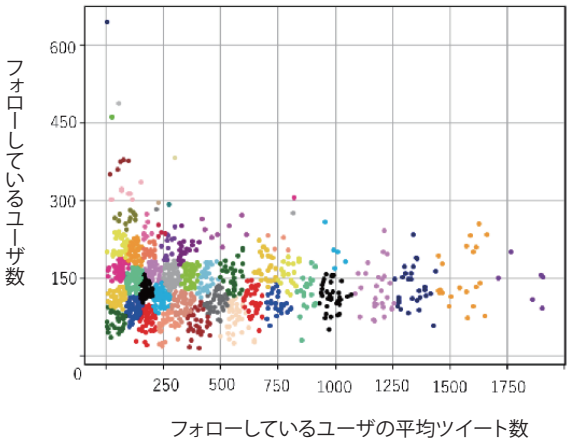

図 5　データセットに含まれるユーザをグループ分けした結果

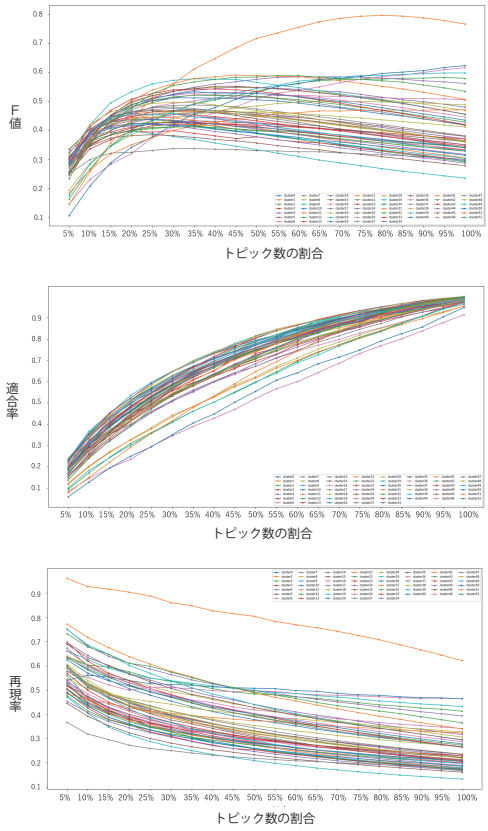

図 6　トピック抽出割合に対する各グループのトピック抽出精度

各グループのトピック抽出精度は図 6 が示すように，同じトピック抽出割合でも，F 値，適合率，および再現率が異なり，

さらに評価値が最も高いトピック抽出割合も各グループで異なることがわかった．F 値を評価の基準として考えた場合では，トピック抽出割合が最大のグループでは 0.95 に設定され，その F 値は 0.54 である．一方で，トピック抽出割合が最小のグループでは 0.15 に設定され，その F 値は 0.43 である．また F 値が最高値 0.84 のグループではトピック抽出割合が 0.70 で，F 値が最小値 0.34 のグループではトピック抽出割合が 0.20 であった．以上から，ユーザグループごとに最適なトピック抽出割合の設定が算出された．その Twitter データの統計について，表 3 に示す．

6.2.2 ツイート除去のための名詞出現割合

ツイート除去のための名詞出現割合について，0.1, 0.2, 0.3, および 0.4 の設定値と，設定なしを用いたトピック抽出精度について調査した．その結果を図 7 に示す．

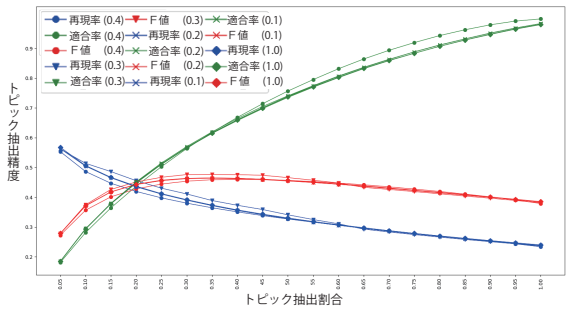

図 7　ツイート除去のための名詞出現割合に対するトピック抽出精度

F 値を基準と考えた場合に，名詞出現率割合の設定を 0.3 にしたときに，トピック抽出割合 0.05 から 0.60 までの間で，最も高いトピック抽出精度が得られた．その次に高いトピック抽出精度は，名詞出現率割合の設定を 0.1, 0.2 および設定なしにしたときに，トピック抽出率 0.05 から 0.60 までの間で，等しい結果であった．そのため，ツイート除去のための名詞出現割合の設定では 0.3 が，能動的ユーザに対する興味トピック抽出精度に最適であることがわかった．

6.2.3 能動的ユーザを対象とした興味トピック抽出の精度

能動的ユーザを対象とした興味トピック抽出の精度について，提案手法に “対象ユーザごとのトピック抽出割合”および “ツイート除去のための名詞出現割合”の最適値を適用した結果を求めた．その結果について，図 8 に示す．

提案手法を適用したトピック抽出精度は適合率，再現率，および F 値について，それぞれ 0.48，0.68，および 0.47 であった．その一方で，トピック抽出割合 0.3 を適用した比較手法は適合率，再現率，および F 値について，それぞれ 0.38，0.56，および 0.38 であり，トピック抽出割合を設定しない比較手法は適合率，再現率，および F 値について，それぞれ 0.24，1.00，および 0.34 であった．これら手法の F 値および適合率を比較すると，提案手法が最も高い結果が得られ，トピック抽出割合 0.3 および 1.0 の比較手法に対し，それぞれ有意水準 1%で有意差が確認された ($F(1, 1250) < 883.98, p < .01$，および

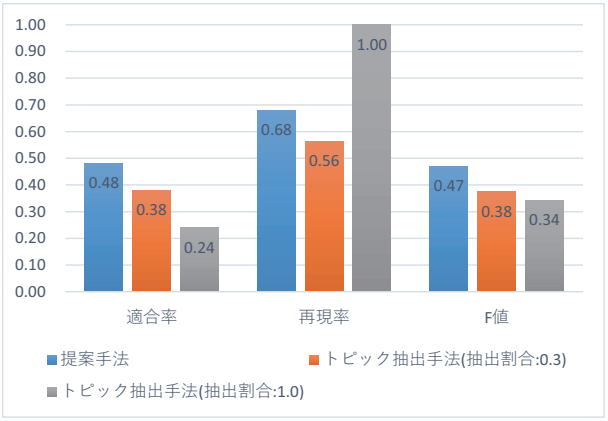

図 8　能動的ユーザを対象とした興味トピック抽出の精度

$F(1, 1250) < 2898.02, p < .01$)．また提案手法とトピック抽出割合 0.3 を適用した比較手法の再現率を比較すると，提案手法が高い結果が得られ，有意水準 1%で有意差が確認された ($F(1, 1250) < 727.81, p < .01$)．

以上から，提案手法は能動的ユーザの興味トピック抽出において，適合率, 再現率，および F 値 の点で有効であることがわかった．

6.3 受動的ユーザを対象とした興味トピック推定

トピックに対する興味評価付きデータの内訳，および受動的ユーザを対象とした興味トピック推定の精度について，それぞれ図 9 および図 10 に示す．本研究ではトピックに対する興味評価付きデータのなかで，ユーザの興味があるトピックを 4 以上に評価された結果を正解データとして用いた．つまり，全トピック 240 に対し，興味のあるトピックは 84 である．

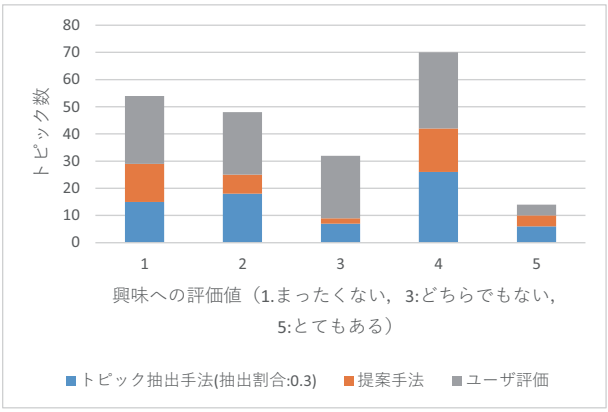

図 9　トピックに対する興味評価付きデータの内訳

提案手法を適用したトピック抽出精度は適合率，再現率，および F 値について，それぞれ 0.53，0.59，および 0.50 であった．一方，トピック抽出割合 0.3 の比較手法を適用したトピック抽出精度は適合率，再現率，および F 値について，それぞれ 0.42，0.36，および 0.38 であり，トピック抽出割合 1.0 の比較手法を適用したトピック抽出精度は適合率，再現率，お

表 3　各ユーザグループの最適なトピック抽出割合の設定値の統計

グループ数	トピック抽出割合				フォローしている平均ユーザ数				フォローしているユーザの平均ツイート数 *			
	平均	標準偏差	最大	最小	平均	標準偏差	最大	最小	平均	標準偏差	最大	最小
53	0.31	0.02	0.95	0.15	311.30	293.72	1163.00	13.00	184.83	30.83	279.99	135.63

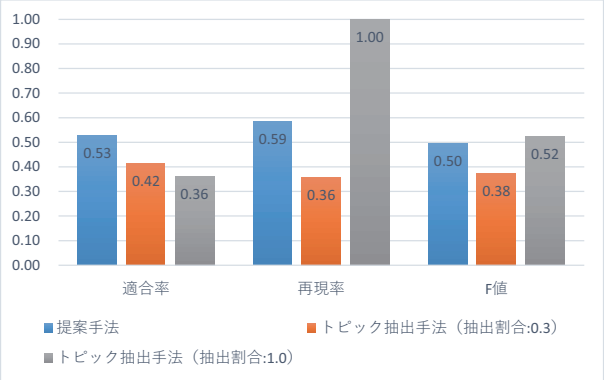

図 10　受動的ユーザを対象とした興味トピック推定の精度

よび F 値について，それぞれ 0.36，1.00，および 0.52 であった．これら手法の適合率を比較すると，提案手法が最も高い結果が得られ，トピック抽出割合 0.3 および 1.0 の比較手法に対し，それぞれ有意水準 5%および 1%で有意差が確認された ($F(1,11) < 7,08, p < .05$，および $F(1,11) < 13.68, p < .01$)．また提案手法とトピック抽出率 0.3 を適用したトピック抽出手法の再現率を比較すると，提案手法が高い結果が得られ，有意水準 10%で有意差が確認された ($F(1,11) < 4.59, p < .10$)．F 値に関しては，提案手法とトピック抽出率 0.3 を適用したトピック抽出手法を比較すると，提案手法が高い結果が得られ，有意水準 5%で有意差が確認された（$F(1,11) < 7.72, p < .05$）．その一方で，提案手法とトピック抽出割合 1.0 を適用した比較手法と比べると，その比較手法が高い結果が得られたものの，有意差が確認されなかった ($F(1,11) < 0.68, p > .10$)．

提案手法は，能動的ユーザを対象とした興味トピック抽出技術を，受動的ユーザを対象とした興味トピック推定技術に転用するアプローチに基づいている．能動的ユーザと受動的ユーザを対象した提案手法の精度は，適合率および F 値の点で同様の傾向があり，その差はそれぞれ 0.05 と 0.02 であった．また提案手法は両方の対象にして場合にも有効であることが確認された．したがって，提案手法は，能動的ユーザの興味トピック抽出を転用することで，受動的ユーザの興味トピックを推定することに有効であることが示唆された．

提案手法をさらに改良するために，今回の実験で評価の低いトピックの理由についてヒアリング調査を行った．その結果，ユーザにとって未知の単語がいくつか含まれているトピックに対しては，ユーザがそのトピックを興味なしと判断している可能性が高いことがわかった．この問題に対処するために，ユーザの興味を推定する前に，ユーザにとって未知の単語を考慮する必要があるが，それについては新たな手法を検討する必要があり，本研究の対象外である．

7　結　　論

ソーシャルネットワーク上のユーザが生成したコンテンツに基づくユーザモデリングの多くの研究は，積極的にツイートを投稿する能動的ユーザを対象としてきた．しかしながら，ツイートを投稿しない受動的なユーザは一定数存在するにも関わらず，受動的ユーザの興味を推測する手掛かりが不足しているため，ほとんど研究開発されてこなかったのが現状である．そこで，本研究では能動的ユーザの興味抽出手法を転用することで，フォローしているユーザのツイートに基づく受動的ユーザの興味推定手法の開発を目的とし，実施した．評価では，12 人のユーザの興味トピック評価データをもとに，単純なトピック抽出手法と比較することで，提案手法の有効性を確認した．

今後の課題としては，提案システムのトピック解析を改善し，応用システムで評価することが挙げられる．提案システムのトピック解析を改善するためには，入力ツイートに含まれる単語をより一般的なものにする必要がある．そうすることで，ユーザの未知の単語を減らすことができ，ユーザに受け入れられやすいトピックモデルが得られる．また，提案手法を情報推薦システムなどの応用システムに組み込むことで，その有効性を検証する．

文　　献

[1] Abel, F., Gao, Q., Houben, G.J., and Tao, K.: Analyzing user modeling on twitter for personalized news recommendations. In Proceedings of the 19th international conference on User modeling, adaption, and personalization (UMAP'11), pp.1–12. Springer (2011).

[2] Besel, C., Schl¨ otterer, J., and Granitzer, M.: Inferring Semantic Interest Profiles from Twitter Followees: Does Twitter Know Better Than Your Friends? In Proceedings of the 31st Annual ACM Symposium on Applied Computing. pp. 1152–1157 (2016).

[3] Blei, D. M., NG, A. Y., and Jordan, M. I.: Latent dirichlet allocation. the Journal of machine Learning research, 3, pp. 993–1022 (2003).

[4] Faralli, S., Stilo, G., and Velardi, P.: Recommendation of microblog users based on hierarchical interest profiles. Social Network Analysis and Mining 5(1), pp.1–23 (2015).

[5] Hamerly, G., and Elkan, C.: Learning the k in k-means, In Proceedings of 16th International Conference on Neural Information Processing Systems (NIPS'03), pp.281–288 (2003).

[6] Kapanipathi, P., Jain, P., Venkataramani, C., and Sheth, A.: User interests identification on twitter using a hierarchical knowledge base, In: Presutti, V., d'Amato, C., Gandon, F., d ' Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 99–113. Springer, Heidelberg (2014).

[7] Orlandi, F., Breslin, J., Passant, A.: Aggregated, interoperable and multi-domain user profiles for the social web. In: Proceedings of the 8th International Conference on Semantic Systems, pp. 41–48 (2012).

[8] Piao, G., and Breslin, J.G.: Exploring Dynamics and Semantics of User Interests for User Modeling on Twitter for Link Recommendations. In Proceedings of 12th International Conference on Semantic Systems, pp.81–88 (2016).

[9] Piao, G.: User Modeling on Twitter with WordNet Synsets and DBpedia Concepts for Personalized Recommendations. In Proceedings of 25th ACM International Conference on Information and Knowledge Management, pp.2057–2060 (2016).

[10] Piao G., and Breslin J.G.: Inferring User Interests for Passive Users on Twitter by Leveraging Followee Biographies. In: Jose J. et al. (eds) Advances in Information Retrieval. In Proceedings of 39th European Conference on Information Retrieval; Lecture Notes in Computer Science, vol 10193, Springer, pp.122–133 (2017).

[11] Sheth, A., and Kapanipathi, P.: Semantic Filtering for Social Data. IEEE Internet Computing 20(4), pp.74–78 (2016).

[12] Siehndel, P., and Kawase, R., TwikiMe!: user profiles that make sense, In Proceedings of the 2012th International Conference on Posters & Demonstrations Track-Volume 914. pp.61–64 (2012).

[13] Teh, Y. W. and Jordan, M. I., Hierarchical bayesian nonparametric models with applications, Bayesian Nonparametrics (eds. N. L. Hjort, C. Holmes, P. Muller and S. G. Walker), pp.158–207 (2010).