

メンション情報を利用した Twitter ユーザプロフィール推定

奥谷 貴志^{†1} 山名 早人^{†2†3}

^{†1} 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1

^{†2} 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

^{†3} 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: {okutani, yamana}@yama.info.waseda.ac.jp

あらまし 日々様々な情報発信が行われている Twitter では、有用な情報を発信しているユーザを発見したり、マーケティングに利用したりするために、Twitter ユーザの所属や趣味といったプロフィールを知ることが重要である。従来のプロフィール推定では、ユーザ間のフォロー情報によって構築される交友関係のグラフからプロフィール推定を行っているが、1) 友人・有名人・ニュースのような混在したフォロー目的を区別できない、2) 一度フォローしたユーザに対して現在も交流があるかなど最新の交友関係を判断しにくい、という問題があった。こうした問題を解決するため、本研究ではフォローに代えてユーザ間でやり取りされるメンションを解析することで、Twitter ユーザのプロフィールを最新の交友関係に基づき詳細に推定する手法を提案する。実験の結果、フォロー情報を用いたプロフィール推定と比較して Precision@10 が 48.6% から 58.6%、MRR が 1.55 から 1.86 に向上した。

キーワード Twitter, ソーシャルグラフ, プロフィール推定

1. はじめに

近年、様々なソーシャルネットワークサービス (SNS) 上で情報発信、情報共有が行われている。代表的な SNS である Twitter[1] は、ツイートと呼ばれる 140 字以内の文章を投稿できるマイクロブログと呼ばれる種類の SNS である。Twitter では、ツイート投稿による情報発信だけでなく、フォローと呼ばれる機能によって他のユーザのツイートを購読したり、メンションと呼ばれる機能によってユーザ間でツイートをやりとりしたりすることによって、ユーザが興味を持つ情報の収集、共有を可能としている。Twitter は 5 億人以上の登録ユーザを持ち[2]、2013 年 9 月時点における月間アクティブユーザ数も 2 億人以上[3]であり、非常に活発に利用されている SNS である。そのため、Twitter ユーザの所属、興味といったプロフィールを知ることができれば、ユーザ推薦のような Twitter 自体の利便性向上だけでなく、企業製品の購買層を知るためのマーケティングや政党の支持層を知るための世論調査、ユーザにマッチした広告を提供するためのターゲティングといった SNS 以外の企業活動、インターネット選挙にも活用することができると考えられる。しかし、Twitter に用意されているプロフィール欄を詳細に記述しているユーザは少ないため、プロフィールテキスト以外にツイートやフォローのような情報を利用してユーザ分類やプロフィール推定を行う必要がある。

現在、Twitter 上の膨大なデータから有用な情報を得るための様々な研究が行われている。その中で、Twitter ユーザのプロフィール推定や、ユーザ分類を行う研究として、大きく分けて 2 つのアプローチが存在している。

1 つは、主にユーザのツイート情報を解析することでプロフィールを推定する研究である。しかし、Twitter ユーザはツイート毎に様々なトピックに関して言及しているため、中心となるトピックに関連したプロフィールは埋もれてしまいやすい。そのため、ユーザの位置情報の推定[4][5]や、ユーザの政治的指向、民族の推定[6]を行う研究のように、特定の属性に関する推定を行う研究が中心となっており、あるユーザのプロフィール全体を推定することは難しいという問題がある。

もう 1 つは、ユーザ間の繋がりによって構成されるソーシャルグラフを解析することでユーザ分類を行う研究である。Twitter ユーザ間のフォロー関係によってソーシャルグラフを構築しクラスタリングを行った研究[7]では、抽出されたコミュニティに対して属性となるキーワードを付与することでプロフィールの推定を行っている。しかし、単純にクラスタリングを行っただけでは 1 人のユーザが 1 つのコミュニティに所属しているものとして抽出されてしまい、複数の属性を持つプロフィールに対応できないという問題がある。また、フォローを利用したソーシャルグラフでは、ユーザ間の繋がりがフォローの有無だけで表現されるため、以前フォロー関係を構築したが現在では交流の少ないユーザと、現在でも活発に交流している親密なユーザを区別できないという問題がある。

そこで本稿では、ユーザ間でやりとりされるメンション情報を利用して Twitter ユーザのプロフィールを推定する手法を提案する。プロフィール推定対象ユーザ周囲に存在する複数のコミュニティから属性タグを抽出することで、ツイートからプロフィール推定を行う研究[4][5][6]において特定の属性に関するプロフィ

ールしか推定できないという問題を解決する。また、メンション情報を利用してソーシャルグラフを構築することで、メンション頻度やメンションの持つ時系列情報を反映させ、フォローから交友関係を推定する研究[7]において最新の交友関係を判断できないという問題を解決する。

提案手法では、まずプロフィールを推定したいユーザを起点ユーザとして設定し、起点ユーザとメンションによる繋がりを持つユーザ集団によるソーシャルグラフを構築する。次に、構築したソーシャルグラフについて、ユーザ間のメンション傾向によってクラスターリングを行う。最後に、各クラスター内において特徴的な単語を属性タグとして抽出し、この属性タグの集合を起点ユーザのプロフィールとして出力する。各クラスターから複数の属性タグを抽出することで、所属や興味といった複数の属性を持つプロフィールの推定を実現した。また、時系列情報を持つメンションを利用し、頻度による重み付けを行ったソーシャルグラフを構築することで、最新の交友関係やユーザ間の親密さに基づくプロフィール推定を実現した。

本論文では以下の構成をとる。まず 2 節で関連研究を紹介し、次に 3 節で提案手法について説明する。4 節で実験と評価を行い、最後に 5 節でまとめを述べる。

2. 関連研究

2.1. ツイート情報とグラフによるプロフィール推定手法

Pennacchiotti らは、機械学習とソーシャルグラフの組み合わせによって Twitter ユーザのプロフィールを推定する手法を提案した[6]。まず、対象ユーザのツイート傾向、ツイート本文、フォロー関係に対してブースティングの一種である Gradient Boosted Decision Trees (GBDT) を用いた分類を行った。ここでは、ツイート傾向としてユーザの総ツイート数と、リツイート、メンション、ハッシュタグ、URL の割合を利用し、ツイート本文の情報としてツイート中の単語、ハッシュタグを利用し、フォロー関係としてフレンド（誰をフォローしているか）を利用している。次に、GBDT によって出力されたスコアに対して、対象ユーザのフレンドのスコアを利用したスコア更新を行った。ここでは、対象ユーザとフレンド間でのメンションとフォロー密度によって重み付けを行っている。提案手法では、機械学習で出力されたスコアとソーシャルグラフによって更新されたスコアの平均スコアを最終的なスコアとして出力している。

実験では、政治的指向（民主党支持か共和党支持か）、民族（アフリカ系アメリカ人かそうでないか）、特定のビジネスへの親近感（スターバックスファンかそうでないか）について分類を行っている。提案手法と、機

械学習のみを用いた場合、ソーシャルグラフによるスコアのみを利用した場合それぞれの分類結果について評価を行った結果、政治的指向、スターバックスファンの判定において提案手法の F 値が最も高く、民族の判定でも機械学習に次ぐ 2 番目に高い F 値になっており、それぞれの属性におけるユーザプロフィールをうまく分類できていることがわかる。しかし、この手法はあらかじめ特定の属性についてのみ分類を行うものであり、またある属性を持っているかの二値分類を行うものとなっているため、属性の種類を特定せず対象ユーザがどのような属性をプロフィールに含んでいるかを知る、という目的に利用することができないという問題がある。

2.2. フォロー関係を用いたユーザコミュニティ抽出

Java らは、Twitter ユーザの地理的な分布や、利用形態、ユーザ間のネットワーク構造に関する調査を行い、その中で Twitter ユーザ間のフォロー関係を利用したコミュニティ抽出を行った[7]。まず、Twitter ユーザをノード、Twitter ユーザのフォロー関係をエッジとしてソーシャルグラフを構築している。ここでは、友人関係によって構築されるコミュニティを抽出するために、相互フォロー関係のみをエッジとした単純無向グラフとしている。次に、構築されたグラフに対して、フォロー関係の密度が小さい部分を発見するために、Girvan-Newman アルゴリズムを利用してグラフの分割を行った。密度が低い部分のエッジを取り除いていくことで Twitter ユーザから構成されるグラフを分割し、複数のコミュニティを抽出した。

実験では、87,897 人の Twitter ユーザに対してソーシャルグラフの分割を行った。抽出されたコミュニティのうち、58 人のユーザによって構成されるユーザコミュニティについて、所属するユーザのツイートから単語を抽出し、出現回数の多い単語を Key Terms として出力している。「xbox」、「game」、「halo」といった Key Terms により、コミュニティに所属しているユーザはゲームに関連したプロフィールを持っていると考えることができる。しかし、1 人のユーザが 1 つのコミュニティに所属しているものとして抽出されているため、コミュニティ毎の Key Terms をそのままユーザのプロフィールとして考えた場合、複数のコミュニティに所属し複数の属性を持っているユーザのプロフィールの一部分しか推定することができないという問題がある。また、フォロー関係によって構築したソーシャルグラフを利用しているため、フォローの有無という重み無しの情報のみでユーザ間の繋がりが表現されるため、以前フォロー関係を構築したが現在では交流の少ないユーザと、現在でも活発に交流している親密なユーザといったユーザの親密度を区別できないという問題が

ある。

2.3. Girvan-Newman アルゴリズム

Newman らは、トップダウンアプローチによる階層型クラスタリング手法として Girvan-Newman アルゴリズムを提案した[8]。トップダウンアプローチでは、ノード全体によって構成されるコミュニティを起点とし、コミュニティを分割していくことでクラスタリングを実現する。一方で、ボトムアップアプローチでは1つのノードを1つのコミュニティとし、個々のコミュニティを結合して1つのコミュニティにしていくことでクラスタリングを実現する。

Newman らは、コミュニティ内のエッジ密度が高く、コミュニティ間のエッジ密度が低いクラスタリングが良いクラスタリングとし、クラスタリング結果を評価する指標として式(1)に示す modularity Q を定義した。

$$Q = \sum_i (e_{ii} - a_i^2) \quad (1)$$

ここで、 e_{ij} は総エッジ数に占めるコミュニティ i からコミュニティ j へのエッジ数の割合、 $\sum_j e_{ij} = a_i$ である。Newman らは、各エッジの shortest-path-betweenness を計算し、最も高いスコアを持つエッジを切断していくことでクラスタリングを行った。クラスタリング結果に対し、 Q が最も高くなるコミュニティ数によってグラフを分割し出力する。

Girvan-Newman アルゴリズムは、メンション関係によるソーシャルグラフのような多重有向グラフにも適用することができ、また modularity によって最適なコミュニティ数を自動的に選択することができる。しかし、エッジ切断の度に shortest-path-betweenness を残存する全エッジについて計算する必要があり、ノード数を n 、エッジ数を m としたときの最悪計算量が $O(m^2n)$ になってしまい、大規模グラフに適用するのが難しいという問題がある。

2.4. Newman アルゴリズム

Newman は、Girvan-Newman アルゴリズムにおける計算量問題を解決するボトムアップ型クラスタリング手法として Newman アルゴリズムを提案した[9]。Girvan-Newman アルゴリズムでは、shortest-path-betweenness に基づいたエッジ切断を行うことで modularity が高くなるコミュニティ抽出ができていたが、Newman は、modularity を高くするという目的に着目し、modularity が最も増加するようなコミュニティ結合の組み合わせを選んで結合していくことを提案した。ある時点での modularity Q と、コミュニティ i, j を結合した後の modularity Q' の差分 ΔQ は以下に示す式(2)で表すことができる。

$$\Delta Q = e_{ij} + e_{ji} - (2a_i a_j) \quad (2)$$

Newman アルゴリズムでは、個々のノード1つのみからなるコミュニティを起点とし、次に ΔQ の最も高いコミュニティの組み合わせを結合する。コミュニティ数が1となるまで ΔQ の再計算とコミュニティ結合を繰り返し、最終的に Q が最も高くなるコミュニティ数によってグラフを分割し出力する。

Newman アルゴリズムでは、結合したコミュニティの周囲のエッジのみについて ΔQ を再計算すればよい。ため、最悪計算量が $O((m+n)n)$ となっており、Girvan-Newman アルゴリズムと比較すると大規模なグラフに対しても実用的なアルゴリズムとなっている。本研究では、Newman アルゴリズムを用いたコミュニティ抽出を行う。

3. 提案手法

本節では、提案手法であるメンション情報を利用した Twitter ユーザプロフィール推定について述べる。従来のツイート情報を用いてプロフィールを推定する研究[6]では、特定の属性についてのプロフィールしか推定できないという問題が存在したが、提案手法ではプロフィール推定対象ユーザの周辺ユーザ情報から複数の属性を抽出することで解決する。また、フォロー関係を利用してユーザ分類を行った研究[7]では、ユーザの親密度を区別したクラスタリングを行うことができないという問題が存在したが、提案手法ではメンションによってユーザ間の繋がりの強さを反映したソーシャルグラフを構築することで解決する。

3.1. メンション情報によるグラフ

本項では、本研究で利用するメンション情報について説明する。本文中に「@ユーザ名」を含むツイートはメンションツイートと呼ばれ、「ユーザ名」で示されるユーザのタイムラインに表示される。そのため、メンションツイートはユーザ同士の会話において用いられる。また、リツイートを広義のメンションツイートとして扱う場合があるが、本研究ではリツイートはメンションとして扱わず、メンションツイートのみをメンションとして扱う。

メンションは、ユーザからユーザへのツイートであるため、フォロー関係と同様にユーザ間の繋がりを示す情報として利用することができる。フォロー関係と比較した場合、メンションには以下に示す3つの特徴が存在する。

1. 個人対個人の関係が中心となっている
 2. メンション頻度によって親密度を表現できる
 3. 1つ1つのメンションが時系列情報を持つ
- 1つ目については、友人関係、ニュースアカウント、

有名人といった複数の目的を持つフォロー関係と異なり、メンションはユーザ間の会話であるため、友人のような個人対個人の関係を中心とした絞ったソーシャルグラフを構築することができる。2つ目については、メンションはフォローの有無のような単一の繋がりではないため、頻繁にメンションをやりとりするユーザ同士はより親密であると考えることができる。3つ目については、メンションはツイートそのものであるため、投稿時間に基づいて直近のメンションを利用することで、最新の交友関係に基づいたソーシャルグラフを構築することができる。

3.2. 提案手法の概要

提案手法では、プロフィール推定の対象となるユーザがどのようなユーザと交友関係を持っているか、という情報を元にプロフィール推定を行う。対象ユーザではなく周辺ユーザのプロフィールテキストを用いることで、対象ユーザがプロフィールテキストを詳細に記述していない場合に対応する。また、周辺ユーザによって構成される複数のコミュニティから属性を抽出することで、複数の属性を含むプロフィールに対応する。

提案手法の流れを示す。まず、プロフィール推定対象のユーザを起点ユーザとして設定する。次に、起点ユーザとメンションによる繋がりを持つユーザ集団を1次ユーザとして抽出する。同様に、1次ユーザと繋がりを持つユーザ集団を2次ユーザとして抽出する。起点ユーザ・1次ユーザ・2次ユーザで構成されたソーシャルグラフに対し、Newman アルゴリズムによってクラスタリングを行い、コミュニティを抽出する。次に、抽出された各コミュニティに対して、所属するユーザのプロフィールテキストから形態素解析によって単語を抽出し、スコアの高い単語を属性タグとして選択する。最終的に、スコア上位の属性タグ集合をプロフィール推定対象ユーザのプロフィールとして出力する。

3.3. クラスタリングによるコミュニティ抽出

本研究では、メンションによるユーザ間の繋がりを表すために、ユーザ i からユーザ j へのメンション数 m_{ij} を用いて、グラフを表現する隣接行列 A の要素 A_{ij} を以下の式(3)のように定義した。

$$A_{ij} = \sqrt{m_{ij} \times m_{ji}} \quad (3)$$

式(3)より、 A_{ij} は m_{ij} と m_{ji} の幾何平均であるため、相互メンション関係を持つユーザ間でのエッジのみを扱うことになる。このため、お互いにメンションを頻繁にやりとりしているユーザ同士ほど強く重み付けされている。また、 $A_{ij} = A_{ji}$ より、無向グラフとなっている。

式(3)を利用し、ソーシャルグラフの構築とクラスタリングを以下の手順によって行う。

1. プロフィール推定対象ユーザを起点ユーザ u_0 として設定する
2. u_0 との間に式(3)で表されるエッジを持つユーザ N 人の集団を、1次ユーザ群 $U_1 = \{u_{11}, u_{12}, \dots, u_{1N}\}$ として抽出する
3. 同様に、 U_1 に含まれるユーザとの間に2本以上エッジを持つ、 u_0 , U_1 以外のユーザ M 人の集団を、2次ユーザ群 $U_2 = \{u_{21}, u_{22}, \dots, u_{2M}\}$ として抽出する
4. u_0 , U_1 , U_2 によって構成されるソーシャルグラフに対して、Newman アルゴリズムによるクラスタリングを行う
5. クラスタリング結果に対し、modularity Q が最も高くなるコミュニティ数によってグラフを分割し出力する

手順3.では、 U_1 との間に2本以上エッジを持つユーザのみを U_2 として抽出している。これは、 U_1 のうち1人としか繋がりを持たないユーザは、起点ユーザと u_0 のプロフィールとの関連性が低いユーザだと考えたためである。以上の手順によって出力されたコミュニティの集合を、起点ユーザ周囲のTwitter ユーザコミュニティ群とする。

3.4. 属性タグの選択

3.3項で抽出されたユーザコミュニティ群に対して、属性タグの抽出を行う。本研究では、各コミュニティに所属するユーザのプロフィールテキストに対し、形態素解析エンジンである MeCab[10]を利用して名詞抽出を行った。また、MeCab で利用される単語辞書について、最新の固有名詞や流行語、略語に対応するため、はてなキーワード[11]と Wikipedia の項目名[12]による拡張を行っている。

次に、各コミュニティ c から抽出された単語 w について、本研究では式(4)に示す TF-IDF によるスコア $S(w, c)$ を付与した。

$$\begin{aligned} tf_{wc} &= \frac{n_{wc}}{\sum_k n_{kc}} \\ idf_{wc} &= \log \frac{D}{d_w} \\ S(w, c) &= tf_{wc} \cdot idf_{wc} \end{aligned} \quad (4)$$

ここで、 n_{wc} はコミュニティ c 中の単語 w の出現回数、 D は削減後の総コミュニティ数、 d_w は単語 w を含むコミュニティ数である。TF-IDF によって、どのコミュニティでも頻出するコミュニティと関連性の低い単語のスコアを低くし、各コミュニティ特有の単語のスコアを高くしている。

しかし、3.3項の手順で抽出されたコミュニティには、 u_0 と1次の繋がりを持たない、 U_2 に所属するユー

ザのみからなるコミュニティが存在する．このようなコミュニティは起点ユーザ u_0 との関連性が低いと考えられるので，式(4)で示される $S(w, c)$ は計算せず，属性タグを付与しない．同様に，極端にユーザ数の少ないコミュニティは，起点ユーザ u_0 との関連性が低いと考えられる一方で， $\sum_k n_{kc}$ が小さくなるため単語のスコアが高くなりやすいという問題がある．そのため，閾値 cth を設定し，所属ユーザ数がソーシャルグラフ全体のユーザ数の $cth\%$ を下回るコミュニティは $S(w, c)$ を計算しない．また，起点ユーザ u_0 が持つプロフィールテキストの詳細さに依存しないプロフィール推定を行うため，本研究では u_0 のプロフィールテキストは除外してスコア $S(w, c)$ の算出を行っている．

また，式(4)の計算において， $n_{wc} = 1$ となるような単語，つまりコミュニティ中に一度しか出現していない単語はコミュニティとの関連性が低く，属性タグとしては不適当だと考えられる．一方で，除外されなかったユーザ数が $cth\%$ を超えるコミュニティにおいても，短いプロフィールテキストを持つユーザが多数である場合は $\sum_k n_{kc}$ が小さくなるため $n_{wc} = 1$ でも tf_{wc} が高くなり属性タグとして選択される可能性がある．そのため， $S(w, c)$ の計算条件に合致したコミュニティであっても， $n_{wc} = 1$ となる単語は属性タグの候補から除外している．

予備実験の結果，本研究では $cth = 5\%$ とした．また，起点ユーザ u_0 が持つプロフィールテキストの詳細さに依存しないプロフィール推定を行うため，本研究では u_0 のプロフィールテキストは除外してスコア $S(w, c)$ の算出を行っている．

最後に，スコア上位 10 件の単語を属性タグとして抽出し，この属性タグ集合を起点ユーザのプロフィールとして出力する．

4. 実験・評価

4.1. 使用データ

本研究では，メンションによって構築されたソーシャルグラフを利用する．しかし，ソーシャルグラフ構築のためにデータを収集する際，片方向のメンションしか一度に取得できないという問題がある．これは，フォロワーとフレンドを同時に取得できるフォロー関係と異なり，メンション情報はユーザのツイートから取得するため，収集対象のユーザから他のユーザへのメンションしか取得できないためである．そのため，対象となるユーザへのメンションを収集することが難しい．

この問題を解決するため，本研究では Twitter API[13]を利用し，事前に Twitter ユーザのツイートを網羅的に収集するクローラを開発した．クローラでは，

Twitter Streaming API[14]によって取得した日本語設定の Twitter ユーザについて，ユーザ 1 人あたり最大 2,000 ツイートを収集している．Streaming API を利用することで，最近ツイートを投稿したユーザを優先的に収集することが可能になっている．

実験では，2013 年 1 月 1 日から 2013 年 12 月 31 日にかけてツイートを投稿した，日本語設定ユーザ 7,955,714 アカウントのツイートを利用した．収集した総ツイート数は 8,756,608,942，総メンションツイート数は 3,257,674,537 であった．メンションツイート中に現れるユーザをメンションユーザと定義し，ユーザあたりのメンションツイート数，メンションユーザ数，フォロー数の平均値・中央値を表 1 に示す．

表 1 ユーザあたりのメンションツイート数・メンションユーザ数・フレンド数の平均値・中央値

	平均値	中央値
メンションツイート	409	260
メンションユーザ	102	62
フォロー	225	110

表 1 より，フォロー数と比較した場合，メンションユーザ数は少なくなっているが，メンションツイート数は多くなっていることがわかる．このことより，同一ユーザへの複数回のメンションツイートを重みとして考えることで，グラフ構築における情報量を増加させられることがわかる．また，各ユーザの総メンションユーザのうち，データセット中に含まれている割合を調査した結果，平均で 77.8%のユーザ ID をカバーできていることがわかった．

4.2. ユーザコミュニティの抽出

実験では，提案手法と比較するため，グラフを表現する隣接行列 A について以下の 3 手法を利用したプロフィール推定も行った．

- ① 相互フォローによる単純無向グラフ
- ② 提案手法に対してエッジ重みを無くした単純無向グラフ
- ③ メンション数をそのままエッジの重みとした多重有向グラフ

手法①では，従来のフォロー関係によるソーシャルグラフと比較するため，[7]の研究で利用される相互フォロー関係によってグラフを構築した．手法②は，提案手法と同様のリンク関係においてフォローのようにエッジ重みが無い場合と，比較手法①と同様の単純無向グラフをメンションで表現した場合の結果を比較するためのものである．手法③は，ユーザ間のメンション関係をそのままグラフとした場合について比較するためのものである．

提案手法に以上 3 手法を加えた 4 手法に対して、理系大学生・大学院生 7 名の Twitter ユーザに対するプロフィール推定を行った。また、グラフ構築の際に 2 次ユーザを利用する有用性を評価するため、3.3 項の手順 3. を行った場合と行わず 1 次ユーザのみでグラフを構築した場合におけるプロフィール推定を行った。クラスタリング結果についての要約を表 2、表 3 に、modularity を図 1、図 2 に示す。

表 2 1 次ユーザのみにおけるクラスタリング結果

手法	平均ユーザ数	総エッジ重みの平均	平均クラスタ数
提案手法	58	3,610	11
手法①	97	987	7
手法②	58	259	9
手法③	138	5,316	10

表 3 2 次ユーザを含めたクラスタリング結果

手法	平均ユーザ数	総エッジ重みの平均	平均クラスタ数
提案手法	623	56,083	17
手法①	3,522	191,985	8
手法②	623	6,277	11
手法③	12,481	2,282,442	105

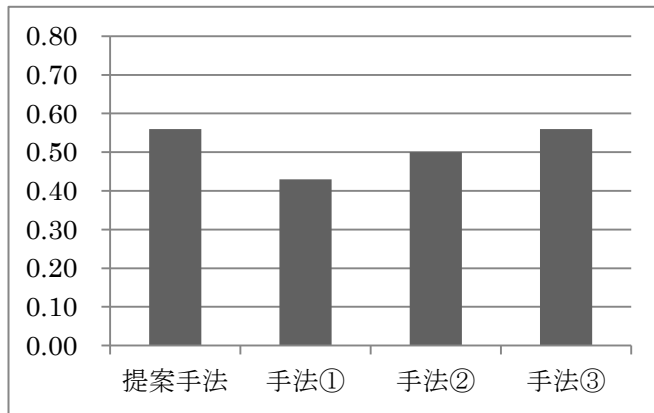


図 1 1 次ユーザのみにおける modularity

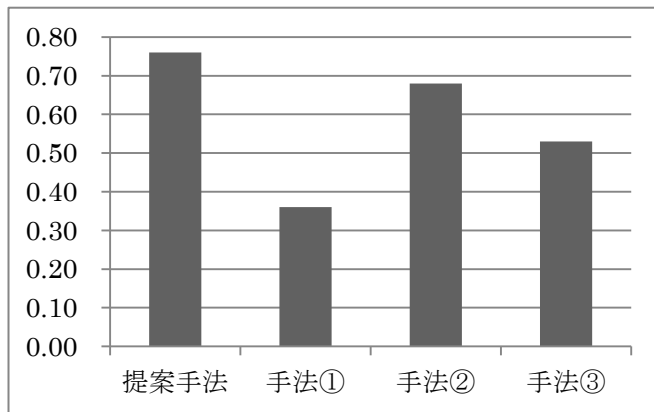


図 2 2 次ユーザを含めた modularity

表 2、表 3 より、手法②のメンションによる単純無向グラフでは、手法①のフォロー関係によるグラフよりもエッジが少なく、クラスタリングに利用できる情報が少ないことがわかる。一方で、エッジに重みを付与することで、手法③の多重有向グラフでは手法①以上の総エッジ重みとなっていることがわかる。表 2 では提案手法のエッジ重みが手法①より多く、表 3 では少ないが、これは提案手法の平均ユーザ数が手法①より少ないため、エッジ密度においては提案手法と手法①は同程度となっている。また、図 1、図 2 より、提案手法はフォローを用いた手法①よりも modularity が高く、特に 2 次ユーザを含めたグラフにおいては 0.7 を超える高い modularity となっている。結果的に、相互フォローを利用した手法①と比較すると modularity は 0.43 から 0.76 まで向上した。以上より、modularity に着目した場合、相互メンション関係を利用し 2 次ユーザを含めたときに質の高いグラフが構築できることがわかる。

4.3. 属性タグ選択とプロフィール推定

次に、被験者の Twitter ユーザ 7 名について、提案手法と手法①、②、③で出力されたプロフィール推定結果の評価を行った。まず、各手法で出力された全属性タグを集めた属性タグリストを被験者ごとに作成した。次に、この属性タグリストから被験者自身によってプロフィールに関連している単語を抜き出してもらい、正解セットとした。そして、各実験での出力結果について、抽出された全属性タグについてプロフィール推定対象ユーザ自身による判定を行い、自身のプロフィールに関連していると考えられる属性タグ集合を正解セットとして、正解セットに対する正答率 (Precision@10: 上位 10 件の属性タグ集合中の正答率) を調べた。また、正解となる属性タグをどれくらい高い順位で正解できていたかを評価するため、正解セットに含まれる属性タグ k の順位を $rank_k$ として、式(5)によって推定結果の MRR (平均逆順位) を計算した。

$$MRR = \sum_k \frac{1}{rank_k} \quad (5)$$

式(5)より、同じ正答率の場合より高い順位で属性タグを出力するほど MRR が高くなる。

まず、提案手法で抽出された属性タグの例を表 4、表 5 に示す。

表 4 属性タグの例 1 (太字が正解)

順位	属性タグ (2 次ユーザ含む)	属性タグ (1 次ユーザのみ)
1	情報工学	学園
2	モータースポーツ	大学院生
3	パーツ	劇団
4	早稲田大学	初心者
5	アキバ	情報工学
6	Twitter	必死
7	理工	修士
8	大学院生	女子
9	情報処理	チョコレート
10	事業	修羅

表 5 属性タグの例 2 (太字が正解)

順位	属性タグ (提案手法)	属性タグ (手法①)
1	アイマス	真美
2	理工	グリマス
3	MIS	ゆるキャラ
4	DJ	グル
5	ラブライブ!	美希
6	早稲田大学	今井麻美
7	Python	春香
8	アニソン	四条貴音
9	基幹	木蓮
10	Android	あずさ

表 4, 表 5 では, 正解と判定された属性タグを太字で示している. 表 4 では, 「大学院生」「情報工学」「修士」というユーザの所属に関する属性タグがプロフィールとして抽出できている. しかし, 理工系の大学院生という単一の属性に関する属性タグしか抽出できていない. 一方で, 1 次ユーザに加えて 2 次ユーザも用いた場合には「モータースポーツ」「アキバ」といった所属以外の趣味, 興味に関連した属性タグも抽出できている. また, 表 5 では, 相互フォローを用いた手法①では「真美」「美希」のような 1 つのゲームに登場するキャラクター名ばかりが属性タグとして抽出されているが, 提案手法では「アイマス」のように当該ゲーム名に加えて, 「早稲田大学」「Python」のような所属や興味に関連した属性タグも抽出できている.

このことから, 2 次ユーザまで拡張してグラフを構築することで, フォロー情報を利用した場合と比較したとき, プロフィール対象ユーザが持つ複数の属性に関連した属性タグの抽出ができていたことがわかった.

次に, 属性タグの Precision@10 と MRR の平均を比較した結果を図 3, 図 4 に示す.

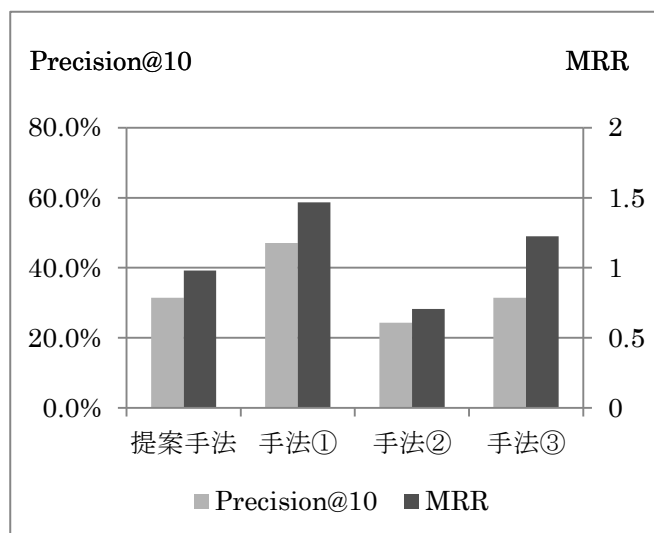


図 3 1 次ユーザのみにおける Precision@10・MRR

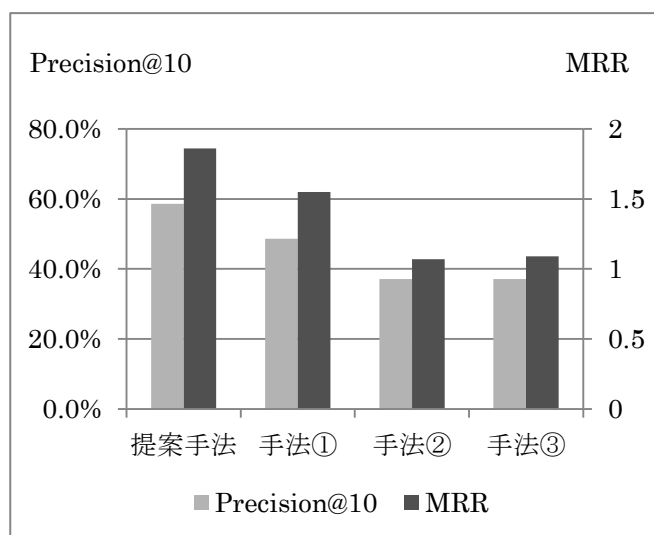


図 4 2 次ユーザを含めた場合の Precision@10・MRR

図 3 では, 1 次ユーザのみを用いた場合においては相互フォローを利用した手法①がもっとも良い結果となっている. また, 多重有向グラフを用いた手法③は提案手法より高い MRR となっており, 一方で単純無向グラフである手法②の Precision@10 と MRR がもっとも低い. 表 2 における平均ユーザ数も考慮すると, 提案手法や手法②は 1 次ユーザのみを用いた場合にはグラフ構築に利用できるユーザ情報が少なくなるために, 表 4 のように複数の属性に対応した属性タグを抽出することができず Precision@10 や MRR が低下していると考えられる.

一方, 2 次ユーザを含めてグラフを構築した図 4 では, 提案手法の Precision@10 と MRR がそれぞれ 58.6%, 1.86 となっており, それぞれ 48.6%, 1.55 だった手法①を含め手法①, ②, ③のどれよりも高くなっている.

また、手法②と手法③の差がほとんど無くなっている。これは、2 次ユーザを含めることで表 4 で示されているように提案手法と手法②で利用できるユーザ情報が増加し、複数の属性タグを得られた結果として正答率向上に繋がったものと考えられる。

また、図 3、図 4 と図 1、図 2 の modularity を比較すると、フォローを利用した手法①の modularity はメンションを用いた提案手法、手法②、③より一貫して低いものの、Precision@10 と MRR は手法②、③より一貫して高く、高い modularity が必ずしも高い正答率に結びついていないことがわかった。

最終的に、全体では図 4 における提案手法が Precision@10 と MRR ともにもっとも高いものとなっており、相互フォローを利用した手法①と比較して Precision@10 が 48.6%から 58.6%に、MRR が 1.55 から 1.86 に向上しており、フォロー関係に替えてメンションを利用することの有効性が確認できた。

5. まとめ

本論文では、メンション情報を利用した Twitter ユーザのプロフィール推定手法を提案した。従来のツイートを利用したプロフィール推定や、フォロー関係を利用したクラスタリングによるユーザ分類では任意・複数の属性にユーザを対応させられないという問題があった。提案手法では、プロフィール推定対象ユーザの周辺ユーザのクラスタリングを行うことで対象ユーザの所属や興味といった複数の属性に対応したプロフィールを推定することができた。また、周辺ユーザによるグラフの構築にメンション情報を利用することで、最新の交友関係に基づくユーザ間の親密さを反映させた。実験の結果、従来のフォローを用いた場合のプロフィール推定結果に対し、Precision@10 が 48.6%から 58.6%、MRR が 1.55 から 1.86 に向上した。また、modularity もフォロー関係を用いた場合と比較し 0.43 から 0.76 まで向上し、質の良いソーシャルグラフを構築することができた。

一方で、高い modularity を持つグラフであっても高い正答率となるような属性タグを出力するとは限らず、質の高いクラスタリングを属性タグ抽出によるプロフィール推定にうまく反映しきれていないという問題もあった。そのため、より精度の高い属性タグの抽出を行うことが今後の課題となった。

参 考 文 献

- [1] Twitter, <http://twitter.com/> (2014 年 1 月 6 日アクセス)
- [2] SemioCast - Twitter reaches half a billion accounts - More than 140 millions in the U.S., http://semioCast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US (2014 年 1 月 6 日アクセス)
- [3] Amendment No. 4 to Form S-1, <http://www.sec.gov/Archives/edgar/data/1418091/000119312513424260/d564001ds1a.htm> (2013 年 12 月 12 日アクセス)
- [4] Z. Cheng, J. Caverlee and K. Lee: "You are where you tweet: a content-based approach to geo-locating twitter users", Proceedings of the CIKM'10, pp.759-768, 2010.
- [5] J. Eisenstein, B. O'Connor, N. A. Smith, E. P. Xing: "A latent variable model for geographic lexical variation", Proceedings of the EMNLP'10, pp.1277-1287, 2010.
- [6] M. Pennacchiotti and A. M. Popescu: "Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter", Proceedings of the KDD'11, pp.430-438, 2011.
- [7] A. Java, X. Song, T. Finin and B. Tseng: "Why We Twitter: Understanding Microblogging Usage and Communities", Proceedings of the WebKDD, pp.56-65, 2007.
- [8] M. E. J. Newman and M. Girvan: "Finding and evaluating community structure in networks", Physical Review E, 69(2):026113, 2004.
- [9] M. E. J. Newman: "Fast algorithm for detecting community structure in networks", Physical Review E, 69(6):066133, 2004.
- [10] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html> (2014 年 1 月 6 日アクセス)
- [11] はてなキーワード一覧ファイル - Hatena Developer Center, <http://developer.hatena.ne.jp/ja/documents/keyword/misc/catalog> (2014 年 1 月 6 日アクセス)
- [12] Index of /jawiki/latest/, <http://dumps.wikimedia.org/jawiki/latest/> (2014 年 1 月 6 日アクセス)
- [13] Documentation | Twitter Developers, <https://dev.twitter.com/docs> (2014 年 1 月 6 日アクセス)
- [14] Streaming API | Twitter Developers, <https://dev.twitter.com/docs/streaming-api> (2014 年 1 月 6 日アクセス)