

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おける 単語重要度算出手法の考察

上里和也^{†1} 田中正浩^{†1} 浅井洋樹^{†1†2} 山名早人^{†3†4}

Twitter のような大規模なソーシャルサービスにおいて、ユーザの興味や所属などのプロフィールを知ることは、効果的なマーケティングを行う上で重要である。このような背景から、Twitter におけるプロフィール推定に関する研究が行われてきた。従来のプロフィール推定手法では、フォロー情報によって構築されるソーシャルグラフからコミュニティを抽出し、対象のユーザが属するコミュニティの属性を推定することでプロフィール推定を行なっている。しかし、各々のフォローの目的や、活発な交流があるかという点を考慮することができないため、実際に親密な関係を持つユーザ群をコミュニティとして抽出することが困難であるという問題が存在する。それに対して奥谷らは、フォローに代えてメンション情報を用いてソーシャルグラフを構築することで、これらの問題を解決する手法を提案している。しかし同手法には、プロフィール推定の対象となるユーザの周辺ユーザのプロフィールに幅広く共通して出現する単語が、プロフィールとして出力されにくいという問題がある。そこで本論文では、奥谷らのプロフィール推定手法における単語の重要度の算出方法を変更し、Twitter ユーザ全体からランダムにサンプリングした 100,000 ユーザのデータを利用して一般語をフィルタリングすることで、この問題を解決する手法を提案する。6 人の被験者による実験の結果、奥谷らの手法と比較して、Precision@10 が 0.37 から 0.78、MRR が 1.44 から 2.61 に向上した。

1. はじめに

Twitter^{a)}は、5 億人以上の登録ユーザを持つ^{b)}大規模なソーシャルネットワークサービスである。Twitter のような大規模なソーシャルネットワークサービスにおけるユーザの興味や所属などに代表されるユーザ属性は、ユーザの興味に合致する広告のターゲティングや、製品の購買層調査などの分野で有用である。Twitter ではプロフィールをユーザ自身が登録できるが、近年個人を特定し得る情報の公開に対する懸念が高まっており、ランダムにサンプリングした 1,000 ユーザから、「所属」、「趣味」、「出身地」などの複数の種類の属性を同プロフィール文に記述するユーザを手作業によって抽出した結果、これに該当したのは 23.0%のユーザのみであった。さらに、プロフィール文に自分の属性をいっさい記述しないユーザが 30.5%存在していた。このような状況の中、本人のプロフィール文を用いることなく当該ユーザの興味等の属性を推定する手法の重要性が高まっている。

これまでもユーザのプロフィールを推定する手法に関する研究が行われてきた。これらの既存手法は、利用する情報の種類によって、2 つに大別される。

1 つは、プロフィール推定の対象となるユーザが発信するツイート自体を用いる手法[1][2][3][4][5]である。しかし、同一アカウントから発信されるツイートであっても、それぞれのツイートは多種多様な話題に言及しており、プロフィール推定対象のユーザの興味等を的確に推定することは困難である。そのため、支持政党や、ある飲食店の支持者

であるか、といった特定の属性を推定する研究が中心となっており、推定対象のユーザのプロフィール全体を推定することが難しい。さらに、ツイートをほとんど発信しないユーザに対しては適用自体が困難となる。

もう 1 つは、ソーシャルグラフを利用する手法である。フォロー情報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う研究[6][7]では、クラスタリングによって得られたそれぞれのコミュニティに対し、キーワードを付与することで、そのコミュニティに属するユーザのプロフィール（属性）を推定している。しかし同手法では、それぞれのユーザが属しているコミュニティに付与されたキーワードのみをそのユーザのプロフィールとして出力しているため、それぞれのユーザに対して、1 つのコミュニティの特徴を示す限定された属性しか付与することができないという問題がある。またフォロー情報を利用する場合には、フォローしている相手が実際に友人であるのか、有名人であるのか、といったフォローの理由を考慮することができないため、実際に親密な関係を持つユーザ群をコミュニティとして抽出することが困難である。そのため、異なる属性を持つユーザ群をコミュニティとしてしまい、的確に推定対象ユーザの属性を推定できない可能性がある。

これらの問題を解決する手法として、奥谷らは、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、直接メッセージのやりとりをしているユーザ群をコミュニティとして抽出し、当該コミュニティに属するユーザのプロフィール情報を用いて推定対象ユーザの属性を求める手法を提案[8]している。具体的には、抽出されたコミュニティに属する複数ユーザのプロフィール情報に共通して表れる単語を推定対象ユーザの属性として出力する。抽出されるコミュニティは複数あるため、当該ユーザに対して、より多種多様な属性を付与することができる。また、ツイートの投稿が少なくても、メンションによりやりとりをして

†1 早稲田大学大学院 基幹理工学研究科

†2 早稲田大学グローバルエデュケーションセンター

†3 早稲田大学 理工学術院

†4 国立情報学研究所

a) Twitter, <https://twitter.com/>. (2014 年 6 月 20 日アクセス)

b) Twitter reaches half a billion accounts More than 140 million in the U.S., http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US. (2014 年 6 月 13 日アクセス)

いるユーザ群（コミュニティ）が得られれば当該ユーザの属性推定が可能となる。

一方で同手法では、多くのコミュニティから共通して抽出される単語を、それぞれのコミュニティの特徴を示さない一般的な単語として見なし、これの重みを小さく設定している（全抽出コミュニティを全体集合と考え、各コミュニティ内で出現するキーワードに idf により重み付けを行っている）。しかし、このような単語の中には、周辺ユーザ全体に共通する属性を表す単語が含まれている可能性があり、これらの単語を除外する奥谷らの手法は周辺ユーザの重要な属性を除外している可能性がある。

そこで本稿では、idf を計算する対象を抽出された全コミュニティではなく、Twitter ユーザ全体のプロフィール集合（実際にはランダムサンプリング）とした。これにより、一般的な語をフィルタリングすることで、抽出された複数のコミュニティに属するユーザ全体のプロフィール文に共通して出現する単語を推定対象ユーザの属性として抽出しやすくする。

以下、2 節で関連研究を紹介し、3 節で提案手法である Twitter ユーザのプロフィール推定手法について説明する。そして、4 節で提案手法及びベースラインとなる奥谷らの手法を用いた実験及び評価について述べ、最後に 5 節で本稿をまとめる。

2. 関連研究

2.1 ソーシャルグラフ及びツイートを利用した Twitter ユーザの属性推定

Pennacchiotti らは、ソーシャルグラフ及びツイートに関する情報を用いて、機械学習の分類器によるユーザ分類をし、ユーザの特定の属性を推定する手法を提案[2]している。ツイートに関する情報としては、総ツイート数や、URL やハッシュタグなどを含むツイートの割合及びハッシュタグとツイート本文に含まれる単語を用いている。手順としては、まずツイートに関する情報及びフォロー関係を用いて、機械学習の分類器である Gradient Boosted Decision Trees (GDBT) によってユーザ进行分类する。そして、GDBT から得られたスコアと、対象ユーザのフレンドとのメンション数及びフォロー関係を用いて新たに算出したスコアの平均を最終的なスコアとして分類結果を更新する。

評価実験では、ユーザが支持している政党、ユーザ自身の民族、ユーザがスターバックスコーヒーの支持者であるかについて、2 値分類を行っている。機械学習、フレンド情報のいずれかのみを用いた分類手法をベースラインとして実験を行った結果、支持政党及びスターバックスコーヒーに関する分類実験では、提案手法が最も F 値が高く、民族の分類では、機械学習のみを利用する手法に続いて、提案手法が 2 番目に高い F 値となっている。しかし Pennacchiotti らの手法では、特定の属性についての推定を

行っており、対象の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推定することができない。

2.2 ソーシャルグラフを用いたユーザコミュニティ抽出

Java らは、フォロー関係を用いてソーシャルグラフを構築し、グラフデータのクラスタリング手法である Girvan-Newman アルゴリズムを適用することで、ソーシャルグラフからユーザコミュニティを抽出[6]している。同手法では、友人関係とみられるコミュニティのみを抽出するため、相互フォロー関係になっているユーザ間のみをエッジで接続した単純無向グラフをソーシャルグラフとして構築している。

実験では、87,897 人の Twitter ユーザのデータからソーシャルグラフを構築し、コミュニティの抽出を行った。そして、ここで得られたあるコミュニティに属する全ユーザのツイートに含まれる単語の出現頻度を取得し、出現頻度の高い単語群をキーワードとして抽出している。抽出したキーワードには、ゲームに関係する単語が多く含まれており、このコミュニティは、ゲームに関連するユーザによって構成されていると考えられる。しかし Java らの手法では、それぞれのユーザがただ 1 つのコミュニティに属していると仮定しているため、当該コミュニティに関係する限られた属性しかユーザに付与することができない。また、フォロー情報によってソーシャルグラフを構築しているため、メンション情報によってソーシャルグラフを構築する場合と比較して、メンションのやりとりの数などの、ユーザ同士の親交の深さを考慮することができない。そのため、実際に親密な関係を持つユーザ同士をコミュニティとして抽出することが困難であり、異なる属性を持つユーザ群をコミュニティとしてしまい、的確な属性推定が行えない可能性がある。

2.3 メンション情報を利用した複数のユーザ属性推定

奥谷らは、メンション情報を利用し、プロフィール推定の対象となるユーザの複数の属性を推定する手法を提案[8]している。まず、メンション情報を利用して、プロフィール推定の対象となるユーザ及びその周辺ユーザからなるソーシャルグラフを構築している。そして得られたグラフに対してグラフデータのクラスタリング手法である Newman アルゴリズムを適用することで、コミュニティを取得し、各コミュニティから抽出したキーワード全体を推定対象のユーザのプロフィール（属性）として出力する。ここで、コミュニティに分割しているのは、多様な属性を抽出するためである。また、周辺ユーザとは、プロフィール推定の対象となるユーザと相互にメンションを送り合っているユーザ（1 次ユーザ）及び、2 人以上の 1 次ユーザとメンションを相互に送り合っているユーザ（2 次ユーザ）全体である。キーワードとして抽出する単語は周辺ユーザのプロフィール文から抽出している。具体的には、各コミュニティに属するユーザのプロフィール文から抽出され

た単語 w に対し、式(1)に示す TF-IDF によるスコアを付与している。

$$tf_{wc} = \frac{n_{wc}}{\sum_k n_{kc}}$$

$$idf_{wc} = \log \frac{D}{d_w} \quad (1)$$

$$S(w, c) = tf_{wc} \cdot idf_{wc}$$

ここで、 n_{wc} はコミュニティ c に属するユーザのプロフィール文中の単語 w の出現回数、 D は総コミュニティ数、 d_w は単語 w を含むプロフィール文を持つユーザが属するコミュニティ数である。それぞれのコミュニティのプロフィール情報により多く出現する単語は tf_{wc} がより高い値となり、より少数のコミュニティにのみ出現する単語は idf_{wc} がより高い値となる。したがって、それぞれのコミュニティの特徴を示すようなキーワードのスコア $S(w, c)$ の値が大きくなるスコアリング手法である。奥谷らの手法では、スコア $S(w, c)$ の値が最大となる 10 語をプロフィール推定の対象となるユーザのプロフィールとして出力している。

実験では、フォロー情報を用いてソーシャルグラフを構築する手法をベースラインとして7人のユーザに対してプロフィール推定を行い、Precision@10及び平均逆順位(MRR)によって評価を行っている。実験の結果、Precision@10及びMRRのどちらにおいても、提案手法の値がベースラインとしている手法の値より高くなっている。

しかし、奥谷らの単語のスコアリング手法では、多くのコミュニティのユーザのプロフィールに共通して出現する単語のスコアを低くしているため、周辺ユーザ全体のプロフィール文に共通して出現する単語を、対象ユーザのプロフィールとして出力することが困難である。多くの周辺ユーザのプロフィール文に共通して現れる単語は、周辺ユーザ全体の特徴を表す単語であると考えられるため、むしろ推定対象のユーザのプロフィールとして出力すべき単語であると考えられる。

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手法について述べる。奥谷らの手法では、一般的な単語のスコアを低くするために、複数のコミュニティのプロフィール情報に共通して出現している単語のスコアを低減させている。しかし、このスコアリング手法が原因となり、周辺ユーザのプロフィール文に広く共通して出現する、推定対象ユーザの中心的属性を示す単語のスコアが低くなり、的確な属性推定ができないうちがある。そこで、提案手法では、周辺ユーザに代えて、Twitterユーザ全体からランダムにサンプリングした一般ユーザのプロフィール文に広く出現する単語のスコアを低減させることで、この問題を解決する。

また奥谷らの手法では、多様な属性を抽出するために、

周辺ユーザをクラスタリングし、得られたそれぞれのコミュニティから抽出したキーワードを推定対象ユーザの属性として出力している。しかし奥谷らの手法では、それぞれのコミュニティに所属しているユーザのプロフィール文内における出現頻度の高い単語に高いスコアを付与している。そのため、少数のユーザによって構成されているコミュニティにおいて出現頻度が高い単語には、その他の大多数の周辺ユーザのプロフィール文に出現しない単語であっても、高いスコアが付与されてしまう。このような単語は、推定対象のユーザに強く関連する属性を示している可能性が低いと予想される。そこで提案手法では、奥谷らの手法で行われている周辺ユーザのクラスタリングを行わず、周辺ユーザ全体のプロフィール文内により広く出現する単語に高いスコアを与えることで、推定対象ユーザにより関連が深い属性の抽出を図る。なお、提案手法では周辺ユーザのクラスタリングを行わないが、推定対象ユーザが多様な属性に深く関連していれば、それに伴った多様な属性が抽出されると考えられる。

3.1 提案手法の概要

本項では、提案手法の流れを示す。まず、プロフィール推定の対象となるユーザを起点ユーザとし、メンション情報を用いてその周辺ユーザからなるソーシャルグラフを構築する。次に、ランダムにサンプリングした100,000人のTwitterユーザを一般ユーザとして取得する。そして、ソーシャルグラフに含まれるユーザのプロフィール文に対して形態素解析を行い、ここで取得した単語に対してスコアリングを行い、ソーシャルグラフに含まれるユーザのプロフィール文における出現頻度が高く、かつ一般ユーザのプロフィール文における出現回数が少ない単語をキーワードとして抽出する。ここで得られたキーワード群をプロフィール推定の対象となるユーザのプロフィール(属性)として出力する。

3.2 ソーシャルグラフの構築及び一般ユーザのサンプリング

本研究では、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。具体的には、ユーザ i からユーザ j へのメンション数を m_{ij} とし、ソーシャルグラフの隣接行列 A の要素 a_{ij} を式(2)によって定める。

$$a_{ij} = \sqrt{m_{ij} \cdot m_{ji}} \quad (2)$$

このとき、ユーザ i 及びユーザ j が相互にメンションを送り合っている場合にのみ $a_{ij} \neq 0$ となり、エッジが張られることに注意されたい。また、より多く相互にメンションを送り合っているユーザの間のエッジの重みがより大きくなる。ここで、 $a_{ij} = a_{ji}$ となるため、提案手法におけるソーシャルグラフは無向グラフとなる。

式(2)を用いて、次の手順によってソーシャルグラフを構築する。

1. プロフィール推定の対象となるユーザを起点ユーザ

u_0 とする

2. ユーザ u_0 との間に式(2)で表されるエッジを持つ N_1 人のユーザからなる集合を 1 次ユーザ集合 $U_1 = \{u_{11}, u_{12}, \dots, u_{1N_1}\}$ とする
3. 起点ユーザ u_0 ではなく、かつ 1 次ユーザ集合 U_1 に属さないユーザの中で、1 次ユーザ集合 U_1 に属する 2 人以上のユーザとの間にエッジを持つ N_2 人のユーザからなる集合を 2 次ユーザ集合 $U_2 = \{u_{21}, u_{22}, \dots, u_{2N_2}\}$ とする
4. ユーザ u_0 , 1 次ユーザ集合 U_1 及び 2 次ユーザ集合 U_2 に属するユーザをノードとするソーシャルグラフを構築する

またソーシャルグラフとは別に、Twitter API^{c)}によって収集したデータから、100,000 人のユーザをランダムにサンプリングし、それらのユーザからなる集合を、一般ユーザ集合 $U_g = \{u_{g1}, u_{g2}, \dots, u_{g100,000}\}$ とする。

3.3 単語の取得と重要度の算出

本項では、3.2 項で構築したソーシャルグラフに含まれるユーザのプロフィール文から、推定対象ユーザの属性を示す単語を取得する方法を示す。本項の内容が奥谷らの手法と大きく異なる部分であり、本稿の重要なポイントである。

まず、ソーシャルグラフに含まれるユーザのうち、ユーザ u_0 を除く、1 次ユーザ集合 U_1 及び 2 次ユーザ集合 U_2 に属する N_1+N_2 人のユーザからなる集合を周辺ユーザ集合 $U_s = \{u_{11}, u_{12}, \dots, u_{1N_1}, u_{21}, u_{22}, \dots, u_{2N_2}\}$ とする。周辺ユーザ集合 U_s に属するユーザのプロフィール文に対し、形態素解析エンジンである MeCab^{d)}を利用して名詞を抽出する。なお MeCab で利用する辞書は、固有名詞や流行語や略語に対応するため、はてなキーワード^{e)}及び Wikipedia の項目名^{f)}を利用して拡張している。また、実験時は奥谷らの手法においても同様に辞書の拡張を行う。

周辺ユーザ集合 U_s に属するユーザのプロフィール文から MeCab による形態素解析によって得られた単語 w に対し、式(3)に示すスコア $S(w)$ を付与する。

$$lfp_w = \frac{lp_w}{|U_s|}$$

$$igfp_w = \begin{cases} \log \frac{100,000}{gp_w} & (gp_w \neq 0) \\ \log 100,000 & (gp_w = 0) \end{cases} \quad (3)$$

$$S(w) = lfp_w \cdot igfp_w$$

ここで、 lp_w は周辺ユーザ集合 U_s に属するユーザのプロフィール文のうちの単語 w を含むプロフィール文の数、 gp_w は一般ユーザ集合 U_g に属するユーザのプロフィール文のうちの単語 w を含むプロフィール文の数を示す。式(3)では、ランダムにサンプリングした一般ユーザのプロフィール文に頻出する単語のスコアを低減させることで、一般的な単語のフィルタリングを行うと同時に、推定対象となるユーザの周辺ユーザに特有な単語のスコアを高くしている。

周辺ユーザ集合 U_s に属するユーザのプロフィール文から得られたすべての単語についてスコア $S(w)$ を算出し、最終的にスコア $S(w)$ が最も高い 10 語を、推定対象ユーザのプロフィール (属性) として出力する。

4. 実験・評価

4.1 データセット

本研究で扱うデータは、Twitter API を用いて収集した。具体的には、Streaming API^{g)}によって、2013 年 1 月 1 日から同年 12 月 31 日にツイートを投稿したユーザの内、ユーザインターフェースの言語設定を日本語にしているユーザの ID を取得し、ここで得られた ID を元に、REST API^{h)}によって、そのユーザのツイートデータを最大 2,000 ツイート収集した。その結果、7,957,324 ユーザのデータを収集できた。3.2 項で述べた一般ユーザ集合 U_g のデータは、この方法で収集した 7,957,324 ユーザを母集団とし、ランダムに 100,000 ユーザをサンプリングすることで取得している。

4.2 実験および評価の方法

被験者の Twitter ユーザ 6 人を対象として、提案手法及び奥谷らの手法を用いて、プロフィールの推定実験を行い、その結果に対して被験者による評価を行った。具体的には、我々の提案手法及び奥谷らの手法によって被験者のプロフィールとして出力された 10 語ずつのキーワードをそれぞれ被験者本人に確認してもらい、その中から自身に関連するキーワードを選出してもらうことで、それぞれのキーワードが被験者の属性として正しいか否かの 2 値分類による評価を行った。そして、ここで得られた 6 人分の評価結果の Precision@10 及び MRR を算出した。

また Precision@10 及び MRR の値についての考察に加え、提案手法及び奥谷らの手法によって出力された属性そのものについての考察を行った。出力された属性については、それぞれの手法で抽出した属性が示す概念の広さ及び属性の多様性についての評価を行った。概念の広さについては、実験を行った結果、手法ごとに抽出される属性の概念の広さに傾向がみられることがわかったため、その考察を行う。また属性の多様性については、奥谷らの手法では多様な属性を抽出するために周辺ユーザのクラスタリングを行って

c) Documentation | Twitter Developers, <https://dev.twitter.com/docs>. (2014 年 6 月 22 日 アクセス)

d) MeCab: Yet Another Part-of-Speech and Morphological Analyzer, <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>. (2014 年 6 月 22 日 アクセス)

e) はてなキーワード一覧ファイル, <http://developer.hatena.ne.jp/ja/documents/keyword/misc/catalog>. (2014 年 6 月 22 日 アクセス)

f) Index of /jawiki/latest/, <http://dumps.wikimedia.org/jawiki/latest/>. (2014 年 6 月 22 日 アクセス)

g) The Streaming APIs | Twitter Developers, <https://dev.twitter.com/docs/api/streaming>. (2014 年 6 月 22 日アクセス)

h) REST API v1.1 Resources | Twitter Developers, <https://dev.twitter.com/docs/api/1.1>. (2014 年 6 月 22 日アクセス)

いるため、提案手法がクラスタリングを行わずに、いかに多様な属性を抽出することができているかについての評価を行う。

4.3 Precision@10 及び MRR による評価

6 人の被験者それぞれの評価結果をもとに算出した Precision@10 及びその平均を図 1 に示し、同様に MRR を図 2 に示す。また図 1 及び図 2 には、提案手法及び奥谷らの手法からそれぞれ取得された 10 個の単語が示すユニークな属性の種類の数も示しているが、これについての説明及び評価は 4.5 項で述べる。

図 1 を見ると、Precision@10 の値は、ユーザ D においては等しいが、その他のすべてのユーザにおいて提案手法の方が奥谷らの手法より高いことがわかる。また図 2 を見ると、MRR の値は、すべてのユーザにおいて提案手法の方が奥谷らの手法より高いことがわかる。平均の値は、Precision@10 が奥谷らの手法では 0.37、提案手法では 0.78、MRR が奥谷らの手法では 1.44、提案手法では 2.61 となっている。

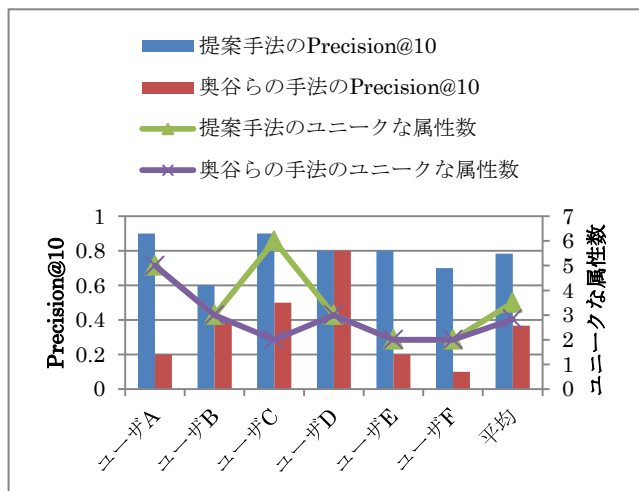


図 1 6 人の被験者による評価結果の Precision@10 及びユニークな属性数

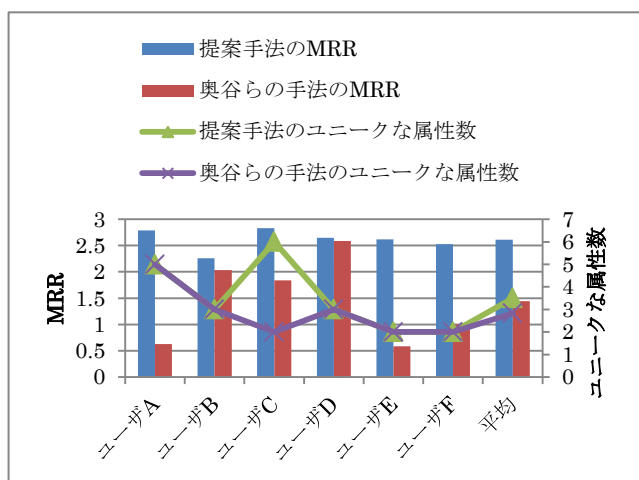


図 2 6 人の被験者による評価結果の MRR 及びユニークな属性数

また提案手法と比較して、奥谷らの手法における Precision@10 及び MRR の値は、ユーザによって大きく異なっている。提案手法の Precision@10 及び MRR の標準偏差はそれぞれ 0.12 と 0.21 であり、それに対して奥谷らの手法の Precision@10 及び MRR の標準偏差はそれぞれ 0.26 と 0.83 である。

これらのことから、提案手法では、奥谷らの手法と比較して、プロフィール推定の対象となるユーザの選び方によらず、安定して高い精度でプロフィールを推定できていることがわかる。

4.4 出力された属性が示す概念の広さについての考察

本項では、提案手法及び奥谷らの手法によるプロフィール推定結果の例を示し、プロフィールとして抽出された属性が示す概念の広さについての考察を行う。推定結果の例として、2 人の被験者に対する推定結果をそれぞれ表 1、表 2 に示す。

表 1 プロフィール推定結果の例 1

順位	提案手法		奥谷らの手法	
	キーワード	正誤	キーワード	正誤
1	情報	○	小学校	×
2	早稲田	○	修士	○
3	早稲田大学	○	動画	×
4	理工	○	艦これ	×
5	アニメ	○	ボカロ	×
6	大学生	○	応用数理	×
7	学科	×	初心者	×
8	東方	○	東京	○
9	ゲーム	○	百合	×
10	野球	○	GTO	×

表 2 プロフィール推定結果の例 2

順位	提案手法		奥谷らの手法	
	キーワード	正誤	キーワード	正誤
1	水樹奈々	○	奈々	○
2	理工	○	理工	○
3	アニメ	○	水樹奈々	○
4	早稲田大学	○	学園祭	○
5	なのは	○	物理	×
6	会員	×	修士	○
7	早稲田	○	GROUP	×
8	ゲーム	○	PHP	○
9	大学	×	Web	○
10	奈々	○	大学院	○

表 1 及び表 2 を見ると、所属や肩書に関する属性として、奥谷らの手法では「大学院」「理工」「修士」などが抽

出されており、提案手法ではそれらに加えて、「早稲田大学」「大学」などのより大きなコミュニティを示す属性が抽出されていることがわかる。

また、アニメやゲームなどの趣味に関する属性においては、提案手法の方が奥谷らの手法より正答率が高い。奥谷らの手法では「DDR」、「艦これ」、「ボカロ」などのゲームや音楽の具体的なタイトルやカテゴリを示す属性が抽出されているのに対し、提案手法では「アニメ」、「ゲーム」、「野球」などのより上位の概念を示す属性が抽出されている。その結果、より幅広いユーザに該当する上位の概念を示す属性を抽出する提案手法の正答率が高くなっていると考えられる。

提案手法及び奥谷らの手法で抽出した属性の間にこれらの差異が現れる要因は2つ考えられる。1つは、周辺ユーザのクラスタリングを行うか否かの差異である。奥谷らの手法では、周辺ユーザをクラスタリングし、得られたそれぞれのコミュニティ内で、より出現頻度の高い単語を推定対象ユーザの属性として抽出している。それに対して、提案手法では周辺ユーザ全体に共通して出現する単語に高いスコアを与えているため、多くの周辺ユーザに共通するような、上位の概念を示す属性が抽出されていると考えられる。

もう1つは、一般的な単語のフィルタリング方法の差異である。奥谷らの手法では、周辺ユーザをクラスタリングして得られた複数のコミュニティ間に共通して出現するような単語のスコアを低減することで、一般語のフィルタリングを行っている。その結果、多くの周辺ユーザに共通している、上位の概念を示す単語のスコアが低下し、より下位の概念を示す単語を属性として抽出していると考えられる。それに対して、提案手法では一般語のフィルタリングに、Twitter ユーザ全体からランダムにサンプリングした一般ユーザを利用しているため、周辺ユーザ全体に共通する、上位の概念を示す単語のスコアを低減させず、これを推定対象ユーザの属性として抽出することができていると考えられる。

4.5 出力された属性の多様性の評価

本項では、提案手法が、周辺ユーザのクラスタリングを行わずに、いかに多様な属性を抽出することができるかの評価について述べる。図1及び図2に示した「ユニークな属性数」は、それぞれの手法によって出力された10個の単語のうち、同様な属性を示す単語をまとめて1つの属性としてカウントしたときの出力属性数である。例えば表1における、「早稲田大学」、「理工」はいずれも推定対象ユーザが所属している大学に関する属性を示すため、これらをまとめて「大学」という種類の属性としている。図1及び図2を見ると、ユーザCについては提案手法の方が、多様な属性を抽出できていることがわかる。またその他の

ユーザに関しては、等しい数のユニークな属性を取得できていることがわかる。

これらの結果から、提案手法ではクラスタリングを行っていないが、奥谷らの手法と同等の多様性を持つ属性を抽出できていることがわかる。

5. おわりに

本稿では、奥谷らが提案している、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における、推定対象となるユーザの周辺ユーザ全体のプロフィール文に広く出現する単語を属性として抽出することが困難であるという問題を解決する手法を提案した。

6人の被験者による実験の結果、提案手法では、奥谷らの手法と比較して、Precision@10 及び MRR において安定して高い値が得られることを確認することができた。また提案手法では、周辺ユーザのクラスタリングを行っていないが、奥谷らの手法と同等の多様性を持つ属性を抽出することが確認できた。

一方で、本稿における手法の評価では、プロフィールとして出力された属性の概念的な広さについては定量的な評価を行っていない。そのため、属性の示す概念の広さに対する定量的な評価方法を考案することが今後の課題となる。

また提案手法では、一般的な単語のフィルタリングに利用した100,000 ユーザというサンプル数が十分であるかの統計的な検証を行っていない。そのため、一般ユーザのサンプル数を統計的な理由付けの元で決定していくことも、今後の課題である。

参考文献

- 1) F. Abel, Q. Gao, G. J. Houben and K. Tao: "Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web", The Semantic Web, Research and Applications, Springer Berlin Heidelberg, pp. 375-389, 2011.
- 2) M. Pennacchiotti and A. M. Popescu: "Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter", Proc. of the KDD'11, pp.430-438, 2011.
- 3) M. Pennacchiotti and A. M. Popescu: "A Machine Learning Approach to Twitter User Classification", ICWSM'11, pp. 281-288, 2011.
- 4) M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini and F. Menczer: "Predicting the Political Alignment of Twitter Users", Proc. of the International Conference on Social Computing, 192-199, 2011.
- 5) Z. Cheng, J. Caverlee and K. Lee: "You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users", Proc. of the CIKM'10, pp. 759-768, 2010.
- 6) A. Java, X. Song, T. Finin and B. Tseng: "Why We Twitter: Understanding Microblogging Usage and Communities", Proc. of the WebKDD, pp.56-65, 2007.
- 7) K. H. Lim and A. Datta: "Finding Twitter Communities with Common Interests using Following Links of Celebrities", Proc. of the MSM'12, pp. 25-32, 2012.
- 8) 奥谷貴志, 山名早人: "メンション情報を利用した Twitter ユーザプロフィール推定", DEIM'14, B2-3, 2014.