待遇表現に着目したマイクロブログにおけるユーザのクラスタリング

畑本典宣 黒澤義明 目良和也 竹澤 寿幸
広島市立大学大学院 情報科学研究科 知能工学専攻
[hatamoto,kurosawa,mera,takezawa]@ls.info.hiroshima-cu.ac.jp

1.はじめに

近年のインターネットの発展や多くのコミュニケーションツールが登場したことで，インターネット上で他者とコミュニケーションを図ることが日常化してきた．そのコミュニケーションツールのなかでも，近年注目されているのが，Twitter[1]に代表されるマイクロブログである．

Twitter は，ユーザが簡単に自由な情報を発信できるため，簡易ブログとしての機能を有する．また，Twitter では，フォロー・フォロワーという関係が存在する．ユーザが他のユーザの発信するツイートを自動的に取得する機能をフォローという．フォローされたユーザにとって，そのフォローしてきたユーザはフォロワーとなる．この関係が非常に多くのユーザのネットワークを構成しているため，Twitter は Social Networking Service としての機能も有する．

Twitter には，様々なユーザが存在し，Twitter に 140 文字以内の短文(以下，ツイート)を投稿することで，他のユーザとコミュニケーションを図り，多種多様なコミュニティを形成している．そのコミュニティのなかには，「上司と部下」，「先輩と後輩」，「教員と学生」などの，“タテのつながり”，また，「会社の同期」，「学校の友達」などの“ヨコのつながり”が存在する．

そこで本研究では，前述の“タテのつながり”に着目し，待遇表現を用いて Twitter のなかに形成された“タテのつながり”を示す上位関係の抽出を目的とする．具体的には，ツイート中の「です」・「ます」という表現を手がかり語として利用し，ユーザ間の返信にこの手がかり語が一方向に大量に存在すれば，上位関係とみなす．この上位関係を抽出することが出来れば，影響力のあるユーザを特定することが可能になり，そのユーザのツイートを情報推薦のソースとして利用，またそのユーザを中心にクラスタリングを行い，“タテのつながり”に基づいたコミュニティの抽出が可能になると考えられる．

なお，本学に関係する小規模なデータで実験を行う理由として，大学では学部や学科，学年などでコミュニティが構成されているため，解析が容易であること，実際に情報推薦などを行う際に，対象を決定しやすいなどの理由が挙げられる．

[1] http://twitter.com

2. 関連研究

本章では，本研究に関連する研究を紹介する．

2.1. Twitter ユーザのクラスタリング

2.1.1. 返信行動に基づいたクラスタリング

黒澤ら[1]は，Twitter における返信行動に着目し，その元投稿および返信の内容から投稿内容ベクトルなる指標を用いてユーザのクラスタリングを行い，結果を SOM を用いて視覚化している．黒澤らは，ユーザのクラスタリングを有効に行うために，ユーザ間の興味の共通化を行っている．しかし，この手法で得られるクラスタは，ある興味に基づいたクラスタであり，上位関係を抽出するにあたっては，上下関係にあるからといって二者間の興味が一致するとは限らず，この手法は不向きであると考えられる．

2.1.2.フォロー関係を基づいたクラスタリング

畑本ら[2]は，Twitter におけるフォロー・フォロワー関係をソーシャルグラフのエッジとみたててグラフを作成し，そのグラフを用いてユーザのクラスタリングを行った．また，得られたクラスタがどのような特徴を持っているのかを特定するために，各クラスタに属するユーザのツイートを解析し，tf-idf 法を改良した手法を用いて，そのクラスタ内で重要と思われる名詞句を抽出し，これをクラスタの特徴を示す特徴語としている．この手法で得られるクラスタは，グラフを作成する際にフォロー・フォロワー関係に着目しているため，各ユーザは平等に扱われる．しかし，実世界のコミュニティ同様，Twitter の世界には“ヨコのつながり”だけでなく“タテのつながり”が存在し，ユーザが情報の発信力に関しては必ずしも平等とは言えない．ゆえにこの“タテのつながり”に着目することで，影響力のあるユーザを特定すれば，より現実に即したコミュニティ抽出が可能になると思われる．

2.2. マイクロブログにおけるユーザ意思の解析

Java ら[3]は，なぜ多くの人がマイクロブログを利用しているのか，また，どのようにマイクロブログを利用しているのかを解析している．その一つとして，Twitter ユーザの意思を解析している．Java らは，ユーザ意思解析のために，HITS アルゴリズムを用いて，ネットワークにおける Twitter ユーザの“Authority”と

“Hubs”を算出している. この 2 つはフォロー・フォロワー関係をもとに算出している. Java らは, この 2 つの値によって, ユーザの意思を以下の 3 つに大別している.

1. friendship-wise relationship
2. information seeking
3. information sharing

“friendship-wise relationship”は Authority と Hubs が両方高い場合, “information sharing”は Authority が高く, Hubs が低い場合, “information sharing” は Authority が低く, Hubs が高い場合である. HITS は元々重要性の高い Web ページを抽出するアルゴリズムであるので, この手法を用いれば, Twitter ユーザの上下関係を抽出できる可能性はあるが, フォロー・フォロワーに基づいている解析を行っているので, ツイートそのものには触れていない. ゆえに, 実際に抽出された結果が上下関係といえるかは定かではないといえる.

3.提案手法

本研究の提案する手法は, Twitter のユーザ群に対して, 収集した各ユーザのツイートから待遇表現を手がかりとし, Twitter のなかでの“タテのつながり”を示す上下関係を抽出することである. 以下にその詳細を述べる.

3.1.待遇表現

待遇表現とは, ある話者が自分自身と聞き手, またはその話題に登場する第三者との社会的な強さ, 状況, 感情などに応じて行う表現のことである. 待遇表現には, 敬意表現や軽卑表現などが含まれる. さらに, この敬意表現には, 尊敬語, 謙譲語および丁寧語が含まれる. 本研究では, 解析が比較的簡単な丁寧語に焦点を当てる. 丁寧語は, 聞き手の方が話者より上位の関係にあるときに使われる語である. 具体的には, 「です」, 「ます」およびその活用形を上位関係の抽出するための手がかり語として用いる.

3.2. 上下関係の抽出方法

上下関係の抽出方法を述べる. まず, 各ユーザの収集したツイートから, 文頭が”@”で始まるツイート(以下, リプライ)を抽出する. 通常, リプライはあるユーザのツイートに対しての返信に用いられる. そして, そのリプライに対して形態素解析器 MeCab[2]を用いて形態素解析し, 手がかり語の有無を調べ, ユーザ間の上下関係を判定する. 例えば, あるユーザ A が, 「@B 今から研究室に向かいます!!」というツイートを投稿したとすると, ツイート中に「ます」という手がかり語が含まれているので, B というユーザは, ユーザ A より上位の関係にあると判定する. しかし, ツイートは非常に短い文章であることが多く, 仮に上下関係にあるユーザ間であっても, ツイート中に上下関係を表すような語が含まれない可能性は大いにある. またその逆も然りで, “ヨコのつながり”であってもツイート中に手がかり語が含まれることも少なくない. これらの理由から一度のリプライで上下関係を判断することはできないと考えられる. そこで本手法は, あるユーザ X があるユーザ Y に対して投稿したリプライすべてに対して手がかり語の有無を調べ, 手がかり語が含まれているリプライの数が, 手がかり語の含まれていないリプライの数を上回ったときに, ユーザ Y はユーザ X の上位の関係にあると判定する. これをすべてのユーザ間で調べ, 各ユーザ間の上下関係を抽出する. そして, この上下関係をソーシャルグラフとして示す.

3.3.視覚化におけるノードの大きさの設定

上下関係をソーシャルグラフとして表す際に視覚的にわかりやすくするため, ノードの大きさをユーザの社会的強さに応じて変更する必要がある. そこで, 本手法では, 入次数中心性という指標を用いる. 入次数中心性とは, あるノードに対して接続されているエッジが多ければ多いほど高くなる中心性指標の一つである. この入次数中心性を正規化した値をノードの大きさとしてグラフを作成する.

4.実験結果

3 章で述べた手法を用いて, 実際に本学の関係者と思われるユーザに対して, 上位関係を抽出した結果を示す.

なお, 図の作成には統計解析言語 R[3]というツールを用いた.

本学の関係者と思われる約 400 人の Twitter ユーザのアカウントに対して, 例えばユーザ A よりユーザ B が上位の関係にあると判定されたとすると, ユーザ B からユーザ A に向かってエッジを張ると定義する. このような方法で, 作成したグラフを図1に示す. また, その比較手法として, フォロー・フォロワー関係に着目し, ユーザ A がユーザ B をフォローしていた場合, ユーザ A からユーザ B に向かってエッジを張ると定義する. また, ノードの大きさは提案手法と同様に, 入次数中心性を用いて設定した. このようにして作成したグラフを以下の図 2 に示す. なお, グラフ作成に際し, エッジが張られていないノードに関しては, 予めど

[2] http://mecab.sourceforge.net

[3] http://www.r-project.org

ちらの図からも除いている.

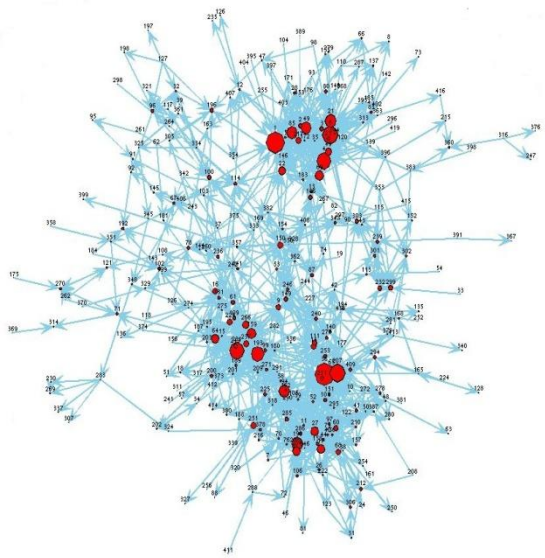

図 1:上位関係の抽出結果を示すグラフ

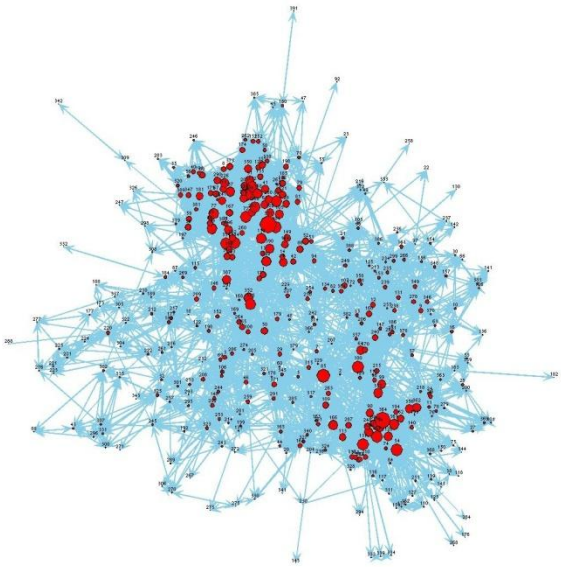

図 2:フォロー・フォロワー関係に基づいたグラフ

5.実験結果

5.1. 視覚的考察

4 章にて述べた上位関係抽出の実験結果について考察を行う. まず, 視覚的観点から考察を行う. 以下に, 提案手法である図1と, 比較手法である図 2 のノードの数とエッジの数を表1に示す.

表1:図1と図 2 のグラフのノード数とエッジ数

	ノード(個)	エッジ(本)
提案手法	326	1,057
比較手法	392	6,098

表1からもわかるように, ノードの数は66 個ほど減少し, エッジの本数は比較手法のグラフと比べ, およそ 1/6 までに減少している. このことから, グラフから“ヨコのつながり”を示すエッジが削除されグラフが簡素化されたということがわかる. また, ノードの大きさは図2 ではあまり差異は見受けられないが, 図1 では比較的に分散している. よって視覚的にも, どのノードが上位の位置に属しているのか一目で判断できるようになったといえる. これらのことから, 比較手法よりも, 視覚的に上下関係を判別することが容易になったといえる.

5.2. ノードの大きさについての考察

続いて, ノードの大きさ(直径)についての考察を行う. ノードの大きさは, 3 章でも述べたとおり, 入次数中心性を正規化した値を用いている. 以下の表 2 に, 入次数中心性を正規化した値が高かったユーザ, 上位 5 件についての入次数中心性の値と広島市立大学における属性を示す. なお, この属性は人手で付与している.

表 2:入次数中心性値, 上位 5 件のユーザ情報

ユーザ	入次数中心性	当校における属性
A	1.00	教員
B	0.94	教員
C	0.79	博士前期
D	0.79	博士後期
E	0.72	博士前期

表 2 から, 入次数中心性の値が大きいユーザは, 大学というコミュニティではあるが, 社会的に強ければ強いほど高いということがわかる. また, 今回はスペースの都合上, 入次数中心性の値が高かった上位 5 件についてのみ掲載しているが, 上位 20 件程度は, 社会的に強い属性が付与されているユーザだったので, この手法は上位関係を抽出するにあたって, 有効であるといえる.

5.3. 5.1.節と 5.2.節の総合的考察

以下に, 今回の実験の有効性を示すため, 今回の実験で得られたグラフ図 1 と図 2 を拡大したグラフを図 3, 図 4 にて示す. なお、どちらの図も縮尺は同一である.

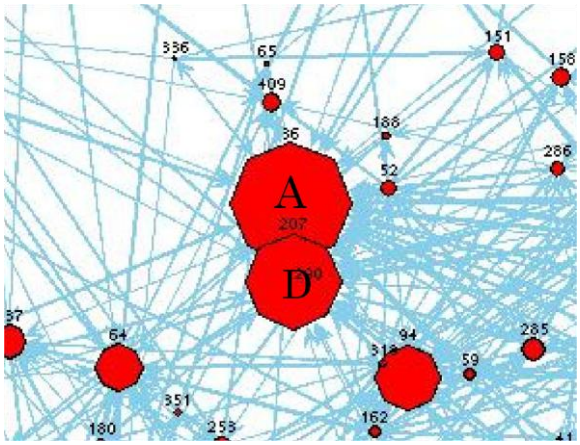

図 3:提案手法におけるユーザ A 付近の図

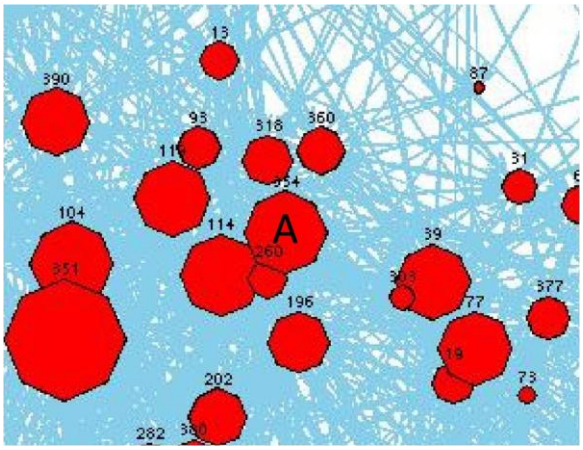

図 4:比較手法におけるユーザ A 付近の図

図 3 および図 4 を比較すると，5.1 節で述べたように，エッジの本数が激減しているのがわかる．また，図 3 にてノード A は多くのエッジが張られているのが見てとれ，ノードの大きさも図 4 より大きくなっているのがわかる．ゆえに，提案手法において上位関係の抽出が視覚的にわかりやすくグラフ化されているといえ，提案手法の有効性を示しているといえる．

5.4. 入次数中心性の考察

この節では，提案手法と比較手法，それぞれの入次数中心性の値についての考察を行う．各ノードについて，提案手法および比較手法における入次数中心性を算出しその差を計算すると，差が 0.1 以上あったユーザは，わずか 4 件であった．以下に，その 4 件についての入次数中心性の値と広島市立大学における属性を示す．なお，表中の数値は入次数中心性の値を示しており，表 2 と同じユーザの表記がなされている場合，表 3 にても同一ユーザを示す．また，属性についても表 2 と同義である．

表 3:各手法における入次数中心性と属性

ユーザ	提案手法	比較手法	属性
A	1.00	0.68	教員
B	0.94	0.72	教員
D	0.79	0.51	博士後期
F	0.58	0.32	教員

表 3 から，各ユーザとも比較手法よりも提案手法の方が入次数中心性の値が 0.2 以上大きくなっている．ゆえに，当校においてかなり上位の社会的強さを持っているといえる．ゆえに，本手法の有効性が示されたともいえるが，裏を返せば他のユーザに関しては，差がなかったとも言える．これは，提案手法ではリプライに着目して上位関係の抽出を行った．しかし，そもそもリプライはお互いにフォローしているユーザ同士が行う行為であるため，提案手法と比較手法の間に差が生じなかったと考えられる．この点は今後の課題として，別の指標などを検討する必要がある．

6.おわりに

本研究では，広島市立大学の関係者と思われる Twitter ユーザ群に対して上位関係の抽出を行った．その結果，視覚的および入次数中心性という観点からも，良好な結果が得られたので，本手法の有効性が示されたといえる．今後の課題として，本手法は待遇表現として，敬意表現の一種である丁寧語の「です」および「ます」という語を手がかり語として実験を行った．しかし，待遇表現には丁寧語だけではなく，他に多くの種類が存在する．今後は，その他の表現にも着目し，手がかり語を増やし，上位関係の抽出の行っていく必要がある．また，今回の実験は広島市立大学というコミュニティに対して行ったが，今後は他のコミュニティに対しても実験を行い，この手法の有効性を示す必要があると考える．

謝辞

この研究の一部は，平成 23 年度広島市立大学特定研究費(一般研究)の補助を得ている．関係各位に感謝申し上げる．

参考文献

[1] 黒澤ら，“マイクロブログサービスの返信行動に着目した投稿及びユーザの分類”，言語処理学会第 17 回年次大会発表論文集，pp.460-463，2010.

[2] 畑本ら，“マイクロブログにおけるユーザのクラスタリングとその特徴語抽出”，言語処理学会第 17 回年次大会発表論文集，pp.280-283，2010.

[3] Java et al.，“Why We Twitter : Understanding Microblogging Usage and Communities,” Procedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop on Web mining and social network analysis，2007.