社会ネットワーク分析に基づく新規指標及びリプライ回数に着目した決定木による マイクロブログユーザの分類とその分析

河下勇太 黒澤義明 目良和也 竹澤 寿幸
広島市立大学情報科学部広島市立大学大学院 情報科学研究科

1. はじめに

近年，ネットワーク構造を用いた研究が多く行われている．その中でも，ノードのリンクに張られている近傍の情報を用いた「リンクに基づく分類」では，その精度を上げるためにネットワーク構造を用いた様々な指標が考案されている．例えば，唐門らの研究[1]では，社会学の一分野である社会ネットワーク分析に基づく指標から新たな指標を生成し，論文データベースとコミュニティサイトについてノードの分類を行い，指標の有用性について述べている．「リンクに基づく分類」において新しい指標の生成は必要であると考えられる．

本研究では，ユーザの分類行う．新しい指標を用いることにより，その精度が上がれば，ユーザのグループ化が可能となり，そのグループが興味を持ちそうな内容の情報推薦が可能となる．

2. Twitter

本研究では，Twitterを用いて研究を行う．Twitterには，他のブログサービスとは異なり，フォローという記事の閲覧を簡単に行う機能がある．従来のブログでは，Facebookのように，相手の投稿した記事をタイムラインに表示させるためには双方向から許可しないと記事を取得できないのに対して，Twitterでは，片方向から興味のあるユーザをフォローするだけで，非公開にしてないユーザをフォローしない限りは，記事を自動的に取得することが可能になる．ゆえに，Facebookなどに代表される双方向コミュニケーションが基本であるSNSとは一線を画している．

また，Twitterでは記事を投稿する際に，140字以内の文字制限がある．記事の内容が短いため，記事の投稿が容易であり，Twitterはチャットのような役割をする場合もある．ユーザが投稿した記事に返信する機能としてリプライという機能がある．

本研究は，Twitterのフォロー関係とリプライに着目する．

3. 関連研究

1章で述べた唐門らの研究では，社会学の一分野である社会ネットワーク分析でよく用いられる8つの指標を3段階のステップに分解し，組み合わせ，新しい指標を生成し，ノードのカテゴリ分類について生成した指標が有用であるか考察している．8つの指標とは，社会ネットワーク密度，次数中心性，近接中心性，媒介中心性，平均パス長，クラスタ係数，構造同値，構造空隙である．ここで問題がある．唐門らの研究では無向グラフのデータにしか実験を行っていない点である．Hatamoto et al.[2]の研究では，Twitterの記事から「です」,「ます」などの文末表現に着目し，ユーザ間のリンクにはそれぞれ上下関係があると考察している．Twitterには独特のリンク関係がある．本研究では，有向グラフのデータに焦点を当て，Twitterを用いてTwitterユーザに対して分類を行い，新規指標の有用性について述べる．

Twitterは単なるリンクではなく，岩木ら[3]は，Twitterユーザのフォロー・フォロワー関係とリプライ回数に着目し，ユーザの近接度を求め，ユーザにとって読む価値のある記事を効率よく発見する方法を考案している．しかし，本研究の目的は，Twitterユーザの分類である．本研究では，ユーザの近接度を求める手法をリンクの分析に応用する．

4. 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，フォロー関係からリンク関係を求め，新規指標とユーザの近接度を用い，ユーザの分類を行う．新規指標の有向グラフに対する有用性，リプライ回数におけるユーザの分類の変化について分析する．

4.1. 社会ネットワーク分析から生成した指標

指標は，前章で述べた唐門らの生成した指標のステージ 3 までを用い，正のノード集合は用いないことにした．リンク関係は，フォローしていたら 1，フォローしていなければ 0 で表す．指標は，以下の表 4.1 に示す．

4.2. ユーザの近接度計算

Twitter のリンク関係には，フォロー関係という人とのつながりを表す概念があるが，つながりの度合いに関する情報は持たない．他人の投稿を読むだけのつながりなのか，それとも連絡を取り合うつながりなのか知るためにはフォロー関係以外の情報も必要となる．

よって，本研究では，ユーザの書き込みに対する返信である，リプライに着目し，リプライ回数からユーザの詳細なリンク関係を求める．詳細なリンク関係のことを以後，ユーザ近接度と言う．以下に，ユーザ近接度について求める式(1)を示す．

$$y = \frac{1}{x+1} \qquad (1)$$

y　：ユーザ近接度

x　：リプライ回数

式(1)において算出される値は，リプライ回数が多ければ多いほど，ユーザ近接度の数値が低くなる．つながりが強いほどユーザ間の距離が短くなるように設定する．また，フォロー関係が 0 であるのに対して，リプライ回数が多いと 0 に値が近づくためユーザ近接度でリンクがない場合は，∞(無限大)と設定する．ユーザ近接度を用いて，社会ネットワーク分析から生成した指標と組み合わせ新しく指標を生成する．以下で，生成した指標を表 4.1 に示す．また，重み付けの区別を行うため重み付けをしてなければ n，していれば w，と指標を区別する．

本実験で用いる指標は，4 つのステージで組み合わされた計 64 の指標である．

また，指標の表記方法はステージ 4 から順にハイフンで組み合わせる．表記例を以下に示す．

●指標の表記例

avg-link-ad-n，sum-link-ad-n，max-link-ad-n，min-link-ad-n，・・・・・・．min-shortpass-ac-w

表 4.1 本研究で用いる指標

ステージ	表記	説明
1	n	リンク関係に重み付けなし
1	w	リンク関係に重み付けあり
2	ad	ノード x の近接ノード集合
2	ac	ノード x の到達可能ノード集合
3	link	リンクがあれば 1，なければ 0
3	lenpass	ノードペア間のパスの長さ
3	dis	ノード x とその他のノードの距離
3	shortpass	最短パスがノード x を経由していれば 1，していなければ 0
4	avg	平均
4	sum	合計
4	max	最大値
4	min	最小値

5. 実験

実験データとして，大学に関するユーザを 392 人収集した．Twitter ユーザのアカウントに対して，ユーザをノードとして，どのユーザともリンクが張られなかった場合は，ノードを除外する．

また，リプライに関して本学に関するユーザ 392 人に関するリプライ 74774 件を用いて，有向グラフと無向グラフ両方のリンクに重み付けを行う．

本学に関する小規模なデータで実験を行う理由として，大学のデータには，学部や学科，学年などのグループが構成されているため，解析が容易であることが挙げられる．

4 章で述べた手法を用いて本学の関係者であるユーザを分類する．

5.1. 無向グラフと有向グラフへの有用性

4 章で生成した指標を用い，ユーザの分類を行った．ここで，無向グラフと有向グラフの有用性について述べるために，それぞれのデータを設定する．

有向グラフのデータでは，フォロー関係があれば 1，なければ 0 のリンクを張る．

無向グラフのデータ作成として，

(i) 片方向がフォローしている場合のみ

(ii) 相互にフォローしている場合のみ

以上の 2 通りが考えられる．

(ⅱ)の場合，Twitter の特性として，リンクが片方向にしかない場合に考えられる理由は，そのフォローしたユーザが面白い記事または有用な情報を投稿してくれるから，または，似たようなユーザとして紹介されてフォローした場合も考えられ，両方向のリンクのみのリンク関係を採用すると現実世界とは異なったリンクを除くことができる．

よって，無向グラフのデータとして両方向のリンクがある場合のみ，リンクを張ったグラフを採用することにした．

有向グラフと無向グラフについて，各ノードに対する指標を生成し，c5.0 法を用いて決定木を学習，各ノードが指定したカテゴリに分類されるかどうか推定し，その再現率，適合率，F 値を評価する．今回, 分類は本学の学部である情報科学部とそれ以外，芸術学部とそれ以外で実験行う．なお，情報科学部に属するユーザは 193 人，芸術学部に所属するユーザは 100 人である．以下，情報科学部は情報，芸術学部は芸術と略す.

有向グラフにおける分類の結果を表 5.1.1 に，無向グラフにおける分類の結果を表 5.1.2 に示す.

表 5.1.1 有向グラフにおける再現率，適合率，F 値

重み付け	学科	再現率	適合率	F 値
なし	情報	0.737	0.725	0.731
	芸術	0.983	0.58	0.729
あり	情報	0.9	0.937	0.919
	芸術	0.951	0.78	0.857

表 5.1.2 無向グラフにおける再現率，適合率，F 値

重み付け	学科	再現率	適合率	F 値
なし	情報	0.751	0.802	0.776
	芸術	0.971	0.677	0.798
あり	情報	0.872	0.958	0.913
	芸術	0.97	0.646	0.776

5.2. 小規模なグループにおける分類の変化

本学の関係者のリプライ 74744 件を用いて，本学の研究室である言語音声メディア工学研究室のデータに対して，有向グラフ，無向グラフ，重み付けなし, 重み付けありの 4 通りの再現率を以下の表 5.2.1 で示す．どのようなユーザが分類されるかについて考察する．

表 5.2.1 それぞれの再現率

グラフ	有向	有向	無向	無向
重み付け	なし	あり	なし	あり
再現率	0.791	0.837	0.791	0.744

6. 考察

6.1. 無向グラフと有向グラフへの有用性

F 値に着目すると，表 5.1.1 と表 5.1.2 を見比べると，重み付けを行う前では無向グラフの値が良いことが見てわかる．しかし，リプライ回数から Twitter ユーザの近接度を計算し，指標に取り入れると，有向グラフの値が良いことがわかる．

よって，唐門らの生成した指標は，有向グラフに対して，ユーザ近接度を用いた場合，指標は有用であることが示せた．

6.2. 小規模なグループにおける分類の変化

5.2 節の分類の結果を分析すると，重み付けありにおける無向グラフと有向グラフのデータでは，無向グラフでは分類することができなかったユーザに対して，有向グラフでは，片方向のリンクが多いユーザの分類に成功した．また，片方向のリンクが多いユーザのリンク関係から新たなユーザの分類を行うことができた．以上のことから，有向グラフに関して以上の指標は有用であることが示せた．

7. おわりに

本研究では，Twitter のユーザ群に対して社会ネットワークから生成された新たな指標とユーザ近接度を用いて無向グラフと有向グラフについて分類を行い，その結果について考察した．結果，ユーザ近接度を用いた場合のみに有向グラフの F 値が向上し，詳細なリンク関係を用いると新たな指標が有用であることがわかった．

・参考文献

[1]唐門準，松尾豊，石塚満:“リンク関係に基づく分類のためのネットワーク構造を用いた属性生成”，情報処理学会論文誌 Vol. 49 No.6 2212-2223(June 2008)

[2]Hatamoto,N.,Kurosawa,Y.,Hamada,S.,Mera,K.,and Takezawa,T.:“Finding Social Relationships by Extracting Polite Language in Micro-blog Exchanges”,8th International Conference on Natural Language Processing(JapTAL 2012)

[3]岩木祐輔，アダムヤトフト，田中克己:“マイクロブログにおける有用な記事の発見支援”,DEIM Forum 2009 A6-6