1V-6

フォロー関係に基づく Twitter ユーザの分類

山下　拓也　　佐藤　晴彦　　小山　聡　　栗原　正仁
北海道大学　大学院情報科学研究科

1　研究背景

近年, マイクロブログサービスと SNS 中間の両方の性質を持つ Twitter が注目されている. Tweet と呼ばれる 140 字以内の短文を Web に投稿する機能によりマイクロブログの一面を持ち, また二ユーザ間において, あるユーザが指定したユーザの情報を購読するフォローと呼ばれる機能により SNS の一面も持っている.

二者間で行われるフォローは「友人」「同じ興味」など, 何らかの属性に基づいて行われることが多い. よってあるユーザ集合内にフォローが多数存在する時, 同じ属性に基づいたユーザは互いにフォローしあい, まとまりを構成しているのではないかと考えることができる.

特にあるユーザがフォローしているユーザの集合 (以下, フォロー集合) については, このようなまとまりに分けることは, 分類結果後のそれぞれのクラスタと新しいユーザとのフォローの関係性を見ることによって情報推薦ができたり, あるユーザが第三者のユーザのフォロー集合を見てもどのユーザが共通の属性を持ちまとまっているか一目ではわからない現状の改善にも役立つという点で有用である. 実際後者の問題は, 直接フォロー集合内のユーザにアクセスしプロフィールやツイートを確認することで把握することもできるが, 手間がかかり不便である. 他の方法としてユーザが同じ関係を持つユーザをまとめるリストという機能もあるが, 人手で作らなければならず, フォローしているユーザに比例して作成コストが増えるという問題もある.

2　研究目的

本研究では, フォロー関係に基づいたユーザのフォロー集合の分類を目的とする. この分類は, フォロー集合内において同じ属性に基づいてフォローされたユーザは互いにフォローしあっているという仮定に基づいて行う.

3　類似研究

近年 Twitter に関連した研究が盛んに行われている. 本研究のようにユーザの属性やユーザ間の関係に関する研究としては, ユーザの興味語に関する研究 [1], リスト機能に着目したユーザの属性に関する研究 [2] などが挙げられる.

4　提案手法

本研究では Twitter のユーザ間にできるフォローの関係を、ノードをユーザ, エッジをフォローの有無、重みをユーザ間の類似度とした重みつき有向グラフとして取り扱う.　本システムは入力を Twitter ユーザ, 出力を分類後のクラスタ集合としてクラスタリングを行う. まず入力を受け取ったシステムは TwitterAPI を利用し, フォロー集合とそのフォロー関係を取得する. 次にフォロー関係をグラフとし隣接行列を作成し, フォロー関係からノード間がどれだけ関係が強いか, を表す類似度行列を作成する. 最後にこの類似度行列を用いてクラスタリングを行い出力する. 以下に詳細について述べる.

隣接行列の作成方法について述べる. この隣接行列は本研究ではフォローの有無を表す行列となっている. ユーザを行, 列にとり, フォローの有無を 1, 0 で表す. 具体的には i 行 j 列の要素を, ユーザ i がユーザ j をフォローしていたら 1, フォローしていなければ 0 とする. またユーザ自身へのフォローを表す対角要素は 0 とする.

次に類似度行列の作成について述べる. ユーザが行, 列に対応しており, 値としては 1 が最も関係が強く 0 が関係がない状態を示す. 類似行列の i 行 j 列の要素を, ユーザ i とユーザ j が相互にフォローし合っていれば 1, どちらか一方向へのフォローのみが存在すれば 0.5, フォローがなければ 0 とする. 従ってこの類似度行列は対称行列となる. また同一ユーザに対する類似度 (つまり類似度行列の対角成分) は, 類似度行列の次元数や類似度の値の範囲に応じて十分な大きさの値を入れる.

最後に分類を行うスペクトラルクラスタリングについて述べる. スペクトラルクラスタリングは, 入力が類似度行列とクラスタ数, 出力が分類後のクラスタ集合である. この手法ではまずサブグラフ内が密, サブグラフ間が疎になるようなカットを出す評価関数を設定し, その関数の最適解がある固有値問題の解に対応していることで解くことによりカットを出す. これを複数回再帰的に呼び出すことによってクラスタリングを行う. 本研究では, 評価関数として $Min-maxCut(MCut)$ を用いた. グラ

フを A,B に分ける時の例を以下に示す. ここで $sim(a,b)$ はノード a とノード b の類似度を表す.

$$Mcut = \frac{\sum_{a\in A,b\in B} sim(a,b)}{\sum_{a\in A,b\in A} sim(a,b)} + \frac{\sum_{a\in A,b\in B} sim(a,b)}{\sum_{a\in B,b\in B} sim(b,b)}$$

5 実験

Twitter ユーザである被験者 5 人に依頼し, 実験を行った. 本実験の目的としてはこの手法を用いて理想の分類により近いものが出るかどうか, またこの結果を用いて今後どのように展開していくかを確かめることである. 内容としてはまず被験者に彼らにとってのフォロー集合の理想のクラスタリング結果を正解データとして作成してもらい, それとシステムの出すクラスタリング結果 (考えられうる全てのクラスタ数に関して分類した結果) とを比較して評価値の最も高かったものを出力とした. 評価値には次の *RandIndex* を用いた. クラスタリング結果を A と正解データを B と置き,A と B のそれぞれ全ての要素のペアについて同じクラスタに属するか否かのタグ付けを行う. *RandIndex* は, そのタグが A と B で一致する割合であり以下の式で求められる.

$$Rand\ Index = \frac{A \text{ と } B \text{ でタグが共通する個数}}{\text{全ユーザ数}C_2}$$

6 結果・考察

実験結果を Fig.1 に示す. 分類されたデータを見てみると「友人」や「同じ野球チームが好き」,「競技プログラミング」などのほとんどの属性においてかなり精度よく分類が行えていた. また精度よく行われた分類の特徴として分類されたフォロー集合内のユーザでフォロー数やフォロワー数が極端に多くないユーザが精度よく分類されていた. 逆にフォロー数やフォロワー数が極端に多いユーザ (代表的には芸能人や bot などであり, 以下, オーソリティユーザと呼ぶ) はあまり分類できていなかった. 具体的にどのような点ができていなかったかについて考察とともに述べる.

まず, オーソリティユーザが一つのクラスタとして分類されるという問題が見られた. この原因として多くの被験者が「有名人」のような属性でオーソリティユーザをひとくくりにまとめて正解データを作成していたが, このことが同じ属性を持つユーザは互いにフォローし合っているという仮定に反していたからであると考えられる. また結果の表においてユーザの理想の分類のクラスタ数と出力結果のクラスタ数に大きく差があるのはこの問題に起因していると考えられる.

二つ目の問題として, オーソリティユーザが一般ユーザのクラスタに含まれてしまう例がみられた. この原因としては, 本研究の類似行列ではどのフォローも同一の重さで扱っているが, この類似行列の単純さが起因していると考えられた. 例えば一般ユーザから一般ユーザへのフォローと有名人ユーザから一般ユーザへのフォローを比べたときに, 一般ユーザはフォロー数・フォロワー数ともに極端に大きくなく, 有名人はフォローが極端に少なくフォロワーが極端に大きいという傾向があるため, その場合上記のフォローは同じ重みとは考えにくい. よってフォロー数やフォロワー数に応じた類似行列を作成する必要性があると考えられる.

	被験者 I	被験者 II	被験者 III	被験者 IV	被験者 V
フォロー集合の要素数	13 人	55 人	208 人	47 人	226 人
出力結果の Rand Index	1.00	0.90	0.78	0.96	0.73
出力クラスタ数	8	29	20	13	14
正解クラスタ数の Rand Index	1.00	0.89	0.40	0.96	0.72
正解クラスタ数	8	9	10	13	16

図 1: 被験者 5 人に対する結果

7 まとめ

本研究では Twitter におけるフォロー関係からユーザのフォロー集合を分類する手法を提案した. フォロー集合を分類することに関して, フォロー情報のみだけでも望ましい分類に近い分類ができることが判明した. また結果からオーソリティユーザ以外のユーザはフォロー集合は同じ属性をもつユーザは互いにフォローし合っている, という仮定を満たしていることが分かった. 今後の研究の展望として, クラスタリングの自動化, 属性のラベリング, クラスタリング精度の向上をより少ない情報で実現することが考えられる.

参考文献

[1] 齋藤 準樹 and 湯川 高志. ソーシャルブックマークを基にした twitter ユーザの興味語抽出・推薦手法の提案と評価. 情報処理学会研究報告. 情報学基礎研究会報告, 2011(2):1–8, 2011-03-21.

[2] 奥川巧, 大石哲也, 長谷川隆三, 藤田博, 越村三幸, and 倉門浩二. Twitter のリスト機能を用いたユーザの特徴抽出. 全国大会講演論文集, 2011(1):687–689, 2011-03-02.