Twiiterユーザに対するスペクトラルクラスタリングとGN法によるクラスタリングの比較

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は，「Twitter[1]」に代表されるマイクロブログである．

本研究では，SNS としての機能に着目する．Twitter には，様々なユーザが存在し，多種多様なコミュニティを構成している．あるユーザが，どのコミュニティに属するのか，またそのコミュニティがどのような分野に興味があるのかが判明すれば，そのコミュニティに属するユーザに対して，その分野に関係のある情報の推薦，その分野に関係のない情報の遮断が可能である．そのためには，Twitter のユーザをクラスタリングすることが必要である．

これまで我々は，Twitter ユーザのクラスタリングのために，本学に関係するユーザを約 300 人収集し，そのユーザ群に対して，前述のフォロー・フォロワーという関係をもとに，凝集性クラスタリングを行い，ユーザの分類を行った．しかし，凝集型クラスタリングの問題点として，孤立したデータである外れ値やノイズが含まれるデータを扱えないこと，いったん形成されたクラスタはその後に分割できないなどが挙げられる．

そこで，この問題点を解決するため，クラスタリング手法の一つである，スペクトラルクラスタリングという手法に着目する．スペクトラルクラスタリングとは，クラスタリングをグラフの分割問題として扱い，最適なグラフの分割が，ある固有値問題の解から得られることを利用した非常にテクニカルな手法であり，高精度が期待できるクラスタリング手法である．この手法を前述の約 300 人から 100 人程度増やし，そのデータに対して適応して得られたクラスタと，凝集型クラスタリングによって得られたクラスタとを比較し，得られた知見について述べる．

[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. GN 法を用いた Twitter ユーザのクラスタリング

畑本らの研究[1]では，広島市立大学と思われる Twitter のユーザ約 300 人に対して，ソーシャルグラフを作成し，そのデータに対して凝集型クラスタリングを適用し，ユーザの分類を行っている．具体的なクラスタリング手法は，得られたデータに対して M.Girvan と M.E.J.Newman[2]によって提唱された GN 法を適用し，その後，Newman らが提唱したモジュラリティ指標を用いて，用途に対しての最適な分割数を求め，Twitter ユーザを分類している．本研究では，この凝集型クラスタリングを用いた手法が比較対象の一つであるため，3 章で詳しく述べる．

2.2. スペクトラルクラスタリングを用いた話者分類

磯の研究[3]では，録音音声中の発話の話者クラスタリング法として，ノンパラメトリックな発話クラスタ表現として量子化符号の出現頻度ベクトルを用いて，それらの間の余弦をクラスタ間類似度としてスペクトラルクラスタリングを行うことによる話者クラスタリング法を提案している．この手法の有効性を確認するため，磯は話者クラスタリングの一般的な手法として用いられているガウス分布による発話クラスタ表現，尤度比のベイズ推定量による類似度，階層的クラスタリングを組み合わせた手法とを提案手法と比較している．その結果，提案手法が従来手法よりも低い誤り率を与えることを確認している．本研究では，このスペクトラルクラスタリングを用いた手法も比較対象の一つであるため，3 章にて詳しく述べる．

3.提案手法

本研究の提案する手法は，広島市立大学の関係者と思われる Twitter のユーザ群に対して，スペクトラルクラスタリングと凝集型クラスタリングを行うことにより，Twitter のユーザを分類し，その違いと要因を明らかにすることによって，用途に応じたクラスタリングを可能とすることを目標とする．

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

3.1.スペクトラルクラスタリング

3.1.1.類似度行列

スペクトラルクラスタリングを行う場合，類似度行列を用意する必要がある．類似度行列とは，本研究の場合，ユーザ間がどれほど類似性を持っているかを表す行列となる．行列の類似性を示す指標として，コサイン類似度や構造同値性などが挙げられる．本研究では，広く用いられているコサイン類似度を採用する．まず，前述のフォロー・フォロワー関係を用いて隣接行列を作成する．隣接行列は，ノード間にリンク関係が存在すれば 1，存在しなければ 0 となる行列である．この隣接行列に対してコサイン類似度を算出し，類似度行列を求める．

$$W_{(i,j)} = \sum_{i,j} \frac{\sum_k (A_{(i,k)} * A_{(j,k)})}{\sqrt{\sum_k (A_{(i,k)})^2} * \sqrt{\sum_k (A_{(j,k)})^2}} \qquad (1)$$

$W_{(i,j)}$ ： 類似度行列，$A_{(i,j)}$ ： 隣接行列

3.1.2. スペクトラルクラスタリングのアルゴリズム

次に，類似度行列からラプラシアン行列を求める．グラフ理論においては，ラプラシアン行列とは，グラフの構造を行列で表現した行列で，対角成分で頂点の価数をその他の成分で隣接関係を表す行列である．今回は類似度行列に対してラプラシアン行列 L を求める必要があるので，次式にて算出する．

$$L = I - D^{-1/2} W D^{-1/2} \qquad (2)$$
$$W_{(i,j)} \geq 0 \, , \; D_{(i,j)} = \delta_{(i,j)} \sum_k W_{(i,k)}$$

次に，式(2)で求めたラプラシアン行列 L の固有値を求める．そして，その固有値の中で，小さい方から Q 個(ただし，固有値が 0 の時は除く)求める．

$$\lambda_1 \leqq \ldots \leqq \lambda_Q \qquad (3)$$

さらに，(3)にて求めた固有値に対する固有ベクトルを求める．

$$\{v_{\mathrm{iq}} \mid \mathrm{i} = 1, \ldots, N \, , q = 1, \ldots, Q\} \qquad (4)$$

次に，式(5)を用いて式(4)で求めたベクトルを規格化する．

$$Y_{(i,q)} = \frac{v_{(i,q)}}{\sqrt{\sum_q (v_{(i,q)})^2}} \qquad (5)$$

式(5)にて求められた規格化された行列 Y の N 個の行ベクトルを，k-means 法を用いて Q 個のクラスタに分類する．

$$C = \cup_{\alpha=1}^{Q} C_\alpha \; , \; \mathrm{C}_\alpha = \{\mathrm{i} \mid \mathrm{Y}_{(\mathrm{i,q})}\} \in \mathrm{C}_\alpha \qquad (6)$$

このようなアルゴリズムで，Twitter ユーザのスペクトラルクラスタリングを行う．

3.2.GN 法

3.2.1.エッジ媒介中心性を用いたクラスタリング

エッジ媒介中心性(edge betweeness)とは，社会学者である L.Freeman により提案された点媒介中心性なる指標を，グラフのエッジに対して適用した指標である．この指標は，あるエッジが頂点間の最短経路上にどの程度存在しているかを示す指標である．式(7)に示す．

$$eb = \sum_{s,t} \{ \frac{\sigma(x)_{s,t}}{\sigma_{s,t}} \} \qquad (7)$$

eb ： エッジ媒介中心性の値
$\sigma_{\mathrm{s,t}}$ ： ノード s,t 間の最短経路の総数
$\sigma(X)_{s,t}$ ： $\sigma_{\mathrm{s,t}}$の中であるエッジ Xを通る最短経路の総数

式(7)において算出される値が高ければ高いほど，そのエッジは多くのノードとノードをつなぐ働きを示す．つまり，この値が高いエッジを順番に切断することで，クラスタ内類似度を高く保ちつつ，クラスタ間類似度を下げながら分割していくことが可能である．ゆえに，比較的精度の高いクラスタを得ることが可能であると考えられる．

3.2.2 モジュラリティ

3.2.1.節にて，任意のステップ数でコミュニティを得ることを示した．しかし，任意のステップ数でコミュニティを得られたとしても，どのステップ数で得られたクラスタが，統計的意味を持つのかは不明である．そこで Newman らが提唱したモジュラリティという指標を用いてネットワークの分割を評価する．以下，モジュラリティの詳細を述べる．まず式(8)を説明する．

$$e_{i,j} = \frac{1}{2M} \sum_{s \in V_i} \sum_{t \in V_i} A_{(s,t)} \qquad (8)$$

M ： エッジの総本数
V_i ， V_j ： コミュニティ i , j
$A_{(i,j)}$ ： 隣接行列

コミュニティ i に属するノードと j に属するノードの間に張られるエッジの数がグラフ全体に張られるエッジに対する割合を $\mathbf{e}$=($\mathrm{e_{i,j}}$)とする．このときの $\mathbf{e}$ のトレースは式(9)にて表される．

$$Tr\, \boldsymbol{e} = \sum_s e_{s,s} \qquad (9)$$

この値は同一コミュニティ内部で張られたエッジの比率を表す．

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

しかし，この値だけでは全ノードが 1 つのコミュニティに属する時，最大値 1 をとるので，コミュニティ分割の評価には使用できない．そこでコミュニティ i に属する頂点につながる辺の比率を式(10)のように表す．

$$a_s = \sum_t e_{s,t} \tag{10}$$

これは **e** の行和と同義である．コミュニティに関係なく等確率でエッジを張ったときの期待値は式(11)で表せる．

$$e_{s,t} = a_s a_t \tag{11}$$

この(10)(11)式を用いて，モジュラリティ指標である Q 値は式(12)のように定式化される．

$$Q = \sum_{i \in .L} (e_{ll} - a_l^2) \tag{12}$$

Q 値はコミュニティ内のエッジが密であり，コミュニティ間のエッジが疎であるほど高い値となる．この Q 値を手がかりとして目的のクラスタを得る．なお，Newman らによれば Q 値は 0.3~0.7 の時，有意であると報告されている．

4.実験結果

3 章で述べたスペクトラルクラスタリング法と GN 法を用いて，本学の関係者のユーザ 400 人を分類した結果を以下の表 4.1 と表 4.2 に示す．

表 4.1： スペクトラルクラスタリング法適用時の結果

ID	Node	学部	学科	ラボ	入学年度
0	2	A : 2/2	F : 1/2	N : 1/2	Y : 1/2
1	43	A : 38/43	E : 10/43	R : 7/43	Y : 20/43
2	4	B : 4/4	I : 3/4	S : 1/4	Y : 2/4
3	3	C : 3/3			U : 3/3
4	50	B : 50/50	D : 32/50		Y : 19/50
5	6	C : 5/6	E : 1/6	T : 1/6	U : 3/6
7	35	B : 33/35	D : 20/35	S : 3/35	U : 24/35
8	13	C : 12/13			V : 13/13
9	70	A : 68/70	E : 53/70	O : 44/70	Y : 13/70
12	2	C : 1/2	I : 1/2	J : 1/2	Y : 2/2
13	10	C : 10/10			W : 7/10
15	17	C : 11/17	G : 1/17	L : 1/17	V : 6/17
17	16	A : 15/16	E : 14/16	L : 10/16	V : 6/16
18	10	A : 10/10	G : 10/10	P : 3/10	AA : 7/10
19	20	A : 18/20	F : 8/20		V : 18/20
20	29	C : 12/29	I : 5/29	J : 4/29	V : 11/29
21	20	A : 19/20	G : 9/20	M : 2/20	X : 14/20
22	15	A : 9/15	D : 1/15	K : 1/15	Y : 11/15
23	8	C : 7/8			Z : 4/8
24	2	C : 2/2			V : 2/2
26	8	A : 7/8	E : 4/8	K : 2/8	AB : 7/8

表 4.2: GN 法適用時のクラスタリング結果

ID	node	学部	学科	ラボ	入学年度
0	109	B : 80/109	D : 47/109	J : 6/109	U : 35/109
1	44	A : 25/44	E : 14/44	K : 8/44	V : 19/44
2	11	jh : 11/11	F : 6/11		V : 11/11
3	4	jh : 4/4			W : 3/4
4	15	A : 14/15	E : 13/15	L : 10/15	V : 6/15
5	8	C : 4/8	G : 1/8	L : 1/8	U : 3/8
6	14	A : 14/14	G : 7/14	M : 1/14	X : 12/14
9	5	C : 5/5			Y : 4/5
10	3	A : 3/3	F : 2/3	N : 2/3	X : 2/3
12	11	A : 6/11			V : 7/11
13	31	A : 31/31	E : 30/31	O : 30/31	Y : 10/31
14	2	A : 2/2	G : 1/2		U : 2/2
15	8	C : 7/8			Z : 4/8
20	9	A : 9/9	G : 9/9	P : 1/9	AA : 9/9
22	4	B : 3/4	D : 1/4		U : 2/4
25	3	A : 3/3	E : 3/3	Q : 2/3	AA : 2/3
31	5	A : 5/5	H : 4/5	R : 4/5	Y : 3/5
32	2	A : 2/2			U : 1/2
36	2	A : 2/2	E : 2/2	O : 2/2	
38	5	A : 5/5			Y : 5/5
40	3	A : 3/3	H : 2/3		AA : 3/3
42	8	C : 8/8			W : 6/8
56	6	C : 6/6			V : 5/6
57	2	A : 2/2			Z : 1/2
59	2	A : 2/2	G : 1/2		X : 1/2
69	2	A : 2/2	F : 1/2		V : 2/2
70	2	A : 1/2	E : 1/2		
81	2	B : 1/2	D : 1/2		Z : 1/2
83	2	A : 2/2	E : 2/2	O : 2/2	V : 2/2
86	3	A : 2/3			W : 2/3

表は，左からクラスタ ID，クラスタを構成しているノード数，以下，分類されたノードが持つ属性である．属性は，広島市立大学の関係者を対象としているので，学部，学科，配属されている研究室，入学年度の 4 つとした．属性の内訳は学部が A から C までの 3 個，学科が D から I までの 6 個，ラボが J から T までの 11 個，入学年度が U から AB までの 7 個である．

表の見方は，例えば表 4.2 のクラスタ ID が 0 の時，構成しているノードは 109 個，その中で最も付与された属性が多かった属性 B の人数が 80 個存在したということである．なお，この属性は人手で付与しており，可能な限り付与しているが，どの属性が適当であるか，判断がつかないノードも多く見受けられた．ゆえに，表中に極端に少ない属性が存在する．また，クラスタを構成しているノードが 1 つだった場合，本研究ではクラスタとはみなさずに除外している．

表 4.1 は，GN 法適用時の結果と比較するため，k-means のクラスタ数を 30 と設定した時の結果である．構成されたクラスタ数は，21 個であった．表 4.2 は GN 法適用時，Q 値が最も高かった 0.428 の時のクラスタリング結果である．総分割数は 95 で，構成されたクラスタ数は 30 であった．

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

5. GN 法とスペクトラルクラスタリング法の比較

表 4.1 および表 4.2 からクラスタリング結果における違いについて考察を行う. まず, クラスタを構成しているノードについて着目する. GN 法における結果では, クラスタを構成しているノードの個数が 2 個のクラスタから 109 個と, ばらつきが見受けられるのに対し, スペクトラルクラスタリング法における結果では, GN 法ほどのばらつきは見受けられない. これは, 各手法の違いが最も見受けられる違いであり, GN 法では, 近傍のノードをとにかく包含していくので, クラスタ数にばらつきが見受けられると考えられる. 対象にスペクトラルクラスタリング法においては, k-means 法を用いているので, クラスタを構成するノードは重心からの距離によるので, ばらつきはさほど見受けられないと考えられる.

次に, 学部における精度に着目する. ここにおける精度とは, 表中に示している, クラスタリングされたノードに対して, 最も付与された属性とする. GN 法を用いた場合の精度は概ね 70%程度であるのに対し, スペクトラルクラスタリング法を用いた場合の精度は概ね 80%を超えていることが分かる. これも, 各手法における違いが要因であると考えられる. GN 法では, 一度でもノードがあるクラスタに包含された場合, そのクラスタから離れることはない. ゆえに, 異なる属性を包含した場合, クラスタリングとして間違っていたとしても, 修正されることはない. それとは異なり, k-means 法では, 重心距離との関係から分類を修正しつつクラスタリングを行うので, 仮にクラスタリングとして間違った結果でも修正されることが多いので, 高い精度を得ることができたと考えられる. また, GN 方では, 隣接関係の 2 値のデータを使用しているのに対して, スペクトラルクラスタリング法では, 類似度行列を利用しているので, 扱っている情報量が多い. この理由からも, スペクトラルクラスタリング法を用いた結果の方が高精度であると考えられる.

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 400 人に対してスペクトラルクラスタリング法とGNを用いてクラスタリングし, そのクラスタリング結果の比較を行った. 結果, スペクトラルクラスタリング法を用いた方が良好な結果が得られることがわかった. 今後は, 付与する属性をさらに正確にすること, また今回行ったクラスタリング手法は, 各ノードが 1 つだけのクラスタに属するハードクラスタリングであるので, 今回用いたスペクトラルクラスタリング法を, 各ノードが複数のクラスタに属するソフトクラスタリング化させるなどが今後の課題である.

謝辞

この研究の一部は, 平成 24 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] 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.

[2] M.Girvan, M.E.J.Newman : “Community structure in social and biological networks,” PNAS, vol.99, no.12 p7821-p7826, 2002.

[3] 磯健一 :“ベクトル量子化とスペクトラルクラスタリングによる話者クラスタリング”電子情報通信学会論文誌, vol.J93-D, no.11 pp2467-pp2473, 2010.

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.