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クラスタの特徴語抽出

畑本典宣　　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(以下，SNS)としての機能も有する．

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し，またそのコミュニティがどのような分野に興味を持ってい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する情報の推薦，その分野に無関係な情報の遮断が可能である．そのためには，Twitterのユーザのクラスタリングが必要である．

今回，Twitterのユーザをクラスタリングするために，本学に関係するユーザを約300人収集した．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う．そして，そのクラスタ群が各々どのような分野に興味が有るのかを特定するため，各クラスタにおいて，そのクラスタを象徴するであろう単語すなわち特徴語を抽出する．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，学年などでクラスタが構成されているため，解析が容易であること，大規模なデータに比べて，より細かいレベルでの情報推薦や，フィルタリングが可能であることが挙げられる．

2. 関連研究

Twitter には，ツイートにアノテーションをするお気に入りという機能がある．眞野ら[1]らは，お気に入りに登録されたツイートを用いて，グラフクラスタリングという手法を適用し，ユーザを分類する手法を提案している．また，その分類されたユーザの嗜好を得る手法も提案している．しかし，本研究が対象とするユーザでは，お気に入り機能の使用率が低く，この手法は適用できないことがわかった．よって，本研究ではフォロー・フォロワーに着目する．

[1] http://twitter.com

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，凝集性クラスタリングを行うことにより，Twitter ユーザを分類する．そして，その分類されたクラスタがどのような性質を持っているのかを判断するために，そのクラスタに分類されたユーザが投稿したツイートを分析し，そのクラスタを象徴する語の抽出を行うことである．

3.1. クラスタリング手法

今回，本研究が用いるクラスタリング手法は， GN 法[1]と呼ばれる手法である．GN 法とは，グラフのエッジに対して媒介中心性なる値を求め，その値が一番高いエッジを切断しクラスタを発見するという手法である．以下に，この手法の詳細を述べる．

3.1.1. エッジ媒介中心性を用いたクラスタリング

エッジ媒介中心性は，ノードに対して提案された点媒介中心性なる指標をエッジに対して適用した手法である．これは，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している．以下にエッジ媒介中心性の算出式(1)を示す．

$$eb = \sum_{s,t} \left\{ \sigma(x)_{s,t} / \sigma_{s,t} \right\} \qquad (1)$$

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

式(1)において算出される値は，エッジ媒介中心性の値が高いほど，そのエッジは多くのノードをつなぐ働きをしていることを示している．以下にアルゴリズムを示す．

ⅰ）グラフにおけるエッジ媒介中心性を求める
ⅱ）エッジ媒介中心性が一番高いエッジを削除
　そのエッジの ID をスタックに push する
ⅲ） ⅰ)ⅱ)をすべてのエッジを削除するまで繰り返す
ⅳ) ID が入っているスタックを pop し，クラスタを形成
ⅴ） ⅳ)を任意のステップ数繰り返す

上記のアルゴリズムを用いれば，任意のステップ数でコミュニティを得ることが可能である．

3.1.2. モジュラリティ

3.1.1.節にて，任意のステップ数でコミュニティを得られることを示した．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(2)を説明する．

$$e_{i,j} = \frac{1}{2M} \sum_{s \in V_i} \sum_{t \in V_j} A(s,t) \quad (2)$$

M :エッジの総数
$V_{i,j}$:コミュニティ i , j
$A(s,t)$:隣接行列

コミュニティ i に属するノード j とコミュニティ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 $e = (e_{i,j})$ とする. なお, 隣接行列とはノード s とノード t を結ぶエッジが存在するときに 1, それ以外の成分が 0 の行列を示す.

このとき, e のトレース $\mathrm{Tr}\, e = \sum_s e_{s,s}$ 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.

しかしながら, この Tr e だけではすべてのノードが 1 つのコミュニティに属するとき, 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は式(3)のように表せる.

$$a_s = \sum_t e_{s,t} \quad (3)$$

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 $e_{s,t} = a_s a_t$ と表せる. この(2)(3)式を用いて, モジュラリティ指標である Q 値は式(4)のように定式化される.

$$Q = \sum_{l \in 1..L} \left(e_{ll} - a_l^{\,2} \right) \quad (4)$$

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 値は普遍的妥当性を持っているわけではないので, この Q 値は目的のクラスタを得る手がかりとして使用する. なお, Newman らは, Q 値は 0.3～0.7 の時に有意であると報告している.

3.2. 各クラスタからの特徴語抽出

この節では, 分類された各クラスタの特徴語を抽出する方法を述べる. 本研究では, クラスタ毎に, そのクラスタに属するユーザの 2010 年 4 月から 2010 年の 12 月までのツイートを収集し, その中から名詞語を収集する. その収集された名詞語に対して象徴度 S を求める. 以下にその詳細を述べる.

3.2.1. 象徴度 S の算出

象徴度 S の算出については, 以下の式(5)を用いた.

$$S_{t,c} = tf_{t,c} \times \log\left(\frac{N}{df} \right) \times C_{t,c} \quad (5)$$

S :語 t のクラスタ c における象徴度
tf :クラスタ c おいての語 t の出現回数
N :総クラスタ数
df :すべてのクラスタ中の語 t を含むクラスタ数
C :クラスタ c のユーザ中で語 t を含むユーザ数
ただし, ユーザ数が 1 のときに 0 とする.

式(5)より象徴度 S は, そのクラスタでの使用頻度が高く, かつ他のクラスタに属するユーザの使用頻度は低い語ほど高い値となる. 補正値である C は, あるクラスタにおいて使用頻度が高い語であっても, 使用しているユーザが 1 人だった場合, そのクラスタを象徴しているわけではないとみなし 0 という値にしている. よって, 式(5)によって算出される象徴度 S の高い語は, そのクラスタの象徴する語, すなわち特徴語を示している.

3.2.2. MeCab 辞書の改良

本研究では, ツイートの形態素解析を行うために MeCab[2]を用いた. 既存の MeCab には非常に多くの語が収録されている. しかし, マイクロブログで用いられている用語に関しては辞書に未登録の語が多い. これを未知語として取り扱う. 本研究では, 未知語への対策として, 1つは市川ら[2]が行った手法, もう1つは wikipedia のタイトル名[3]を追加する方法を用いる. この 2 つの未知語対策を施した MeCab を用いてツイートの形態素解析を行い名詞語の抽出を行う.

4. 実験

3 章で述べた手法を用いて, 実際に本学の関係者であるユーザをクラスタリングし, その結果,得られたユーザ群から特徴語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[4]を用いた.

4.1. クラスタリング結果

本学の関係者と思われる 306 人の Twitter ユーザのアカウントに対して, 例えばユーザ A がユーザ B をフォローしていたとすると, A から B 方向へのエッジが存在すると定義する. この定義をもとにユーザをノードとして, どのノードともエッジが張られなかったノードを除外する.

ここで,3 章にて述べたグラフのエッジに対して媒介中心性を算出し, その値が高いエッジを取り除く手法を用いて, クラスタリングした結果の例として, 初めて Q 値が 0.3 を超えるステップ数 160 のとき, ステップ数 240 のときのクラスタリング結果の図をそれぞれ図 4.1 に示す. なお, 左にある図がステップ数 160, 右にある図がステップ数 240 の図であり, 実線で囲まれたノード群がそれぞれ 1 つのクラスタを表している.

また, 3 章にて述べた Q 値が各ステップ数において, どのような数値になるのかを図 4.2 にて示す.

[2] http://mecab.sourceforge.net
[3] http://download.wikimedia.org/jawiki/latest/
[4] http://www.r-project.org/

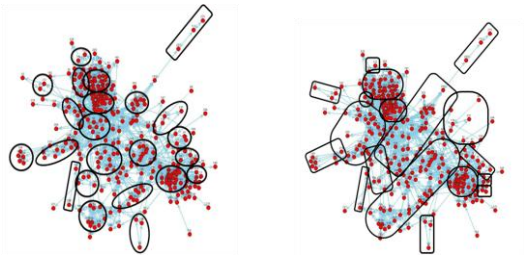

図 4.1:ステップ数 160 のときとステップ数 240 のときのクラスタリング結果

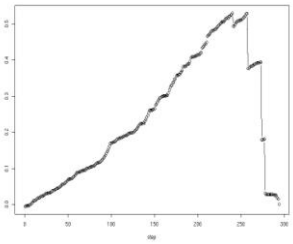

図 4.2:各ステップ数における Q 値のプロット

図 4.2 は，横軸がステップ数，縦軸が Q 値を示している．図 4.2 からステップ数が 240 のとき Q 値は最大値 0.530 である．Q 値が大きく下がっている部分は，クラスタ構成人数が多い 2 つのクラスタが融合し，1 つの大きなクラスタが構成されたことを示している．

次に各ステップ数における精度を示す．今回用いた指標は，学部，学科，入学年度である．以下に精度を算出する式(7)を示す．

$$P_{s,i} = \frac{1}{M_s} \sum_{M} \frac{\max\left(A_{N,i}\right)}{N_M} \qquad (6)$$

$P_{s,i}$:ステップ数 s においての指標 i における精度

M_s :ステップ数 s における総クラスタ数

N_M :クラスタ M を構成する人数

$A_{N,i}$: N の中で指標 i が合致する人数

式(6)は，あるステップ数で分類された各クラスタに対して，指標の要素が合致している人数が一番多い要素の割合を算出し，その平均を算出する式である．精度を算出した結果を図 4.3 に示す．

4.2. 特徴語抽出

この節では，図 4.1 のように分類されたクラスタから例として 4 つのクラスタに焦点を当てる．そのクラスタを A～D として特徴語を抽出した結果の例を表 4.1 に示す．表 4.1 は，2010 年 4 月から 12 月までの間を通してのデータである．しかし，抽出された語の中には短い期間に大量に投稿された語も含まれている可能性があり，4 月から 12 月の長期間のデータとしてはふさわしくない

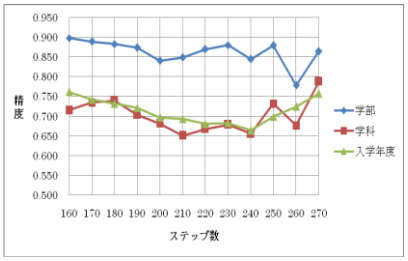

図 4.3:各ステップ数における精度

表 4.1 各クラスタにおける特徴語抽出

クラスタ A

語	S 値
工房	442.1
漆	205.0
作品	196.1
デザイン	137.8
金工	117.2

クラスタ C

語	S 値
練	34.6
ラン	15.9
トライアスロン	12.8
ローラー	12.0
浜	7.0

クラスタ B

語	S 値
ナッキー	185.1
プログラム	82.8
タグ	75.6
データ	58.1
論文	48.6

クラスタ D

語	S 値
尿	51.1
糞	25.8
あゆ	16.0
ゼミ	13.6
まり	9.1

可能性がある．そこで，月ごとにクラスタの象徴度である S 値を算出し，高い順に順位をつける．その順位をその語のスコアとし，もし対象となる月に存在しなかった場合，スコアを 10000 という高い値とする．その月ごとにスコアを算出し，その合計を求め，その語の最終的なスコアとする．つまり，スコアが小さいほどそのクラスタ内において平均的にツイートされているということを示している．語の最終的なスコアを以下の表 4.2 に示す．

表 4.2 月間ごとにスコア付けした語の最終的スコア

クラスタ A

語	スコア
工房	40
作品	55
バイト	115
自分	115
先生	163

クラスタ C

語	スコア
自分	30113
人	40036
雨	50054
先生	50065
バイト	50084

クラスタ B

語	スコア
汗	107
ゼミ	128
大学	244
自分	377
気	380

クラスタ D

語	スコア
学校	117
バイト	10214
自分	10457
先生	20147
人	20221

5.考察

5.1. クラスタリング結果について

5.1.1. クラスタとステップ数の関係について

図 4.1 からステップ数が大きくなると，クラスタが少しずつ大きくなり，やがて大きなクラスタになっていることがわかる．故に，ステップ数が小さいと，規模の小さいクラスタを得ることが可能で，逆にステップ数が大きいと規模の大きいクラスタを得ることが可能である．

しかし，ステップ数が小さすぎると，クラスタに包含されないノードが多く存在するようになる．逆にステップ数が大きすぎると，性質の違うクラスタ同士が結合する可能性が高くなる．そこで，モジュラリティ指標である Q 値を用いた．Q 値を使用する理由は，クラスタリングの結果，分割されたコミュニティが有意であるかどうかを判別する指標にできるからである．Newman らが，Q 値は 0.3～0.7 のときにクラスタリングを行った結果が有意であると報告していることは 2 章の 1.2.節に述べた．本研究の対象とするデータに対して，Q 値を算出すると 0.00～0.53 という値の範囲であった．本研究の手法では，Q 値が 0.3 以上になるのは約 160 ステップから約 270 ステップの間であることがわかる．

5.1.2. クラスタリングの精度について

次に，図 4.3 で示した精度を参照する．ここでいう精度の指標は，学部，学科及び入学年度のことである．学部に対しての精度は，概ねどのステップ数においても 80%を維持し，学部及び入学年度に対しての精度は，70%前後の値をとっていることがわかる．この精度がステップ数 160～270 の時にどう変動しているのかを確認する．まず Q 値が 0.3 を超えている範囲内の前半部，ステップ数が 160～240 の間は少しずつ下降していることがわかる．一方，Q 値が 0.3 を超えている範囲内の後半部，240～270 は概ね上昇傾向である．これはステップ数が 160～240 の間は指標が異なるノードを包含しながらクラスタが形成されていくからと考えられる．しかし，ステップ数が 240～270 の間は，形成されているクラスタそのものが少なくなっており，精度を算出する式の分母が小さくなるため，精度は上昇傾向であると考えられる．

なお，学部に対する精度が他の指標に対する精度より高い理由は，指標の種類が学科及び入学年度より 3 種類と少ないことが考えられる．

5.2. 特徴語の抽出について

表 4.1 は，各クラスタにおいて 4 月から 12 月の間に頻繁にそのクラスタに属するユーザのツイートに含まれており，かつ他のクラスタに属するユーザのツイートには含まれていなかった語を示している．また，表 4.2 は，4 月から 12 月の各月に頻繁にそのクラスタに属するユーザのツイートに含まれていて，かつ他のクラスタに属するユーザのツイートには含まれていなかった語を示している．ここで，クラスタについて説明すると，クラスタ A は芸術学部関係のクラスタ，クラスタ B は本学の言語音声メディア工学研究室のクラスタ,クラスタ C はトライアスロン部のクラスタ，クラスタ D は本学の情報科学部 4 年生のクラスタである．

表 5.1：各語の隔月に対する出現頻度

	APR	MAY	JUN	JLY	AUG
ゼミ	11	23	21	17	18
プログラム	2	2	8	16	22

表 4.1 と表 4.2 をクラスタごとに比較を行う．まず，クラスタ A について述べる．幾らか抽出された語の違いや順位の変動はあるが，大きな差異はない．よって，表 4.1 及び表 4.2 に記載しているクラスタ A から抽出された特徴語は，そのクラスタを象徴していると言える．

次に，クラスタ B について述べる．表 4.1 と表 4.2 をクラスタ A と比較すると,抽出された語に差がある.表 4.2 に示しているスコアから抽出された語はこのクラスタにおいては多くツイートに含まれている語であることは確かである．しかし，その抽出された語をみると，そのクラスタ B を象徴する語が少ないことが分かる．これは，スコアを算出するに当たって，表 5.1 に示している「ゼミ」などの語のように，月ごとに平均的にツイートされる語が，「プログラム」などのように，出現に偏りがある語よりも，式(5)に示している tf の値が各月で大きくなり，象徴度が平均的に高くなるためである．つまり，日常的にツイートされる語のスコアは高くなるためだと考えられる．ゆえに，抽出された語がそのクラスタを象徴しているとは言えない．

クラスタ C 及びクラスタ D については,そのクラスタで使用頻度の高い語の抽出することはできた.しかし,表 4.1 と表 4..2 で示している語の中には,そのクラスタを象徴しているような語が存在するものの,語の種類は異なり,各語のスコアも高いので,特徴語を抽出できたとはいえない.故に,これらのようなクラスタから特徴語を抽出するには,別の尺度からのアプローチが必要だと考えられる．

6.おわりに

本研究では，Twitter のユーザ群の凝集性クラスタリングを行い，さらにその分類された各ユーザ群からそのユーザ群を象徴する語の抽出を行った.その結果，クラスタリングに関する精度では，学部を指標にすると概ね 80%以上，学科及び入学年度を指標にすると概ね 70%前後の値になった．また，特徴語抽出に関しては，多くのクラスタからそのクラスタを象徴する語を抽出することに成功した．したがって，本研究の手法が妥当であることが確認された．

参考文献

[1] 眞野裕也，青山俊弘："ミニブログユーザの記事嗜好を用いたクラスタ発見，"日本高専学会誌，vol.15，no.3，pp.43-46，2010

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

[3] 市川博通，黒澤義明，目良和也，竹澤寿幸．"マイクロブログを用いた音声認識用モデルの構築及び分析，" 言語処理学会年次大会，2011.