マイクロブログにおけるクラスタリング技術の比較

粟　立軍† 廣田 雅春†† 横山 昌平††† 石川 博†††

† 静岡大学大学院情報学研究科 〒 432–8011 静岡県浜松市中区城北 3–5–1
†† 静岡大学創造科学技術大学院 〒 432–8011 静岡県浜松市中区城北 3–5–1
††† 静岡大学情報学部 〒 432–8011 静岡県浜松市中区城北 3–5–1
E-mail: †gs11304@s.inf.shizuoka.ac.jp, ††dgs11538@s.inf.shizuoka.ac.jp ,
†††{yokoyama,ishikawa}@inf.shizuoka.ac.jp

あらまし 近年，Twitter に代表されるマイクロブログが普及し，多くのユーザによって大量の短い文章の記事がマイクロブログに投稿されている．ブログに含まれる文章と比較して，それらの記事は，リアルタイム性が高い．それらの記事は，1 つの記事の文字数が少ないこと，記事に含まれる単語の表記揺れが多いこと，ノイズになるような記事が多いことも特徴としてあげられる．そのため，テキストに対する一般的なクラスタリング手法をマイクロブログに適用した場合，十分な精度が得られない場合がある．本論文では，これについて文章や Twitter のツイートのクラスタリングに用いられることがある LDA，Bisecting K-means，および階層型凝集法を Twitter のツイートに適用し，比較する．

キーワード マイクロブログ，Twitter, クラスタリング

A Comparison of Clustering Techniques on Microblog

Lijun SU†, Masaharu HIROTA††, Shohei YOKOYAMA†††, and Ishikawa HIROSHI†††

† Graduate School of Informatics, Shizuoka University
3–5–1 Jouhoku, Naka–ku, Hamamatsu, Shizuoka, 432–8011 Japan
†† Graduate School of Science and Technology, Shizuoka University
3–5–1 Johoku, Naka-ku, Hamamatsu-shi, Shizuoka, 432-8011 Japan
††† Department of Computer Science, Shizuoka University
3–5–1 Johoku, Naka-ku, Hamamatsu-shi, Shizuoka, 432-8011 Japan
E-mail: †gs11304@s.inf.shizuoka.ac.jp, ††dgs11538@s.inf.shizuoka.ac.jp ,
†††{yokoyama,ishikawa}@inf.shizuoka.ac.jp

Abstract Microblogging services such as Twitter have spread out recently. A huge amount of short texts are posted to these services by users. Compared with traditional documents like a blog, a feature of short texts is more high real-time property than those documents. However, it can also be listed as its features that it is short and noisy and its words and grammars tend to be informal. As these features, it is difficult to get a sufficient accuracy when we perform common clustering techniques to documents in microblogs. Focusing on this point, this paper presents the experimental results on clustering microblogs.

Key words microblog, Twitter, clustering

1. はじめに

近年，Twitter [1] に代表されるマイクロブログが普及し，多くのユーザにより大量の短い記事がツイートとして投稿されている．マイクロブログでは，ユーザが自身の状態や，様々な話題に関する情報を投稿する．Twitter のツイート数は，2012 年 11 月 16 日の発表において，2.5 日間に 10 億件以上，月間アクティブユーザ数は，2012 年 12 月 18 日の発表において，2 億人以上と発表されている(注1)．マイクロブログの特徴として，字数に制約がある場合が多く，Twitter の記事であるツイートの場合，140 文字以内という制約がある．投稿記事が短いことと，携帯端末から利用可能なアプリケーションの普及もあり，

(注1)：http://www.itmedia.co.jp/enterprise/articles/1212/20/news101.html

更新が容易であることが特徴である．ユーザが投稿する記事のトピックは，日常生活に関するものからニュースや，ユーザが視聴しているテレビ番組，参加している行事など多岐に渡る．従来のメディアと比較してこれらの記事は即時性が高く，マイクロブログの記事は，ユーザにとって有益な情報源となっている．それらの情報を利用するためマイクロブログを用いた研究が盛んである．マイクロブログに投稿される情報に基づいて，実世界の動向の予測 [2], [3] や，マイクロブログのトレンドを分析する手法 [4] などがあげられる．これらの研究において，ツイートをトピックに基づいてグループに分けるために，クラスタリング手法を適用することが多く，クラスタリング手法の研究の需要が高まっている．しかし，一般的なクラスタリング手法をマイクロブログに適用した場合，十分な精度が得られない場合がある．これは，ブログに記載されているような文章や，ニュース記事のような文章と比較して，ひとつの記事の文字数が非常に少ないこと，記事に含まれる単語の表記揺れが非常に多いこと，および分析にあまり有用でない記事が多い事などが原因としてあげられる．

近年，文書に適用するクラスタリング手法として，Latent Dirichlet Allocation(LDA) [5] が多くの研究に用いられている．LDA は，文書集合に含まれる文書をトピックに基づいてクラスタリングする手法である．また，文書クラスタリングなどで用いられる手法として，Bisecting K-means [6] や，階層型凝集法などがあげられる．しかし，これらの手法について，マイクロブログに適用する際，マイクロブログに投稿される記事の性質を考慮する必要がある．

そこで，本論文では，マイクロブログの中でも Twitter を対象として，LDA，Bisecting K-means，および階層型凝集法の 3 つのクラスタリング手法を適用し，クラスタリング精度を比較し，それぞれの手法によるクラスタリング結果について検討する．クラスタリング精度を評価するために，本実験では，Twitter から 3 つのトピックに関するツイートを収集し，データセットを用いて精度を評価する．

本論文の構成は以下のようになっている．2 章では，本論文でマイクロブログに適用する 3 つのクラスタリング手法について述べる．3 章では，3 つのクラスタリング手法を Twitter から取得した 3 つのデータセットに適用し，評価実験を行い，考察する．4 章では，本論文で得られた成果と今後の課題について述べる．

2. クラスタリング手法

本章では，本論文で比較する 3 つのクラスタリング手法についての説明を述べる．本論文では，文書クラスタリングやツイートのクラスタリング手法として用いられることが多い LDA，Bisecting K-means，および階層型凝集法の 3 つのクラスタリング手法を用いる．

2.1 LDA

LDA は，文書をトピックに基づいてクラスタリングする際，よく利用される手法である．それぞれのトピック (クラスタ) が確率に従って文書に含まれるとし，ひとつの文書は複数のトピックに関連しているとした確率的トピックモデルである．また，LDA は，記事が複数のクラスタに属することを許すソフトクラスタリング手法である．LDA の文書の生成過程を以下に示す．

（1） Poisson 分布に従い文書の単語数 N を選択

（2） Dirichlet 分布 $Dir(\alpha)$ に従い文書が持つトピックの重み割合ベクトル θ を選択

（3） 単語数（$n = 1, 2..., N$）だけ以下の (a)(b) を繰り返す

（a） 多項分布 Multinomial(θ) に従いトピック $topic_i$ を生成

（b） トピック $topic_i$ と β から単語 w_i を生成

ここで，α は，文書が各トピックに属する重みベクトル θ に関する Dirichlet 分布のパラメータを表す．また，β は，Dirichlet 分布に用いる単語が，あるトピックから生成された文書に含まれる確率 ϕ に関する Dirichlet 分布のパラメータを表す．ここで，LDA のパラメータを推定するための手法として，ギブスサンプリング [7] を用いる．

そして，推定したパラメータに基づいて，それぞれのツイートをトピックに割り当てる．本研究では，LDA をツイートに対してハードクラスタリングを行うために適用する．そのため，ツイートの θ に関する分布に基づいて，ツイートが属する 1 つのクラスタを決定する必要がある．本研究では，ツイート a の所属するクラスタ (トピック) は，θ_i が最大となるトピック $topic_i$ とする [8].

$$topic_i = \underset{i \in |topic|}{\arg\max}(\theta_i) \tag{1}$$

ここで，$|topic|$ は，トピック数を表す．また，θ_i は，ツイート a の θ に含まれる i 番目の値を表す．

2.2 Bisecting K-means

Bisecting K-means は，文書のクラスタリングなどに使われる k-means を改良したハードクラスタリング手法のひとつである．Bisecting K-means のアルゴリズムを以下に示す．

（1） データセットに対して，クラスタ数を $k = 2$ とし，k-means を実行し，2 つのサブクラスタを作成する

（2） クラスタに含まれるデータのペア間の距離の平均値が最も小さいクラスタを選ぶ

（3） (2) で選ばれたクラスタに対して，クラスタ数を $k = 2$ とし，k-means を実行し，2 つのサブクラスタを作成する

（4） 指定したクラスタ数になるまで，(2)，(3) を繰り返す

2.3 階層型凝集法

階層型凝集法は，階層型クラスタリングの代表的な手法である．階層型凝集法は，最初に，それぞれのツイートをひとつのクラスタとみなし，それぞれのクラスタ間の距離が最も近いクラスタのペアを凝集することを繰り返す．2 つのクラスタ c_a, c_b の距離 $distance(c_a, c_b)$ は，以下の式で求める．

$$distance(c_a, c_b) = \frac{\sum_{t_{xa} \in c_a} \sum_{t_{yb} \in c_b} cos(t_{xa}, t_{yb})}{|c_a| * |c_b|} \tag{2}$$

ここで，t_{xa}, t_{yb} は，それぞれ c_a, c_b に含まれるツイートを表す．また，$cos(t_{xa}, t_{yb})$ は，クラスタ c_a, c_b に含まれるツイー

表 1　Twitter から取得したデータセットの概要

データセット	ツイート数	名詞の種類数
天皇杯	1958	1662
コミケ	3468	4788
紅白	4617	4339
全ツイート	10043	8877

表 2　データセットのラベル数

データセット	ラベル数
天皇杯	282
コミケ	676
紅白	501
全ツイート	1459

ト t_{xa}, t_{yb} のコサイン類似度を表す．また，$|c_a|, |c_b|$ は，それぞれのクラスタ c_a, c_b に含まれるツイート数を表す．

3. 評 価 実 験

本論文では，Twitter から取得したツイートの集合に対して 3 つのクラスタリング手法を適用し，クラスタリング結果を比較評価する．クラスタリングを適用するデータセットを作成するため，Twitter Search API [9] を用いて，Twitter からツイートを収集する．今回は，4 つのデータセットについて実験を行う．特定の話題に関するツイートによるデータセットとして，サッカーの大会である第９１回天皇杯 (以下，天皇杯)，第８３回コミックマーケット (以下，コミケ)，第６３回 NHK 紅白歌合戦 (以下，紅白) に関するツイートをハッシュタグに基づいて収集した．それぞれのデータセットを収集するために用いたハッシュタグは，”♯天皇杯”，”♯C83”，”♯NHK 紅白”である．また，複数の話題が含まれるツイートの集合に対して，クラスタリング精度の比較評価を行うために，3 つのデータセットに含まれる全ツイートを結合したデータセットを作成した．

データセットに含まれるツイートにクラスタリング手法を適用するため，それぞれのツイートに対して形態素解析を行う．ツイートの形態素解析は，Kuromoji [10] を用いた．このとき，URL は，形態素解析を行う対象から除外する．解析結果の形態素の中から，[副詞可能名詞，代名詞，サ変可能名詞，形容詞可能名詞，数字] を除く名詞を抽出する．ここで，それぞれのデータセットのツイート数とそのデータセットから得られた名詞の種類数を，表 1 に示す．

それぞれのクラスタリング手法で用いる単語ベクトルや，ツイート間の類似度の計算について述べる．LDA を適用する際，それぞれのツイートの単語ベクトルは，データセットに含まれる単語を含むツイート数に基づいて求められる．また，Bisecting K-means と階層型凝集法について，ツイート間の類似度を用いる．はじめに，ツイートの特徴量として，ツイートから得られた名詞を用いて IDF を求める．全ツイートに含まれる名詞 $W = \{w_1, w_2, ..., w_N\}$ のある単語 w_i の IDF を以下の式で求める．

$$idf_{w_i} = \log \frac{D}{n_{w_i} + 1} \tag{3}$$

ここで，w_i は W に含まれる i 番目の単語，D は全ツイート数，n_{w_i} は単語 w_i が含まれるツイート数とする．そして，IDF に基づいて，ツイート a の特徴量 $\vec{a} = \{t_{1a}, t_{2a}, ..., t_{Na}\}$ を以下の式で求める．

$$t_{ia} = \begin{cases} idf_{w_i} & w_i \in W_a \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

ここで，t_{ia} は，ツイート a における単語 w_i の特徴量，W_a は，ツイート a に含まれる単語の集合である．ツイート間の類似度として，それぞれの単語の IDF に基づいて，コサイン類似度を求める．ツイート a と b の類似度をコサイン類似度に基づいて，以下の式で求める．

$$cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} \tag{5}$$

これらの類似度を用いて，Bisecting K-means と階層型凝集法をデータセットに対して適用する．

3.1 評 価 指 標

本実験では，3 つのクラスタリング手法を適用したそれぞれのクラスタリング結果を評価するため，データセットに含まれるツイートに対してラベルを付与し，正解例を作成した．それらのラベルは，日常的に Twitter 利用している 4 名の実験協力者により作成した．ラベルを付与したツイートは，前述したデータセットに含まれるツイートの 10043 件である．実験協力者は，それらのツイートに対して，1 つの任意のラベルをそれぞれのツイートに付与した．この際，ラベルを付与する際，評価者によってラベルの表記揺れが発生しないように調整を行った．また，ハッシュタグや，ユーザなどにとらわれず，ツイートの内容のみに基づいてラベルを付与するように説明を行った．実験協力者により作成されたラベルの種類数，表 2 に示す．本実験では，このラベルの種類数を正解例のクラスタ数とする．表 3 に，実験協力者により作成されたラベルとツイートの例を示す．本実験では，Fowlkes-Mallows index [11] と variation of information [12] を評価指標として用いた．

Fowlkes-Mallows index は，再現率，適合率を用いた評価を行う．クラスタリング結果と正解例を C，C' とし，すべてのツイートのペアを表 4 のクラスに分ける．以下のように再現率，適合率を求める．

$$W_I(C, C') = \frac{N_{11}}{N_{11} + N_{01}} \tag{6}$$

$$W_{II}(C, C') = \frac{N_{11}}{N_{11} + N_{10}} \tag{7}$$

これらの式を用いて Fowlkes-Mallows index の値を求める．この値が大きいほど 2 つのクラスタリング結果は類似している．

$$FM(C, C') = \sqrt{W_I(C, C') W_{II}(C, C')} \tag{8}$$

Variation of information は相互情報量とエントロピーを用いた評価を行う．エントロピーを求めるため，ランダムに選択したあるツイートがクラスタ k に属する確率を求める．

$$P(k) = \frac{n_k}{n} \tag{9}$$

表 3　ツイートとラベルの例

ツイート	ラベル
【天皇杯】決勝戦 ガンバ大阪 vs 柏レイソル選手入場つっ !!　#天皇杯 #tennouhai	選手入場
無事帰宅。コスプレイヤーさんはみんな可愛いし格好良かった。生きるスタンス的な意味合いも込めて #C83	コスプレ
堀北真希が着物から白いドレスに衣装チェンジしてる！ #NHK 紅白	堀北真希

表 4　Fowlkes-Mallows index におけるツイートのペアのクラス

N_{11}	ツイートのペアが C，C' の両方で同じクラスタに属している
N_{10}	ツイートのペアが C では同じ，C' は異なるクラスタに属している
N_{01}	ツイートのペアが C では異なる，C' は同じクラスタに属している
N_{00}	ツイートのペアが C，C' の両方で異なるクラスタに属している

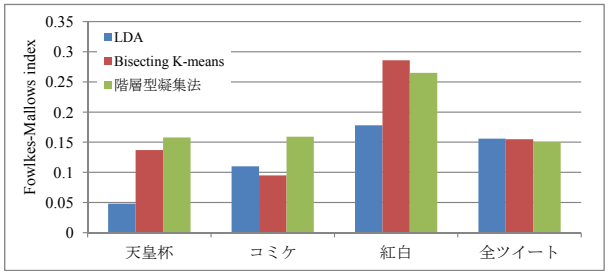

図 1　Fowlkes-Mallows index による各クラスタリング手法の評価値

ここで，n_k はクラスタリング結果 C のクラスタ k に含まれるツイート数，n は全ツイート数である．次に，以下の式でクラスタリング結果 C に関するエントロピーを求める．

$$H(C) = -\sum_{k=1}^{K} P(k) log P(k) \tag{10}$$

次に，2 つのクラスタリング結果 C, C' 間の相互情報量を求める．そのため，ランダムに選択したあるツイートがクラスタリング結果 C_k に属し，クラスタリング結果 C'_k に属している確率を求める．

$$P(k, k') = \frac{|C_k \bigcap C'_k|}{n} \tag{11}$$

クラスタリング結果 C, C' 間の相互情報量を $I(C, C')$ とし，以下の式で表す．

$$I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k) P'(k')} \tag{12}$$

これらを用いて，variation of information の値 $VI(C, C')$ を以下の式で求める．

$$VI(C, C') = [H(C) - I(C, C')] + [H(C') - I(C, C')] \tag{13}$$

Variation of information はあるツイートのそのクラスタ間の関連に着目し，2 つのクラスタリング結果のその関連性の違いを評価している．そのため，variation of information は値が小さいほど 2 つのクラスタリング結果は類似している．

3.2　実 験 結 果

Fowlkes-Mallows index [11] と variation of information [12] による評価結果をそれぞれ図 1，および図 2 に示す．それぞれの図は，3 つのクラスタリング手法を 4 つのデータセットに対して適用し，そのクラスタリング結果に対して評価指標を適用

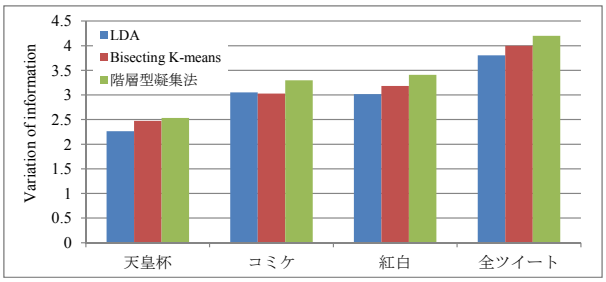

図 2　Variation of information による各クラスタリング手法の評価値

した結果である．Variation of information によるクラスタリング結果の評価において，LDA は，全てのデータセットに対して，最も高い評価である．Fowlkes-Mallows index によるクラスタリング結果の評価において，全ツイートを用いたデータセットに対して，LDA は高い評価である．また，それ以外の 3 つのデータセットにおいて，2 つのデータセットについて，階層型凝集法が，1 つのデータセットについて，Bisecting K-means が良い評価である．

LDA のクラスタリング結果の評価において，Fowlkes-Mallows index の評価が低い理由として，LDA の結果に基づいて，それぞれのツイートをクラスタに割り当てる際，式 (1) を用いたことが原因のひとつとしてあげられる．複数のクラスタが式 (1) を満たす場合，適切でないクラスタが選択される場合がある．図 3 に，それぞれのデータセットにおいて，1 つのツイートにして，2 つ以上のクラスタが式 (1) を満たすツイートの割合を示す．図 3 において，天皇杯データセットが最も割合が高く，約 80%のツイートは複数のクラスタが式 (1) を満たす．これは，天皇杯データセットは，名詞の種類が少なく，それぞれのクラスタの単語の特徴を十分に表すことが困難だったと考えられる．また，他のデータセットは，約 60%の割合である．そのため，LDA を用いてハードクラスタリングを行う場合，クラスタの割り当てを改善する必要がある．LDA は，大量の文書集合から生成される単語分布に基づいたクラスタリング手法であるため，本実験のように文書集合は多い．それぞれの文書に含まれる単語数が少ない場合，それぞれのトピックに対して十分な分布を作成する事ができない場合があると考えられる．また，ツイート数が少ない 3 つのデータセットにおいて，Fowlkes-Mallows index の評価が低く，ツイート数が多い全ツイートを用いたデータセットにおいて，Fowlkes-Mallows index の値は改善したと考えられる．これらのことより，LDA をマイクロブログに適用する場合，ツイート数と名詞の種類数が精度に影響を与えると考えられる．

コミケデータセットに対する Fowlkes-Mallows index の評価において，Bisecting K-means は低い評価である．これは，紅

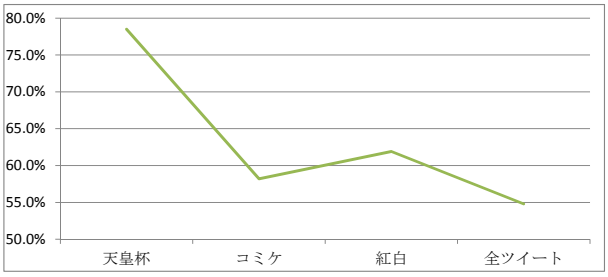

図 3　複数のクラスタが式 1 を満たすツイートの割合

白データセットと天皇杯データセットと比較して，コミケデータセットに含まれる名詞の種類数とクラスタ数が多いことが原因としてあげられる．同一のクラスタに属するべきツイートが共通の単語を含む確率が低く，そのような場合はそれぞれのクラスタ間の境界が曖昧な結果として，Bisecting K-means の精度が低下したと考えられる．

階層型凝集法と比較して，Bisecting K-means のクラスタリング結果の評価は，variation of information の評価において，全てのデータセットで高い．また，Fowlkes-Mallows index において，Bisecting K-means のクラスタリング結果の評価は，2 つのデータセットで高い．これは，それぞれのツイートが全体的に文字数が少ないため，2 つのツイート間の類似度を求める際，それらのツイートが共通する 1 つ単語を持つ場合，それらのツイート間の類似度が高くなりやすい．結果として，共通する単語数が少ない 2 つのツイートの類似度が高くなり，異なるトピックに属するべき 2 つのツイートが同一のクラスタに割り当てられる場合があると考えられる．一方，階層型凝集法と比較して，Bisecting K-means 法は，それぞれのクラスタの重心に基づいてツイートをクラスタに割り当てるため，このような精度の低下が発生しにくいと考えられる．

4. おわりに

本論文では，マイクロブログから取得したツイートに対して，複数のクラスタリング手法を複数のデータセットに対して適用し，比較評価を行った．Variation of information の評価において，LDA が有効であることを確認した．今後の課題として，マイクロブログのデータに対して，LDA をハードクラスタリング手法として用いており，LDA の結果として得られるトピックの分布から適切なトピックを選択するため，式 (1) を改善することがあげられる．また，マイクロブログから取得したツイートに対して，LDA などの複数のソフトクラスタリング手法を適用し，比較評価を行う予定である．

文　献

[1] “Twitter”, http://twitter.com/.
[2] 山中努, 田中祐也, 土方嘉徳, 西田正吾 : “時空間情報を伴うテキストデータを用いた状況把握支援システム”, 知能と情報 : 日本知能情報ファジィ学会誌 : journal of Japan Society for Fuzzy Theory and Intelligent Informatics, **22**, 6, pp. 691–706 (2010).
[3] 藤坂達也, 李龍, 角谷和俊 : “実空間マイクロブログ分析による地域イベントの影響範囲推定”, データ工学と情報マネジメントに関するフォーラム (DEIM2010), pp. D7–4 (2010).
[4] L. Hong, O. Dan and B. D. Davison: “Predicting popular messages in twitter”, Proceedings of the 20th international conference companion on World wide web, WWW ’11, New York, NY, USA, ACM, pp. 57–58 (2011).
[5] D. Blei, A. Ng and M. Jordan: “Latent dirichlet allocation”, the Journal of machine Learning research, **3**, pp. 993–1022 (2003).
[6] M. Steinbach, G. Karypis and V. Kumar: “A comparison of document clustering techniques”, In KDD Workshop on Text Mining (2000).
[7] T. Griffiths and M. Steyvers: “Finding scientific topics”, Proceedings of the National Academy of Sciences of the United States of America, **101**, Suppl 1, pp. 5228–5235 (2004).
[8] Y. Liu and F. Liu: “Unsupervised language model adaptation via topic modeling based on named entity hypotheses”, Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference onIEEE, pp. 4921–4924 (2008).
[9] “Twitter search api”, https://twitter.com/search.
[10] “kuromoji”, http://www.atilika.org/.
[11] E. Fowlkes and C. Mallows: “A method for comparing two hierarchical clusterings”, Journal of the American statistical association, **78**, 383, pp. 553–569 (1983).
[12] M. Meilă: “Comparing clusterings—an information based distance”, J. Multivar. Anal., **98**, 5, pp. 873–895 (2007).