Twitter上で行われる議論構造可視化のための段階的クラスタリングに関する検討

与儀 涼子　　當間 愛晃　　赤嶺 有平　　山田 孝治　　遠藤 聡志

琉球大学大学院理工学研究科 情報工学専攻
琉球大学工学部情報工学科

{k108589, tnal, yuhei, koji, endo}@ie.u-ryukyu.ac.jp

1　はじめに

近年、Ustreamで会議の様子を中継し、専用のハッシュタグを周知することで広くTwitterからの意見を求めるなど、実世界での議論の場にソーシャルストリームを活用するケースが増えている。特に、ソーシャルストリームの一つであるTwitterは、ハッシュタグやRetweetといった機能を持つことから、議論に併用される事が多くなっている。この事から、本研究ではTwitterを研究対象とした。

このような、ソーシャルストリームを併用した議論では、会場へ赴くことなく議論に参加できるという大きなメリットがある。しかし、ソーシャルストリームでの参加人数が増えるほど、短時間に膨大な数の発言が行き交う事になり、その中から人手で有用な意見を抽出する事、話題の推移を把握する事は困難になる。結果、ソーシャルストリームを併用しているにも関わらず会場での参加者のみの意見交換に終始してしまう事もあり得る。

Twitterにおいては、タイムラインの話題を手動で編集するTogetterというサービスが存在するが、やはり議論の関係者が多くなるほど人手での編集コストは高くなる。

そのため、何らかの方法でソーシャルストリームからの有用な意見を拾って実世界の議論にフィードバックすることが望まれる。そのような議論進行支援の技術として、以下の要求を満たす必要があると考えられる。

1. 特定の発言が、どのような役割の発言なのかを同定する
2. 特定の発言が、どの話題に属するかを同定する
3. 特定の発言の重要性を推定する
4. 上記1,2,3を議論と同時進行的に処理する

本稿では、上記2の、特定の発言がどの話題に属するかを自動で分類するというタスクについて検討する。一見、発言内容を形態素解析にかけて特徴ベクトルを生成するなど、文書分類の技術を使うことで容易に対処できる問題に思えるが、今回研究対象とするTwitterは、1エントリーの文字数が140文字に制限されているために特徴次元がスパースになりやすく、この方法だけでは上手く話題毎にクラスタリングする事が難しい。

この問題を解決するため、「ツイート生起時刻が近ければ同じ話題に属する度合いが高い」という仮説に基づき、形態素解析によって得た特徴ベクトル間距離を、ツイートの生起時刻から計算した時間影響度によって調整する。この距離を用いてクラスタリングを行うことで、スパースな特徴ベクトル間の距離を補正することができ、ここで得られたクラスタは、特定の話題を代表する特徴を十分に含んでいると考えられる。

そこで、得られたクラスタ群から特徴ベクトルを抽出し、再度クラスタリングにかける事でツイートを話題毎にクラスタリングできるものと考えている。

2　関連研究

生起時刻を持つ時系列文書に関する研究として、菊池ら[1]の研究が挙げられる。菊池らは、時系列情報を持つ文書から、話題の推移を表現したキーワードを抽出する手法を提案し、電子番組表(EPG)に適用して有用性を検証している。戸田ら[2]の研究では、文書集合のマイニングにおいて、文書内容の類似度に加えて時間的近さを考慮し、話題構造をグラフ化する手法を提案している。

Twitterに関連した研究は近年盛んに行われている。松村ら[3]の研究では、影響伝播モデルIDMをTwitterに応用することで、特定のメッセージやユーザの影響量およびユーザプロファイルの推定やその相互関係を可視化している。また、Twitterに関連する時系列特徴を利用した研究としては、高村ら[4]の研究がある。高村らは、Twitterの時間的特徴に着目し、単一のイベントに対するツイートを要約する手法を提案している。本研究では「実世界上での議論にTwitterを併用する」事を想定している点で異なっている。

3 提案手法

3.1 概要

本手法は、ツイートを形態素解析して得た特徴ベクトルに、ツイートの生起時刻から計算した時間影響度を加味してクラスタリングを行い、得られたクラスタ群から特徴を抽出して再度クラスタリングを行う、という手順を取る。その処理の流れを図 1 に示す。

Step1. ツイートから特徴ベクトルを生成
ノイズを除去した後、ツイートを形態素解析にかけて名詞ベクトルを生成する。

Step2. 時系列特徴を考慮したクラスタリング
Step1 で得た名詞ベクトルから各ツイート間の距離を求め、時間減衰関数を加味した後、階層的クラスタリングにかける。

Step3. 各クラスタから特徴ベクトルを抽出して再クラスタリング
Step2 で得た各クラスタから特徴ベクトルを抽出し、再度クラスタリングにかける。

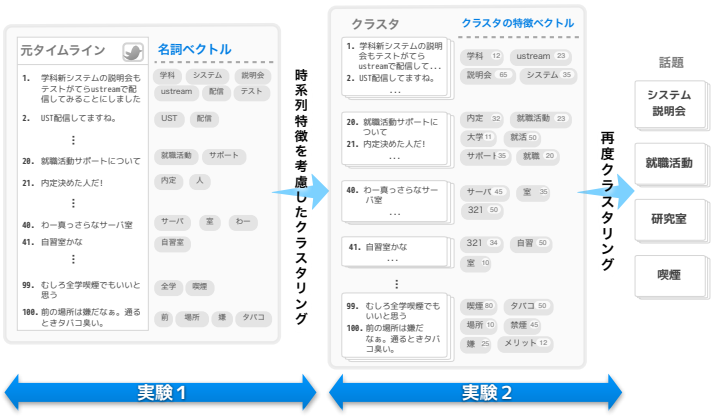

図 1: 提案手法の処理の流れ

3.2 ツイートから特徴ベクトルを生成

各ツイートの類似度を求めるための元情報として、ツイートの名詞ベクトルを利用する。名詞ベクトルは以下のようにして取得する。

1. 形態素解析用辞書を作る
   3. の形態素解析に用いるための辞書を用意しておく。
   Twitter 用語辞書 Twitter 独自の用語に対応するため、Twitter 用語集 [6] に記載されている用語を登録した辞書。
   ユーザ名辞書 Twitter を通して議論に参加しているユーザのスクリーンネームを登録した辞書。
   URL 辞書 議論のタイムラインに挙がった URL を登録した辞書。
   複合語辞書 「情報工学科総会」などの、その議論固有の複合語を登録した辞書。タイムラインに挙がった各語の組み合わせについて連結度を C-value 法 [7] で求め、閾値を超える組み合わせを複合語とみなした。
2. 別のツイートに対する Retweet である場合は除外する
   Retweet を通常のツイートと同様に扱うと、元ツイートとその Retweet のみから構成されるクラスタが乱立してしまい話題毎にクラスタを作る際の阻害要因となる。これを避けるため、データセット中の別のツイートに対する Retweet はクラスタ分析の対象から外す。
3. 形態素解析にかけ、名詞のみ抽出する
   上記 1. の辞書を利用して形態素解析にかけ、名詞を取得する。
4. 得られた名詞のクリーニング
   得られた名詞から、以下の単語を除外する。
   記号
   平仮名のみから構成される 2 文字以下の単語
   その議論を特定するためのハッシュタグ
   会議中継用の Ustream の URL
5. 各名詞に重み付けをし、名詞ベクトルを取得する
   「ツイート中でその名詞が生起した回数」を重みとして名詞ベクトルを作る。

3.3 時系列特徴を考慮したクラスタリング

3.2 で生成した名詞ベクトルと、各ツイートの生起時刻を利用してクラスタリングを行う。

ツイート間の距離を計算する手法としては、ユークリッド距離を用いる。ツイート a とツイート b の距離 d は以下の式 (1) によって表される。ただし X は各ツイートの名詞ベクトル、n はベクトルの次元数である。

$$d(X_a, X_b) = \sqrt{\sum_{i=1}^{n}(X_{ai} + X_{bi})^2} \qquad (1)$$

ここで、得られたツイート間距離を、ツイートの生起時刻の近さによって補正する時間減衰関数を導入する。この時間減衰関数は、「ツイート間の時間的距離が一定量離れる毎に、一定の割合で類似度を減ずる」ものとし、時系列文書を扱う既存の研究 [1] でもしばしば用いられる指数関数モデルを使って定義する。ツイート a とツイート b の距離に重み付けをする減衰関数を以下の式 (2) に示す。t_i はツイートの生起時刻、α は定数を表す。

$$W(a, b) = 1 - \exp(-\alpha(t_a - t_b)^2) \qquad (2)$$

$W(a,b)$ を使って補正した、最終的なツイート間距離は以下の式で定義する。

$$d'(X_a, X_b) = d(X_a, X_b) \quad W(a,b) \tag{3}$$

上記の方法で得られた、生起時刻を考慮した距離を元に、凝集型の階層型クラスタリングによってツイートを話題毎に分類する。クラスタを併合する際、クラスタ間の距離は Ward 法によって求める。Ward 法では、クラスタ C_1 とクラスタ C_2 の距離 d は以下の式 (4) によって表される。

$$d(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2) \tag{4}$$

ただし、クラスタ C_i の質量中心を c_i としたとき、

$$E(C_i) = \sum_{X \in C_i} (X + c_i) \tag{5}$$

である。

3.4 各クラスタの特徴ベクトルを抽出した上での再クラスタリング

3.3 で得た各クラスタについて、特徴ベクトルを取得し、再度クラスタリングにかけることで、最終的な話題クラスタを得る。

クラスタの特徴ベクトル（代表ベクトル）は、そのクラスタの内包するツイートの特徴ベクトルを足し合わせ、ベクトル長で正規化したものとする。計算式は以下の式 (6) で表される。C_i はそれぞれのクラスタを、X はクラスタの代表ベクトルを、x はクラスタに含まれるツイートの特徴ベクトルを表している。

$$X_{C_i} = \frac{\sum_{x \in C_i} (x)}{|\sum_{x \in C_i} (x)|} \tag{6}$$

この特徴ベクトルどうしの距離を計算する手法は、3.3 と同じくユークリッド距離を用い、クラスタリングも同様に Ward 法を用いて行う。

4 評価実験

提案手法の「時系列を考慮したクラスタリング」「各クラスタの特徴ベクトルを抽出した上での再クラスタリング」によって、話題クラスタを得る実験を行う。

4.1 データセット

提案手法の評価実験には、2011 年 4 月 20 日に琉球大学工学部情報工学科で行われた第二回情報工学科総会 [5] における専用ハッシュタグ#ieryukyu 内でのツイート 577 件 (Retweet を除くと 538 件) をデータセットとして利用した。このイベントは、本稿で研究対象としているような、「実世界上の特定の場所に参加者が集合して議論を行い、その様子をインターネットで中継し、Twitter を通して不特定多数の個人からの意見を受け付ける」という会議である。

4.2 実験 1:時系列特徴を考慮したクラスタリング

時系列特徴を利用したクラスタリングによるクラスタリング精度と、得られた各クラスタが、クラスタ中の話題を代表する特徴を含んでいるかどうかを調査する。実験は 4.1 を用いて行った。クラスタ数は、Pseudo-F 値 [8] を参考にし、また目視でも適当と思われた 14 に設定している。

4.2.1 クラスタリング結果

各クラスタに分類されたツイートを見て、そのクラスタでの主な話題を人手で判断した。各クラスタでの主な話題は以下のようになった。

No.	話題	No.	話題
1	システム説明会	8	就職活動
2	システム説明会	9	就職活動
3	システム説明会	10	部屋の使い方
4	システム説明会	11	研究室配属
5	システム説明会	12	研究室配属
6	就職活動	13	研究室配属
7	就職活動	14	喫煙問題

4.2.2 クラスタリング精度の評価

各クラスタに含まれるツイートを 1 件ずつチェックし、クラスタリングの精度を調査した。各ツイートが正しいクラスタに属しているか否かは、以下の基準により人手で判断した。

正解 ツイートの内容とクラスタの話題が一致しており、正しいクラスタに属している。

不正解 ツイートの内容が別のクラスタの話題に関するものであり、間違ったクラスタに属している。

ノイズ ツイート内容がどのクラスタの話題とも関係の無いものである。

結果を以下に示す。

種類	件数	割合
正解	423 件	79%
不正解	18 件	3%
ノイズ	97 件	18%
計	538 件	—

間違ったクラスタに分類されたツイートは全体の 3%程度だった。どのクラスタの話題とも関係のない、ノイズが 18%存在している。

4.2.3 各クラスタから抽出した特徴ベクトルの評価

各クラスタに含まれるツイートの特徴ベクトルから、そのクラスタで話されている話題を代表するような特徴を抽出できるかを確認する。各クラスタに含まれる特徴次元を足しあわせ、その上位 5 件を抜き出した結果を以下に示す。

No.	話題	得られた特徴
1	システム説明会	総会, 学科, 説明, 会, 前向き
2	システム説明会	IP, 6, 資料, 説明, 設定
3	システム説明会	学科, 情報, ssh, ie, 提供
4	システム説明会	コア, サーバ, 1, VM, CPU
5	システム説明会	登録, 学科, シンクライアント, 利用, Windows
6	就職活動	就職, 就活, 一, 最寄り, 万
7	就職活動	就職, 就活, 履歴書, サポート, 年
8	就職活動	笑顔, 就活, サポート, 就職, 大学
9	就職活動	就職, 就活, 大学, 駄目, 気
10	部屋の使い方	321, B, 1, ゴミ箱, 404
11	研究室配属	配属, プログラミング, 1, 研究, 人
12	研究室配属	先生, 人, 先生枠, 1, 枠
13	研究室配属	研究室, 学生, 室, 研究, 単位
14	喫煙問題	喫煙, 喫煙者, 場所, メリット, 者

同一の話題に関するクラスタ間で、共通する特徴が見られる。

4.3 実験 2:各クラスタの特徴ベクトルを抽出した上での再クラスタリング

4.2 で得た各クラスタについて、特徴ベクトルを取得し再度クラスタリングにかけ、最終的な話題クラスタを得る実験を行う。

4.3.1 クラスタリング結果

クラスタ数は、Pseudo-F 値を参考にし、また目視でも適当と思われた 6 に設定している。以下の図 2 に得られた話題クラスタのデンドログラムを示す。

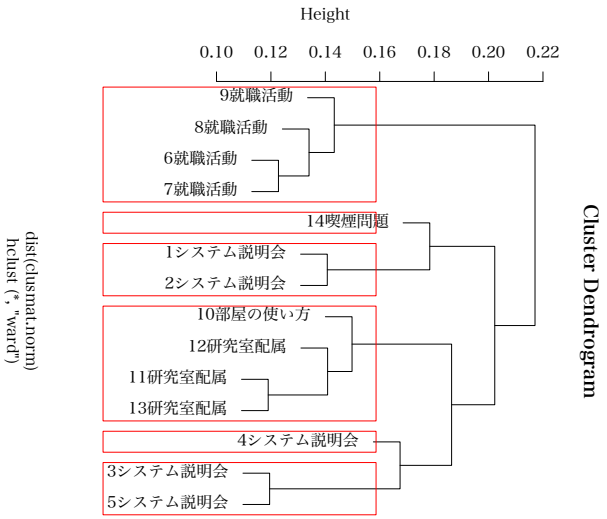

図 2: 話題クラスタリングの結果

話題「就職活動」に関するクラスタは 1 つにまとめることができている。話題「研究室配属」と「部屋の使い方」クラスタが 1 まとめになっているが、この 2 つの話題は生起時間が近く、また “部屋” や “室” といった共通する特徴次元が多い事が原因と見られる。話題「システム説明会」は 3 つのクラスタに分かれてしまっている。この話題のクラスタには専門用語が多く、“サーバ” と “サーバー” などの表記ゆれや、“ブレードサーバ” と “シンクライアント” などの表層形のみではわからない関連語が多く見られた。

5 まとめ

本研究では、時系列特徴を考慮したクラスタリングと、そこで得たクラスタから抽出した特徴ベクトルを使った再クラスタリングによって、ツイートの話題をクラスタリングする手法を提案した。時系列特徴を考慮したクラスタリングによって、79%の精度でツイートをクラスタリングすることができた。次の段階である各クラスタから特徴ベクトルを抽出した上での再クラスタリングにおいて、いくつかのクラスタを正しい話題にまとめる事ができたが、その精度は不十分であった。クラスタからの特徴ベクトルの取得方法を工夫することで、よりよい話題クラスタを得ることが今後の課題である。

参考文献

[1] 菊池匡晃, 岡本昌之, 山崎智弘. “階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出” 日本データベース学会論文誌. 第 7 巻

[2] 戸田浩之, 北川博之, 藤村考, 片岡良治. “時間的近さを考慮した話題構造マイニング” 電子情報通信学会論文誌. D, 情報・システム J90-D(2), 292-310, 2007-02-01

[3] 松村真宏. “影響伝播モデル IDM の線形代数表現と Twitter 分析への応用” 電子情報通信学会第二種研究会資料（Web インテリジェンスとインタラクション）, WI2-2010-1 31, pp.73-78.

[4] 高村大也, 横野光, 奥村学. “Summarizing microblog stream” 人工知能学会第 22 回 SWO 研究会 SIG-SWO-A1001-03, 2010.

[5] 第二回情報工学科総会 http://ie.u-ryukyu.ac.jp/blog/2011/04/08/第 2 回情報工学科総会/

[6] Twitter ヘルプセンター Twitter 用語集 http://support.twitter.com/groups/31-twitter-basics/topics/104-welcome-to-twitter-support/articles/243951-twitter

[7] Frantsi, K. and Ananiadou, S. “Extracting Nested Collocations” COLING 96, pp.41-46, 1996.

[8] Calinski, R.B., Harabasz, J. “A dendrite method for cluster analysis” Communications in Statistics, vol. 3, 1-27.