4I1-R-9-2

マイクロブログの特徴を考慮した文書クラスタリング手法の提案と実装

Proposal and Implementation of Document Clustering Method based on feature of MicroBlog

辻井　由佳[*1]　　西山　裕之[*1]
Yuka Tsujii　　Hiroyuki Nishiyama

[*1]東京理科大学理工学研究科経営工学専攻
Graduate School of Science and Technology, Tokyo University of Science

In recent years, MicroBlog has been used to find out real-time information. However, search by keywords of MicroBlog have some problems which it is difficult to recognize whether posting is useful or not and understand outline by limited number of characters. Document Clustering Method has a beneficial effect on this matter. On the other hand, Document Clustering Method for MicroBlog can't show high accuracy because it is hard to express features due to limited number of characters and word which were made recently. So, The purpose of this study is that we propose and implement Document Clustering Method based on feature of MicroBlog.

1. はじめに

近年，ネットワークの発展により，情報を得るためにインターネットを利用するのが一般的になった [総務省 11]．中でもリアルタイムの情報を得る手段として，自身の状況などを短文で書き込むためのツールであるマイクロブログが挙げられる．その大きな特徴としては字数制限があるほか，発信者の大多数が個人である，ユーザ間でやりとりができることなどが挙げられ，中でもよく知られているのが Twitter[*1]である．Twitter では投稿できる文字数が 140 文字に制限されており，投稿された記事をツイートと呼ぶ．本研究ではマイクロブログとして Twitter を取り扱う．

一般的に，マイクロブログにおいて情報を得るにはキーワード検索を行う．しかし，この結果にはプライベートな投稿と有益な情報を含む投稿が入り混じっているため，有益な情報のみを取得することが困難であり，字数制限によって一件の情報量が少ないため，複数の投稿を読まなければ事態の全容を把握するのが難しいといった問題がある．こういった結果の表示における問題に対し，文書集合にクラスタリングを行うことでわかりやすく結果を表示できることがわかっている [岸田 03]．

けれども，字数制限により特徴を掴みづらい，流行語や造語に対応できないといった理由から，マイクロブログに一般的に使用される文書クラスタリングの手法を適用しても十分に高い精度を得ることは難しいのが実状である．このような文書クラスタリング手法をマイクロブログに適用する試みには，青島らの研究 [青島ら 10] などが挙げられるが，ユーザが制約条件を付加することを目的としている．

そこで本研究では，マイクロブログの特徴を考慮して，同じ話題をクラスタにまとめる文書クラスタリング手法の提案と手法の実装を目的とした．

2. 文書クラスタリング

文書クラスタリングには複数の手法が存在する．ここでは本研究で使用する一般的なクラスタリング手法とそれに関連する技術について確認する．

連絡先: 辻井由佳，東京理科大学理工学部，千葉県野田市山崎 2641，j7412616@ed.tus.ac.jp

∗1　Twitter : https://twitter.com/

2.1 形態素解析

形態素解析とは自然言語処理の基本技術の一つであり，文を単語単位（形態素）に切り分け，品詞の情報，活用形などを付加する作業を示す．

2.2 特徴語抽出

形態素解析で単語を抽出した後に，ツイートの特徴を示す単語である特徴語を選択する．一般的には名詞や形容動詞の語幹，サ変動詞の語幹が特徴語の候補とされる．

2.3 重要度計算（TF-IDF 法）

TF-IDF 法とはある文書内における単語出現頻度 (tf) とその単語が出現する文書の割合 (idf) の積をその単語の重要度とする方法であり，他の文書での出現頻度が低く，ある文書内では出現頻度の高い単語が重要度が高く，特徴的な単語とされる．

2.4 類似度

類似度の測定にはいくつかの方法があるが，文書クラスタリングの類似度計算にはコサイン類似度がよく使用される．

2.5 クラスタリング

クラスタリング手法の主なものとして k-means 法があるが，この方法では前もってクラスタの個数と種子点（seeds）と呼ばれるクラスタの核となる文書を決定する必要があり，この初期設定によって結果が大きく変化することが分かっている．しかし，本研究では話題ごとにクラスタを作るためにクラスタの個数を前もって決めることはできず，最適な初期設定を決定することが難しい．そこで，クラスタリング手法として，与えられた閾値をもとにクラスタリングを行う leader-follower 法を使用した．

3. 提案手法

前述した一般的なクラスタリング手法にマイクロブログの特徴を考慮したいくつかの対策を加えてクラスタリング精度の向上を考える．以下にマイクロブログの考慮すべき特徴と，それに対する対策手法を示した．

3.1 字数制限

マイクロブログのもっとも大きな特徴は字数制限である．字数制限によって取得できる単語数が少なくなるため，ツイートの特徴が掴みづらく，重要度の計算においても問題が生じる．

字数制限による問題として，まず，一文が短いツイートでは話題の中心となる語でも繰り返されることが少なく，単語の出現頻度で重要度を求める TF-IDF 法があまり有効でないことが挙げられる．この問題に対し，TF の値をそのまま重要度とした．IDF を使用しなくても特徴語として品詞を限定している時点である程度一般的に広く使われる語は減らせると推測されることに加え，検索ワードを排除することでその役割を果たせるものと考える．これにより重要度の計算において特徴語が少ないという影響が減らせ，出現頻度をより重視できる．

もともとの単語数が少ないという問題に対して，特徴語候補（名詞や形容動詞の語幹，サ変動詞の語幹）の中から特徴語を選択するのではなく，特徴語候補をそのまま特徴語とした．また，本研究では上記品詞以外に，形態素解析器の辞書に登録されていない語を意味する未定義語も特徴語として使用することを考える．未定義語に分類されるのは意味を成さない言葉や顔文字，または近年作られた造語などの最近になって使われるようになった言葉や人物名である．そこで，未定義語の中から意味を成す言葉を見つけるためにキーフレーズ抽出[*2]を利用した．これにより，最近作られた単語など辞書に登録されていない用語についても特徴として利用することができる．

3.2 ユーザ間のやり取り

Twitter にはユーザ間のやり取りを行うことができる機能がある．”@ユーザ名 ” がついたツイートはそのユーザに向けたコメントを示し，同じワードを含み同一ユーザに対するツイートは関連する話題やそれに対するやりとりだと考えられる．また，ツイートには URL を乗せることも可能である．このとき，同じホームページに対するツイートも関連する話題についてだと考えられる．以上を踏まえ，同一のユーザに向けられたツイート，同一の URL が含まれるツイートをクラスタとする．

leader-follower 法では一番初めの文書を初期クラスタとするが，本研究では上記手法で作られたクラスタ間の類似度を測定し，閾値を超えたクラスタを統合して，初期クラスタとして設定した．もし，同一ユーザ宛てのツイート，同一 URL の記述されたツイートが存在せず，クラスタができなかった場合は従来通り，一番初めのツイートを初期クラスタとする．

4. 評価

従来手法との比較を通じて本研究の提案手法における精度について考察を行う．

精度の評価方法としてどれだけ取りこぼしなく判別できているかという度合いを示す再現率とどれだけ無駄なものが入っていないかを示す指標である適合率を使用し，このふたつをまとめて評価できる F 値を尺度として使い，評価を行った．本研究では手動でクラスタリングを行ったものを正解クラスタリングとしている．

検索エンジン Google が提供する今日の急上昇ワード[*3]からランダムに評価の対象とする検索ワードを 500 件選択した．以上の結果に対し，検索ワード 1 つにつき 100 件のツイートを取得し，各ワードごとに提案手法と従来手法を使用したクラスタリングを行う．各手法における leader-follower 法の閾値を 0.1，0.2，0.3，0.4，0.5，0.6，0.7，0.8，0.9 に設定し，一般的な手法と提案手法における各閾値ごとの F 値の平均を測定する．

結果をまとめたものを以下の表 1 に示す．

表 1: 各手法における閾値ごとの F 値の平均

閾値	従来手法	提案手法
0.1	0.42	0.38
0.2	0.55	0.59
0.3	0.61	0.74
0.4	0.64	0.79
0.5	0.65	0.82
0.6	0.63	0.82
0.7	0.61	0.82
0.8	0.59	0.81
0.9	0.55	0.79

表 1 に示された通り，閾値 0.1 を除くすべての閾値で提案手法による精度の向上が見られ，提案手法では閾値 0.5 から 0.8 の間で 0.8 を超える精度を得ることができた．

また，本実験では検索ワードごとに最も良い精度を示す閾値にばらつきが見られ，閾値を一つに固定するのではなく，最適な閾値を検索ワードごとに選択することでさらなる精度の向上が見込めると推測された．そこで閾値ごとの平均精度ではなく，検索ワードごとに一番高い精度を示す閾値を選択し，その値から F 値の平均を求めた．その結果，従来手法では 0.69，提案手法では 0.88 と，表 1 で示したどの閾値における結果よりも良い精度を得ることができた．ここから，検索ワードが持つ特徴ごとに適した閾値が異なると推測され，より高い精度を出すためには細かく分類したいときには高い閾値を設け，大きく分類する場合は低い閾値を設定する必要があると考えられる．

また，本提案手法では同一ユーザあての発言をまとめる際に，ニュースを発信するようなアカウントでは違う話題であっても同一の話題としてクラスタリングされてしまうといった問題もみられた．

5. おわりに

本研究ではマイクロブログにおける情報取得支援を目的とし，わかりやすい検索結果の提示のためにマイクロブログの特徴を踏まえて新たなクラスタリング手法を提案した．一般的な手法のみのクラスタリングと本研究で提案したクラスタリング手法の精度を比較し，その結果に対して考察を行い，精度の向上を確認することができた．

今後の展望としては，まず，検索ワードごとに最適な閾値の選択を行うことが挙げられる．また，作成されるクラスタにより関心がある話題に関する情報量が増えることを利用して，追加情報の自動取得や，その内容に対する信頼性についても言及したい．

参考文献

[総務省 11] 総務省:情報通信白書 2011 年度版，2011

[岸田 03] 岸田　和明:文書クラスタリングの技法:文献レビュー，Library and Information Science, No. 49, p.33-75，2003

[青島ら 10] 青島傳隼，福田直樹，横山昌平，石川博:マイクロブログを対象とした制約付きクラスタリングの実現，DEIM Forum，B1-3，2010

∗2 Yahoo!キーフレーズ抽出：Yahoo!デベロッパーネットワークが提供する API（http://developer.yahoo.co.jp/）

∗3 Google 今日の急上昇ワード：http://www.google.co.jp/trends