テキストクラスタリングにユーザの意思を盛り込む手法の考察

A Method for Incorporating User's Idea into Text Clustering

笹崎 格　　元木 達也
Itaru Sasazaki　　Tatsuya Motoki

新潟大学大学院自然科学研究科　　新潟大学工学部
Grad. School of Sci. & Tech., Niigata University　　Faculty of Eng., Niigata University

1 はじめに

テキストクラスタリングは、大量にあるテキスト集合を自動的に分類を行う手法であり、主にテキスト検索や大量のテキストの自動要約に応用される。単純なテキストクラスタリングでは単語の度数、共起などの統計量に基づいてクラスタリングが行われ、ユーザの意思が十分に反映されるとは限らない。

そこで、本研究では、ユーザの意思をテキストクラスタリング処理に盛り込むために、特徴語の最終的な選択をユーザに行わせたり、ユーザにヒントを与えさせ複数回クラスタリングを行わせたり、といった処理を行う。

2 テキストクラスタリングでユーザの意思を盛り込む方策

従来、テキストクラスタリングは、(1) 特徴語抽出、(2) 抽出した特徴語に基づくクラスタリングという手順で行われ、ユーザの意思はステップ (2) の直前にヒントを与えたり [2]、ステップ (2) の直後に得られたクラスタの取捨を行ったり [1] することによってわずかに反映できるのみであった。そこで、この研究ではユーザの意思を更に取り込む為に、次のようなクラスタリング手順を考える。

(1) 特徴語抽出 (ユーザの意思も反映)
(2) 抽出した特徴語に基づくクラスタリング
(3) クラスタリング結果に満足できない場合は、ユーザがヒントを与えてステップ (2) へ戻る。

ここで、最初の特徴語抽出のステップにおいては、基本的には tf-idf 値 [5] の高い単語を特徴語として選ぶが、tf-idf 値の高い単語をアルファベット順に並べた 特徴語候補列を参考にユーザは自分の望む分類に利用できそうな特徴語を選択できるとする。分類上での単語の意味を考慮し複数の単語を 1 つの同義特徴語グループとして指名することも可とする。

次のクラスタリングのステップでは、抽出された特徴語を基にベクトル空間を張り、この空間内に配置されたテキストをクラスタリングする。その方法としてここでは、クラスタ数の調整も行う VGA(Variable string-length Genetic Algoritm)[3] を考えクラスタリング候補の良さを測る指標 (適合度関数) として DB-index(Davies-Bouldin index)[3] を用いる。

最後のステップ (3) においては、ユーザはテキストクラスタリングの分類結果を確認する。その結果、本来 1 つにまとめるべきクラスタとテキストの組が見出されたなら、強制的にそのテキストをそのクラスタの重心に配置し直した上で、再度テキストクラスタリングを行う。

3 実験

実験データとしては Reuter-21578[4] という実験用に公開されているデータを用いる。テキスト群はそれぞれ tea,gas,iron-steel,cpu,jobs という 5 つの分類が既にされており、tea の分類のテキストが 15 個、gas が 15 個、iron-steel が 66 個、cpu が 127 個 jobs が 75 個という分布となっている。これらのテキスト群に対して、表 1〜3 のパラメータの下で第 2 節の手順を試した結果を図 1、2 に示す。実験ではユーザが正解のような分類を行いたいものとして情報を与えていく。図 1、2 は VGA の各世代における最良個体 (i.e. 適合度最高の分類結果) がどの程度正しい分類と一致しているかを調べたグラフである。

表 1 特徴語抽出のパラメータ

特徴語候補選定法	tf-idf
特徴語候補の表示数	300
ユーザが選択した特徴語数	6
自動的に決定させた特徴語数	9

表 2 VGA 処理のパラメータ

集団の大きさ	300
終了世代	30
適合度	DB index
選択方法	サイズ 2 のトーナメント選択
交叉確率	0.7
突然変異確率	0.2
再生確率	0.3

表 3 結果修正のパラメータ

修正回数	2
修正したテキスト数	30

図 1 の類似性検出率とは正解のクラスタリングで同一クラスタに分類されているテキストのペアの内、探索結果で正しく同一クラスタに分類された割合、すなわちテキスト間の類似性を正しく検出できた割合であり、図 2 の類似性誤検出率とは正解のクラスタリングで同一クラスタに分類されないテキストのペアの内、探索結果で誤って同一クラスタに分類された割合、すなわちテキスト間の誤った類似性をもたらした割合である。クラスタリングの回数を重ねる毎に類似性検出率、類似性誤検出率のどちらも上がるのは、現在のクラスタリング結果の修正方法がテキストを一つのクラスタに纏める働きしか行っていない為である。

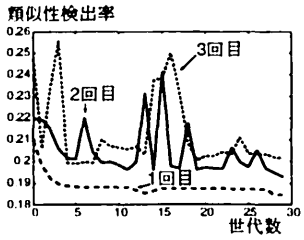

図 1 類似性検出率

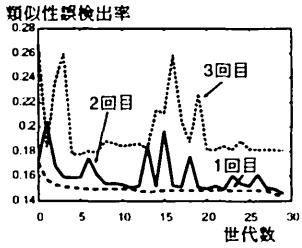

図 2 類似性誤検出率

4 これからの課題

正しい分類に導くために 2 回目以降の VGA にユーザが色々な情報を指示することも可能である。しかし、それだけユーザの負担が大きくなるので、どの程度の処理までユーザに任せるべきなのか検討中である。

参考文献

[1] Manu Konchady, *Text Mining Application Programming*, Charles River Media, 2006.
[2] Kiri Wagstaff, Claire Cardie, Seth Rogers and Stefan Schroedl, Constrained K-means Clustering with Background Knowledge, *Proc. of the Eighteenth International Conference on Machine Learning*, pp.577-584, 2001.
[3] Sanghamitra Bandyopadhyay and UjjwalMaulik. Performance Evaluation of Some Clustering Algorithms and Validity Indices, *IEEE Trans. Pattern Analysis and Machine Intelligence* , Vol.24, No.12, pp.1650-1654, 2002.
[4] Reuter-21578 [http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html]
[5] Savio L.Y. Lam and Dik Lun Lee. Feature Reduction for Neural Network Based Text Categorization,*6th International Conference on Database Systems for Advanced Applications*, pp.195-202, 1999.