アンケートを対象としたテキスト自動分類システムの検討

4N－3

杉崎 正之　大久保 雅且　田中 一男

NTTサイバーソリューション研究所

1 はじめに

アンケートは市場の意識調査や意見収集などの一般的な手段である。旧来は紙や電話等を使って行なわれていたが、時間を選ばず、地理的条件からも解放されるという利点から、インターネット上のWebや電子メールを用いて収集する方法が多く取られるようになった。

年齢や住所などの定型的な情報は、コンピュータを用いることで集計しグラフ化することが容易であるが、「意見」などの自由文回答が可能な非定型的な情報は、一つ一つを人手で確認するのが現状であり、処理できる数と時間が問題となっている。時間の問題はまた、情報の鮮度の問題とも関係がある。せっかくアンケートを大量に収集しても、その収集分析をしているうちに市場の意見が変化しないとも限らない。また、例えばテレビやラジオの放送中にアンケートを受けて、その内容を番組内でリアルタイムに反映したい場合なども処理時間は非常に重要になってくる。時間とデータ数を考えて負荷分散のために人員を多く用意するという方法をとっても、人員間の意識合わせと結果の統合を行なう必要があり、有効な解決策とはなりにくい。

そこで今回、小人数で自由文回答を分析するためのシステムを目指し、処理時間の短縮と分類精度の向上の2点について検討した。

2 処理の高速化

2.1 クラスター分析

自由文の自動分類手法にはクラスター分析を用いている。文書間の類似度を定義し、類似した文書を一つのまとまり(クラスター)に割り当てていくことで文書すべてを葉とする2分木が構成できる。それに閾値を導入することで、複数のクラスターを抽出する。文書の特徴ベクトルを $tf \cdot idf$ の手法 (Salton[1]) を用いて生成した。文書間の類似度は特徴ベクトル間の成す角 ($cos\theta$) とした。なお、単語の切り出しにはInfoBeeの形態素解析エンジン ([2]) を使用した。

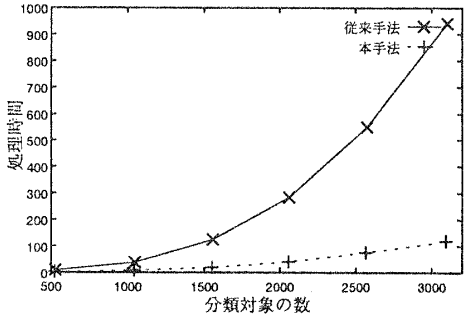

図 1: 処理時間のグラフ

最も類似したクラスター p,q から新たなクラスター r を生成した場合、r とそれ以外のクラスターとの類似度計算を行う。このとき r の特徴ベクトルを p と q に割り当てられた文書すべての特徴ベクトルの平均とし、他の特徴ベクトルとの類似度を計算することにした。

2.2 次候補記録方式

クラスター分析処理において時間がかかるのは、文書間の類似度計算とそれらからの最大値探索であり、今回は後者に注目した。処理の高速化を行なうために、類似度が最大となるクラスター p,q の組の探索途中の結果を記憶して再利用する方法を取る。

クラスターの数を n とすると、それらの類似度は $n \times n$ の行列 (これを A とする) で表現できる。行列 A の各行毎にその行内で類似度が最大となる列とその値を記録しておく (このテーブルを T とする)。値が最大となる行列 A の要素の組はテーブル T を探索することで見つけ出すことができる。

新しいクラスターを生成した場合、行列 A の要素の値の変更が行なわれる。しかし、行列 A で実際に変更されるのは、主にクラスター p,q に相当する行と列、および、新規に生成したクラスターとそれ以外のクラスターとの類似度である。テーブル T の要素の変更点は、(1) クラスター p,q と新規クラスターの要素、(2) 行列 A の各行でラスター p,q が最大値を取る要素、の2点であり、それ以外の要素はそのまま次の最大値探索に利用できる。

A study on text clustering system
for analyzing questionnaire results
Masayuki SUGIZAKI, Masaaki OHKUBO,
and Kazuo TANAKA
{sugizaki, ohkubo, tanaka}@aether.hil.ntt.co.jp
NTT Cyber Solutions Laboratories

2.3 実験および結果

上記のアルゴリズムと従来手法との計算速度を、実際のデータを用いて実験した。その結果を図1に示す。横軸がアンケートの数、縦軸が処理時間である。処理対象が少ないときは差がないが、増加するに従って次候補記録テーブルが有効に働いているのが分かる。

3 類義語の抽出

3.1 分類精度の問題

自然言語で書かれた文書の分類において分類精度の点で重要となるのは、(a) 重要単語の抽出と利用と (b) 類義した単語 (あるいは語句) の利用である。(a) は、分類対象の文書内で分類に有効に働く単語、あるいは、分類基準として意味を成さない単語を見つけることができれば分類時の指標として有効に利用することができる。しかし、ある単語が重要かどうかの判断は、個々の問題に依存する。(b) は、いわゆるシソーラスと呼ばれる辞書などの利用が考えられる。しかし、アンケートの分類に関しては、類似した単語の判断基準はアンケート分析者毎や分析対象の設問毎に異なり、一般的なシソーラスの利用は難しい。

単語の出現頻度等を用いたクラスター分析による分類結果は、人間が意図した分類結果と異なり、なんらかの修正を行いたいという要求が生じる。そこで、分析者が分類結果を評価や修正といった操作をした場合、その情報を再分類時に利用する方法を検討した。

3.2 類似候補の抽出

質問が明確な場合、回答文の特徴として端的なものが多く、結果的に文書の長さは短く、存在する単語の量も少ない。単語の種類も、同じ文書量の新聞記事と比較して非常に限定されている。図2がその確認用の前実験の結果で、アンケートは出現回数の多い単語が多く、逆に新聞記事は出現回数の少ない単語が多かった。すなわち、アンケートの自由文を分類する場合「限られた単語」から分類を行なう必要がある。その条件のなかで分類結果のクラスター数をより少なくするには、前述 (a) の単語の重みを変化させるより (b) の類似した単語を見つけることが有効であると判断した。

類似した単語として、(i) 単語の表記のゆらぎ ("無し" と "なし") や (ii) 同意語 ("風景" と "景色")(iii) 複数の語による同意語 ("海で泳ぐ" と "海水浴") がある。そのうち、(i) は単語の読み、(ii) は同じ意味の単語は字面でも似ているという仮説から、この2点について抽出を検討した。抽出方法として、まず、分析者は抽出されたいくつかのクラスターで類似しているものを選出し、分類システムに対し類似している旨を伝える。システムは、与えられた各クラスター内に存在する単語間で類似している単語を抽出する。類似の判断は (i) 単語の読みと (ii) 字面の2つの評価基準を用いて行なう。(i) は単語の読み方が同じ場合に類似の候補とし、(ii) は DP マッチングの手法を用いて単語間の距離を定義し、距離が近いものを類似候補とする関数を用意した。

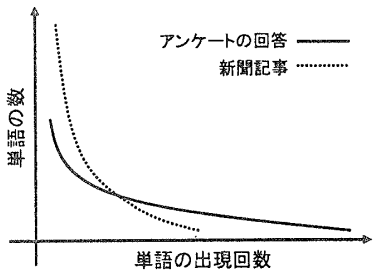

図 2: 単語の分布

3.3 考察

上記のアルゴリズムで、実際のデータから類似した単語の抽出を行なわせた。アンケートの質問は「デジタルカメラでよく撮るもの」である。実際に書かれた回答として「子供の写真」「風景」「ペット」などが多かった。学習前で例えば「子供」と「子ども」、「風景」と「景色」、「愛犬」と「犬」などが別の単語として認識されていたが、学習を行なうとそれぞれ類似単語候補として抽出することができた。

4 今後の課題

今回、計算の高速化と類義語の抽出に焦点を当てた。類義語の抽出方法は、各個人毎の類似語の抽出方法であったが、複数の分析者からの指示を利用して、小人数間での類似語の抽出および重要単語の重みの変更方法を検討したい。

参考文献

[1] G. Salton: Automatic Text Processing, Addison Wesley, 1989

[2] 井上, 大久保, 杉崎: InfoBee テキスト情報検索技術, NTT R&D 10月号, pp.1103-1108, 1997