疑似カテゴリ生成によるテキスト自動分類の高速化について

2V－3

杉崎 正之　森 大二郎　大久保 雅且　田中 一男

NTTヒューマンインタフェース研究所

1 はじめに

近年、インターネットに代表されるコンピュータネットワークの普及により、数多くの電子的なテキストが発信されている。大量のテキストから必要な情報を引き出すためにフルテキスト検索技術が利用され、インターネットではgoo(http://www.goo.ne.jp)やAltavista(http://www.altavista.digital.com)などの不特定多数の人々を対象としたWeb検索サービスが行われている。しかし、得ることの出来るテキスト情報が大量であるために、必要な情報を収集することのみならず、集めた情報の整理や分類に非常に手間がかかる。

テキスト自動分類技術の研究は従来から行われており、分類するための箱(カテゴリ)をあらかじめ用意し、それらのカテゴリに分類する手法(classification)と、テキスト集合から自動的に分類するためのカテゴリを抽出し分類する手法(clustering)の2つに大別できる。

前者の分類手法に対し、数億を超えるとも言われているWWW上のホームページなどの大量のテキストを、多くのカテゴリに分類する場合の処理時間に注目し、分類処理を高速に処理するための手法を検討した。

2 従来の分類手法と問題点

最初に、テキストの自動分類技術について説明する。本報告では、あらかじめ用意したカテゴリにテキストを分類する技術として最近傍決定則(NN法)の手法を用いる。まず、分類したいカテゴリとそこに割り当てられるべきサンプルのテキストから、各カテゴリ毎にテキストを形態素解析処理し単語の抽出を行う。抽出した単語とその出現頻度から、カテゴリ毎に単語の特徴ベクトルを生成する。このとき、各要素の値はSalton[1]の$tf*idf$を用いた。

分類手順は、分類対象となるテキストから単語の特徴ベクトルを生成し、カテゴリの持つ特徴ベクトルとの類似度を計算することで、類似していると判断したカテゴリに分類する。類似度関数として三角関数のcosを用い、分類の基準は、ある閾値を用いて「類似度が閾値より大きなカテゴリすべてに割り当てる」とする。分類処理のイメージを図1に示す。

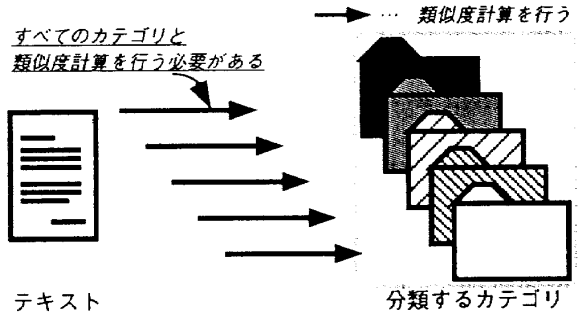

図1: NN法を用いた従来手法

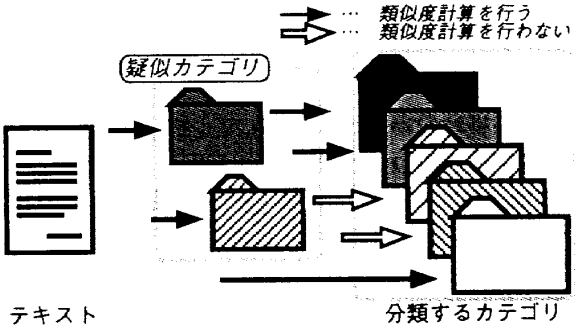

図2: 本手法

今回、対象としている分類手法では、テキスト数がn、カテゴリ数がmのとき、類似度を求めるための計算回数は$m \times n$であり、テキスト数やカテゴリ数に比例して計算回数は増大する。また、分類精度への影響を考慮すると、単語の取捨選択による処理の高速化は困難である。特徴ベクトルの要素数の変更とは別の分類処理の高速化手法について検討した。

3 疑似カテゴリの生成

類似度計算を実際に行う前に、計算すべきかどうかの判断を行うことで計算コストを削減する方法を考える。そのために、あらかじめカテゴリ同士で類似したものをひとまとめにして擬似的なカテゴリを作成し、疑

An efficient text classification method
by a pseudo-category generation
Masayuki SUGIZAKI, Daijiro MORI,
Masaaki OHKUBO, and Kazuo TANAKA
NTT Human Interface Laboratories

似カテゴリと分類対象のテキストとの類似度判断により、実際に類似度の計算を行うカテゴリを選択する手法を取る。図2に、本手法の分類時のイメージを示す。その方法として次のような手法を考えた。

類似しているカテゴリの組を作り出すための基準としてカテゴリ間で共通に出現した単語の割合を用い、クラスター分析 [2] を行って擬似的なカテゴリを生成する。具体的には、カテゴリ C_i、C_j の類似度 Rel_{ij} を、

$$Rel_{ij} = \frac{(\text{共通に存在した単語数})^2}{(C_i\text{の単語数}) \times (C_j\text{の単語数})} \quad (1)$$

とし、類似度の値が大きいカテゴリ同士を一つの疑似カテゴリに割り当てた。疑似カテゴリの特徴ベクトルは割り当てられたすべてのカテゴリの持つ特徴ベクトルの平均とし、新たな疑似カテゴリを生成するたびにカテゴリ間の類似度を再計算するようにした。

この処理では、最終的にすべてのカテゴリが一つの疑似カテゴリに割り当てられてしまうという問題があるため、それを避けるために以下の条件を加える。疑似カテゴリ M_k と疑似カテゴリ M_l において、

$$\forall C_i \in M_k, \forall C_j \in M_l \text{に対し} \quad Rel_{ij} > \alpha \quad (2)$$

を満たす場合のみ、M_k と M_l から新たな疑似カテゴリ M_m の生成を行う。α は0から1の間の実数値とする。

式 (2) では、閾値 α の値を調整することにより、疑似カテゴリ内に割り当てられたカテゴリ相互の関係を決定することができる。すなわち、α の値を大きくすれば、疑似カテゴリ内に割り当てられたカテゴリ相互の関係は密になり、疑似カテゴリでの判断による不要な類似度計算の回数が多くなることを避けることが出来る。値をうまく設定すれば、疑似カテゴリ内の不要なカテゴリを減らすことができ、その結果、類似度計算の回数を減らすことができると考えた。

分類時には、分類対象のテキスト内に存在する単語が疑似的なカテゴリ内に存在するか判断し、一つでも存在する場合は疑似カテゴリに分けられた実際のカテゴリとの類似度計算を行う。

4 計算コストの評価

実験システムを構築し、計算コストの評価を行う。コンピュータや趣味のカテゴリなど約800のカテゴリを用意し、分類対象としてインターネット上の約2000のWebページを利用した。その結果を図3に示す。

類似度計算を、すべてのテキストとカテゴリ間で行った場合は160万回 (図3の (1) + (2)) であり、このうち、実際に類似度計算を行う必要があった回数、いいかえると、実際に類似度が0より大きくなった計算回数

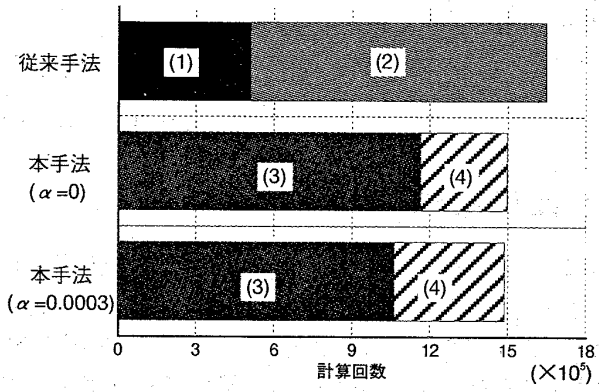

図 3: 計算コストのグラフ

は、約51万回 (図3の (1)) であった。すなわち、図3の (2) の部分をいかに減らせるかが本研究の課題である。

本手法の類似度計算の回数は図3の (3) で、疑似カテゴリとの判断回数は図3の (4) となった。類似度計算の回数は従来手法の約2/3に減少しており、疑似カテゴリでの判断回数を加えると、全体で約1割の計算コストが削減できていることが実証できた。

また、α が0.0003の場合を0の場合とで比較すると、カテゴリとの類似度計算の回数を減らすことが出来ていた。しかし、疑似カテゴリ数が増加し、全体として値が0の場合とほぼ変わらない結果になった。

5 おわりに

あらかじめ与えられたカテゴリにテキストを分類する分類手法において、カテゴリとテキスト間の類似度計算の回数を減少させるために擬似的なカテゴリを用いる手法を検討し、実験を行った。その結果、実際に類似度の計算回数を減らすことができた。

式 (2) において、α の値の調整によって実際に類似度計算の回数を減らせることが出来たが、疑似カテゴリとの計算回数が増えてしまい、閾値 α の効果は明確にはできなかった。α の最適値の抽出は今後の課題である。

また、主成分解析などを用い、カテゴリの分布状況に従った疑似カテゴリ生成による計算コストの削減手法を検討し、本手法と比較評価する予定である。

参考文献

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

[2] 田中, 脇本: 多変量統計解析法, 現代数学社, 1983