D-4-1 テキスト自動分類におけるサブカテゴリの生成による分類精度の改善

An Improvement of Accuracy for Automatic Text Classification by Generating Sub-Category

加藤 大樹 Hiroki Kato　熊谷 貴禎 Kiyoshi Kumagai　小林 学 Manabu Kobayashi　平澤 茂一 Shigeichi Hirasawa

早稲田大学 理工学部 経営システム工学科

Dept. of Industrial & Management Systems Engineering , School of Science and Engineering, Waseda University

1 はじめに

近年日本語テキストの電子化が進み，文献検索等の用途のために少数の専門家によって適切に分類が行われ，また整理されてきた．しかし，非常に大量のテキストが利用可能となった現在，テキストを自動的に分類することの必要性が高まっている．従来のベクトル空間モデルに基づく自動分類手法では，カテゴリごとに一意の重心ベクトルを決定するために，カテゴリ内の文書の分布に多数の偏りがある場合は適切な分類をすることができないという問題点がある．

そこで本研究では，重み付き類似度という新しい指標を導入し，各カテゴリ毎に文書ベクトルのクラスタリングを行い，複数のサブカテゴリ及びガテゴリベクトルを生成する手法を提案する．さらに，この手法をベンチマークデータに適用しシミュレーションによりその有効性を示す．

2 問題設定と従来手法

ベクトル空間モデル[1]では，文書及びカテゴリをキーワード[2]の出現頻度を重み付け[3]したものを要素とするベクトルで表現し，文書とカテゴリの距離をベクトルの類似度[4]で表す．

従来，カテゴリ内の文書ベクトル全ての重心であるカテゴリベクトルは，必然的に一つのカテゴリに一つしか存在し得ない．しかし，専門家によってカテゴリが作られ手作業で分類した結果，多くの場合にカテゴリの分布に多数の偏り（多峰性等）が生じる．例えば図1のように「国際」カテゴリの分布に2つの偏りがある場合，本来ならば距離の近い「国際」カテゴリに分類されるべき新規文書が「政治」カテゴリ分類されてしまうという問題がある．

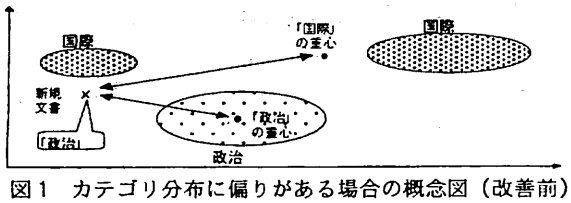

図1 カテゴリ分布に偏りがある場合の概念図（改善前）

3 提案手法

本研究では，カテゴリ毎の分類済み文書データに対しクラスタリングを行うことによりサブカテゴリを生成し，サブカテゴリ毎にカテゴリベクトルを計算するという手法を提案する．それによって，文書分布に多数の偏りがある場合にも，分布に適応したカテゴリベクトルを生成することができるようにする．例えば，図2のように，従来では正しく分類されなかった新規文書を正しく分類できるようにする．

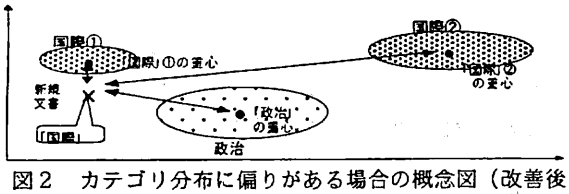

図2 カテゴリ分布に偏りがある場合の概念図（改善後）

3.1 カテゴリ内文書に対するクラスタリング

与えられたカテゴリ内文書に対し凝集法クラスタリング[5]を行い，クラスタリング終了時点でのクラスタをサブカテゴリとし，クラスタの重心ベクトルをサブカテゴリのカテゴリベクトルとする．

3.2 サブカテゴリに応じた文書の自動分類

新規文書の文書ベクトル$\mathbf{d}'$と各サブカテゴリのカテゴリベクトル$\mathbf{c}'$との類似度を求める．ここで，クラスタリング結果であるサブカテゴリは所属する文書の数に大きな格差が生じる可能性が高い．そこで，サブカテゴリに所属する文書数D'の平方根（平方根である根拠は予備実験による）を新規文書とサブカテゴリとの内積に乗じた重みつき類似度$\mathrm{wsim}(\mathbf{d}',\mathbf{c}')$を定義する．

$$\mathrm{wsim}(\mathbf{d}',\mathbf{c}') = \frac{\mathbf{d}'\cdot\mathbf{c}'}{|\mathbf{d}'||\mathbf{c}'|}\times\sqrt{D'} \qquad -(1)$$

最終的に$\mathbf{d}'$に対し，この重み付き類似度の最も高いサブカテゴリの属するカテゴリに分類する．

4 提案手法の評価

シミュレーションには毎日新聞の1年分の記事データを収録したCD-毎日新聞94' データ集[6]を利用した．

従来手法，提案手法で分割クラスタ数を変化させた場合及びクラスタ内分散Vに閾値を設けてクラスタ数の決定を行った場合の，文書のカテゴリへの分類精度のグラフをそれぞれ以下の図3，図4に示す．ここでVは以下のように定義する．$\mathbf{c}$はクラスタリング途中のクラスタ，$\mathbf{d}$はクラスタ所属文書，Dは所属文書数を表す．

$$V = \sum\left(1-\frac{\mathbf{d}\cdot\mathbf{c}}{|\mathbf{d}||\mathbf{c}|}\right)^2 \Big/ D \qquad -(2)$$

今回の評価実験では，トレーニングデータを1万件，テストデータを5千件，抽出するキーワードの種類を1千単語とした．実験に使用するカテゴリ数は9個とした．

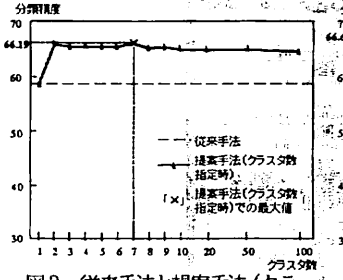

図3 従来手法と提案手法（クラスタ数指定）の分類精度比較

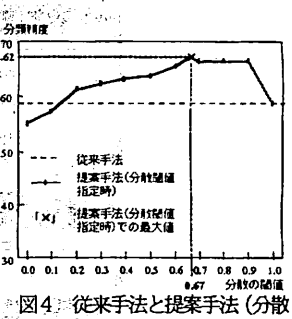

図4 従来手法と提案手法（分散の閾値指定）の分類精度比較

図3はサブカテゴリ数が2以上の時に分類精度が向上することを示している．これより人手による分類をベクトル空間モデル上に再現するには，各カテゴリに複数のサブカテゴリが必要であることがわかる．

図4はクラスタ内分散の閾値を適正に設定することにより，分類精度が向上することを示している．閾値が小さすぎるとサブカテゴリ数が増大し，大きすぎるとサブカテゴリ数が減少するため，いずれも分類精度が落ちる．

また，分散の閾値を設定した場合は，クラスタ数を指定した場合に比べて精度の変動が大きかったが，最大値では0.43ポイント高い精度を示した．分散の閾値を設定した場合の方が，各カテゴリごとの文書の分布に応じたクラスタリングを行えるためだと考えられる．

5 むすび

今回の提案手法である，サブカテゴリ生成と重みつき類似度を併用することにより，偏りのある文書分布において分類精度が向上することが確認された．

<参考文献>

[1]徳永健伸，情報検索と言語処理，財団法人東京大学出版会，1999.

[2]相澤彰子，語と文書の共起に基づく「特徴量」の定義，情報学基礎57-4 自然言語処理136-4，P25-32，2000.

[3]Chyrch, K. and Gale, W.: Inverse Document Frequency(IDF): A Measure of Deviations from Poisson, Kluwer Academic Pub. P283-295, 1999.

[4]Salton, G.& Buckley, C. Term-Weighting approaches in automatic text retrieval. Information Processig Management, 24(5), P513-523. 1988.

[5]宮本定明，クラスター分析入門，森北出版，1999.

[6]CD-毎日新聞94'データ集，日外アソシエーツ，1995.