The 25th Symposium on Information Theory
and Its Applications (SITA2002)
Ikaho, Gunma, Japan, Dec. 10 - 13, 2002

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　‡鈴木 誠 Makoto Suzuki　†平澤 茂一 Shigeichi Hirasawa

Abstract－Recently, many Japanese texts are made as digital documents. For the purpose of reference or so, they have been classified by a few specialists. However, the necessity of automatic classification technique is increasing, because there are too many digital documents. By the conventional method of automatic classification on the vector space model, only one center-of-gravity vector is assumed on each category. Therefore, if the distribution of texts have more than two peaks, an appropriate classification can not be available. In this research, we propose a new method using clustering for each category and make several sub-categories and sub-category vectors in a category. Then, we show the effectiveness of the proposed method by simulations.

Keywords － automatic classification, text-mining, clustering,

1　はじめに

近年，計算機の普及と共に日本語テキストの機械可読化が進んでいる．機械可読化されたこれらのテキストは，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（インデクス付け等）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計算機上で利用可能となりつつある現在，専門家による手作業による分類だけではその作業量に限界がある。このような背景の下，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．

テキストの自動分類には，予め分類すべき分野（カテゴリと呼ぶ）を人手で設定し，各テキストにこのうちのいずれかのカテゴリを割り当てる手法と，カテゴリを予め設定せず類似した文書集合に分割する手法とに分けられる．本研究では前者の手法を扱うこととし，以下ではこれを単に分類と呼ぶ．

自動分類の手法には様々なモデルによるものが提案されている[1]．それらの手法の中で，本研究で扱うベクトル空間モデルによる手法はテキストに出現した単語の頻度に，ある重み付けを行うことによって頻度を補正する手法で，この手法は最も著名なアプローチの一つである．しかし，従来のベクトル空間モデルによる自動分類手法では，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峰性など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．

本研究では，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サブカテゴリベクトルを生成し，これにより，分類精度を向上させる手法を提案する．また，それに付随して文書とカテゴリとの類似度の計算方法にも改良を加え更なる精度の向上を目指す．最後に，この提案手法を人手によって設定されたカテゴリに分類された新聞記事データ[6]に適用し、分類精度が改善することを示す．

2　従来のテキスト自動分類法

2.1　自動分類の手順[1]

ベクトル空間モデルにおける自動分類の手順は以下のようになる。

① カテゴリを特徴付けるキーワードを抽出する．
② キーワードの出現頻度等を要素とする文書ベクトルを計算する．
③ 各カテゴリにおいてカテゴリベクトルを計算する．
④ 新規文書と各カテゴリベクトルとを比較し，類似度の高いカテゴリへ分類する．

2.2　キーワードの抽出[1]

分類済みの全文書中の単語を対象に単語 t と文書集合 D 間の相互情報量を計算し，値の高いものをキーワードとする[2]．相互情報量が高い単語は文書集合 D の分類に影響の大きい単語であると言える．

定義１（相互情報量：$I(t;D)$ の定義）

$$I(t;D)=(f(t)/F)\times\log(M/df(t)) \qquad -(1)$$

$f(t)$:全文書中での単語 t ののべ出現回数
F:全文書での全単語の総のべ出現回数
M:文書総数
$df(t)$:単語 t が出現する文書数

2.3　文書ベクトルの計算[1]

分類済みの各文書について文書ベクトルを計算する．文書ベクトルはキーワードの重みである TF･IDF 値を文書長で正規化したものを各要素とする[3].

定義２（文書ベクトル：$\mathbf{d}$ の定義）

$$\mathbf{d}=(tf\cdot idf(t_1),tf\cdot idf(t_2),\cdots,tf\cdot idf(t_n))/L \qquad -(2)$$

$t_1,t_2,\cdots,t_n$:抽出された n 個のキーワード
L:文書長（$t_1,t_2\cdots t_n$ の総出現回数）

また，ここで TF･IDF 値は以下の式で表される．

定義３（TF･IDF 値：$tf\cdot idf(t)$ の定義）

† 〒169-8555 東京都新宿区大久保 3-4-1 早稲田大学大学院理工学研究科 経営システム工学専攻. Dept. of Industrial and Management System Engineering School of Science and Engineering, Waseda University, Okubo 3-4-1, Shinjuku-ku, Tokyo, 169-8555 JAPAN.
‡ 〒251-8511 神奈川県藤沢市辻堂西海岸 1-1-25 1410-3 湘南工科大学 工学部 情報工学科. Dept. of Information Science, Shonan Institute of Technology, Tujido Nisikaigan 1-1-25 Fujisawa-shi Kanagawa-ken, 251-8511 JAPAN.

$$tf \cdot idf(t) = f'(t) \times \log(M \cdot df(t)) \quad -(3)$$

$f'(t)$: 文書中での単語 t ののべ出現回数

2.4 カテゴリベクトルの計算[1]

各カテゴリについて，所属する全文書の文書ベクトルの重心をカテゴリベクトルとする．

定義4（カテゴリベクトル：Cの定義）

$$\mathbf{C} = (\mathbf{d}_1 + \mathbf{d}_2 + \cdots + \mathbf{d}_n)/n \quad -(4)$$

n: カテゴリ内の文書数

2.5 新規文書の分類[1]

分類すべき新規文書について，2.3節の方法で文書ベクトル $\mathbf{d}$ を計算する．そして、各カテゴリベクトル $\mathbf{C}$ との類似度を計算し，最も高いカテゴリに分類する．

類似度 sim(d,C) は文書ベクトルとカテゴリベクトルとの余弦を用いる[4]．

定義5（類似度 sim(d,C) の定義）

$$\mathrm{sim}(\mathbf{d},\mathbf{C}) = \frac{\mathbf{d}\cdot\mathbf{C}}{|\mathbf{d}\|\mathbf{C}|} \quad -(5)$$

2.6 従来手法の問題点

従来，カテゴリベクトルはカテゴリ内の文書ベクトル全ての重心を求めるため，必然的に一つのカテゴリは一つのカテゴリベクトルしか持ち得ない．しかし，専門家によってカテゴリが設定さ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偏りが生じる．これはひとつのカテゴリ中にも複数の異なる話題の中心が存在するからであり，それは分布の多峰性を引き起こす．

例えば図1のように「国際」カテゴリの分布に偏りがある場合，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．

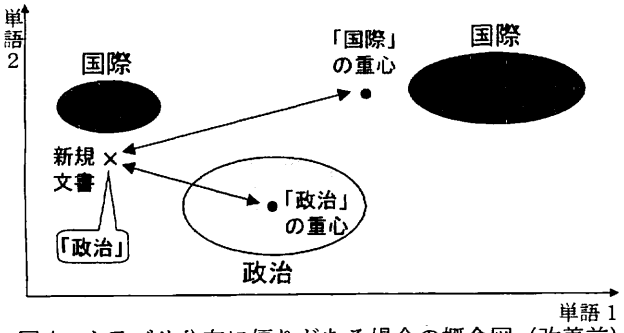

図1　カテゴリ分布に偏りがある場合の概念図（改善前）

3 提案手法

本研究では，各親カテゴリ毎に分類済み文書データに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．（ここで親カテゴリとは，クラスタリングによって生成されたカテゴリをサブカテゴリと呼ぶのに対し，元のカテゴリを表すこととする．）それによって，文書分布に偏りがある場合にも，分布に適応したサブ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．

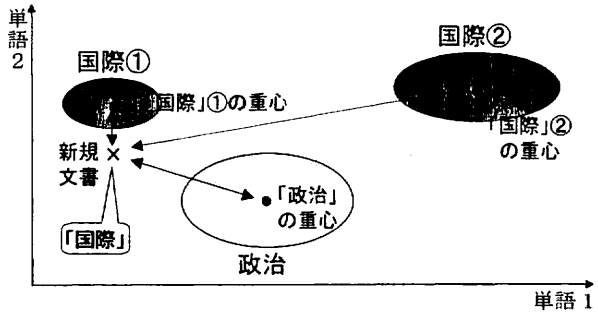

図2　カテゴリ分布に偏りがある場合の概念図（改善後）

3.1 提案手法の分類手順

① カテゴリを特徴付けるキーワードを抽出する．
② キーワードの出現頻度等を要素とする文書ベクトルを計算する．
③ 各親カテゴリ内文書に対するクラスタリングを行い，サブカテゴリを生成する．
④ 各サブカテゴリにおいてサブカテゴリベクトルを計算する．
⑤ 新規文書と各サブカテゴリベクトルとを比較し，類似度の高いカテゴリへ分類する．

3.2 カテゴリ内文書に対するクラスタリング

与えられた各親カテゴリ内文書に対しクラスタリングを行う．本研究ではクラスタリングの手法として凝集法[5]を用いる．この方法は各クラスタ間の類似度を求め，最大の類似度をとる2つのクラスタを併合し1つのクラスタとする方法で，クラスタ数が任意の整数 k になるまで併合を繰り返す．この手法には局所解に陥ることがないという特徴がある．

（手順1）親カテゴリ内の各文書を初期クラスタとする．各文書ベクトル $\mathbf{d}$ から各クラスタの重心ベクトル $\mathbf{c}$ を求め，各クラスタの重心ベクトル $\mathbf{c}_i, \mathbf{c}_j$ 間の類似度を計算する．類似度 $\mathrm{sim}(\mathbf{c}_i, \mathbf{c}_j)$ は以下の式で求める．

定義6（類似度 $\mathrm{sim}(\mathbf{c}_i, \mathbf{c}_j)$ の定義）

$$\mathrm{sim}(\mathbf{c}_i,\mathbf{c}_j) = \frac{\mathbf{c}_i \cdot \mathbf{c}_j}{|\mathbf{c}_i\|\mathbf{c}_j|} \quad -(6)$$

（i, j は任意のクラスタ番号）

（手順2）類似度の最も高い2つのクラスタを併合し，新しい1つのクラスタとする．新しいクラスタの重心ベクトルは所属する文書ベクトルの重心とする．

（手順3）各クラスタ内の分散を計算し，分散が閾値を超えるまで，あるいはあらかじめ設定した最大数を超えるまで，手順1～2を繰り返す．クラスタリングが終了した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 $\mathbf{c}$ をサブカテゴリベクトル $\mathbf{c}'$ とする．

3.3 サブカテゴリに応じた文書の分類

まず始めに，新規文書の文書ベクトル $\mathbf{d}'$ を式(2)により計算し，各サブカテゴリベクトル $\mathbf{c}'$ との類似度を求める．この時，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．もし，文書数によらずに各サブカテゴリを一定に扱ってしまう場合，分類結果が文書数の少ないサブカテゴリ（外れ値）に

影響を受けやすくなってしまう．

そこで，サブカテゴリに所属する文書数を新規文書とサブカテゴリとの類似度に乗算した重みつき類似度を定義する．また，パラメータ p によりその重みを調整する．これにより文書数の多いサブカテゴリを優先して分類を行うことが出来る．

定義７（重みつき類似度 wsim(d′,c′) の定義）

$$\mathrm{wsim}(\mathbf{d}',\mathbf{c}') = \frac{\mathbf{d}'\cdot\mathbf{c}'}{|\mathbf{d}'||\mathbf{c}'|} \times D^p \qquad -(7)$$

D：サブカテゴリ中の文書数

また，従来手法では図４のように各カテゴリを一つのクラスタと考えているため大局的な分布を表していた．しかし，クラスタリングを細かくすると，図５のようにモデルを細かく表すことが出来る反面，外れ値等の局所的な分布に影響されやすい過学習の状態になることが考えられる．

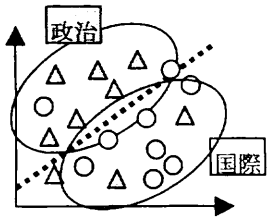

図４　文書モデル概念図（従来手法）

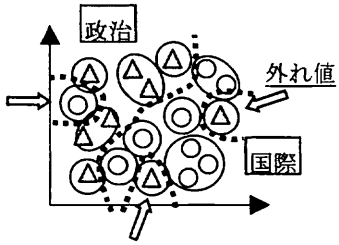

図５　文書モデル概念図（提案手法）

そこで，外れ値の影響を受けにくいよう新規文書と各親カテゴリとの類似度を以下の式のように所属サブカテゴリへの類似度の平均で定義する手法を提案する．

定義８（親カテゴリへの類似度）

$$\mathrm{sim}(\mathbf{d}',\mathbf{C}) = \sum_{i=1}^{m} \mathrm{wsim}(d',c_i')/m \qquad -(8)$$

m：親カテゴリ内のサブカテゴリ数

これにより文書分布全体を考慮できるようになるため，局所的な分布に影響されにくくなると考えられる．下にこの考えに関する例を示す．

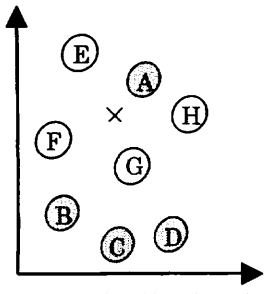

図６　外れ値のある文書モデル概念図

表１　各サブカテゴリへの類似度

サブカテゴリ	×との類似度
A（政治）	3
B（政治）	1
C（政治）	1
D（政治）	1
E（国際）	2
F（国際）	2
G（国際）	2
H（国際）	2

図６のような文書モデルが存在し新規文書×の各サブカテゴリへの類似度が表１のように計算されているとする．従来手法では新規文書は単純に最も類似度の高いサブカテゴリ A のカテゴリである「政治」のカテゴリに分類されるが，文書分布全体を見ると A は外れ値であり新規文書は「国際」のカテゴリに分類されるのが妥当である．そこで定義７のように各サブカテゴリへの類似度の平均を計算すると，「政治」カテゴリへの類似度はサブカテゴリ A から D の平均で，(3+1+1+1)/4=1.5 となり，「国際」カテゴリへの類似度は E から H の平均で，(2+2+2+2)/4=2 となる．するとこの新規文書は「国際」カテゴリに分類される．

このように平均の類似度を用いることにより外れ値の影響を抑えることが出来ると考えられる．

４　提案手法の評価と考察

４．１　利用データと実験方法

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．このデータ集は人手により分類結果が示されている．今回の実験では全データの内，９カテゴリの文書を使用した．

性能の評価には分類精度という尺度を用いた．分類精度とは，分類を試みた全文書数に対して正しく分類できた文書数の割合を示す指標である．

４．２　従来手法と提案手法の比較

４．２．１　実験方法

今回の実験では学習データとして各カテゴリ 1000 文書（合計 9000 文書），テストデータとして各カテゴリ 500 文書（合計 4500 文書）を無作為に抽出し使用した．また，抽出キーワード数は 1000 単語とした．

さらに，サブカテゴリを生成する提案手法においても，式（８）で定義した平均の類似度を使用する場合と使用しない場合の２通りを実験した．これにより平均の類似度を用いることの有効性も同時に検証する．

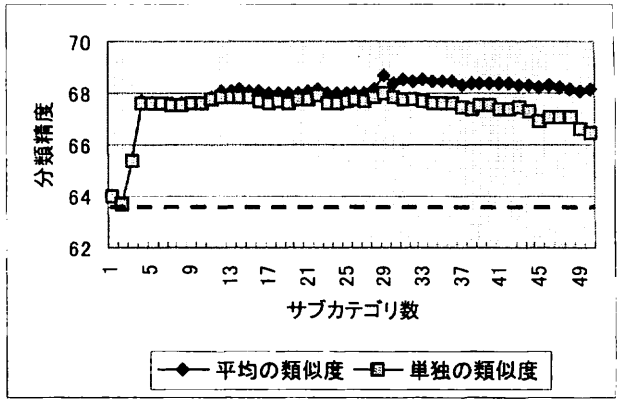

図７　従来手法と提案手法の性能比較

４．２．２　実験結果と考察

(1)提案手法全体の有効性

まず，従来手法と提案手法の性能を実験により比較した．図７は従来の分類手法と本論文での提案手法の性能を比較したグラフである．従来手法による分類精度は 63.9%（図中の点線部分）であった．それに対し提案手法ではほとんど全てのサブカテゴリ数の場合において従来手法の性能を上回った．

提案手法の分類精度の最大値はクラスタ数 29 における平均の類似度を使用した場合で 68.0%を記録し，従来手法

の分類精度を 4.1%上回った．これによりサブカテゴリを生成する本提案手法の有効性が確認されたと言える．

(2)平均の類似度

提案手法において，単独の類似度を使用した場合に比べ，平均の類似度を使用した場合の方が性能がよく，それはサブカテゴリ数が増えるに従い顕著に差が現れた．これは3.2節で述べたように，クラスタリングをより細かく行うにつれて外れ値の影響が大きくなるためだと考えられ，平均の類似度を用いることの有効性が確認されたと言える．

4.3 使用する文書数及びキーワード数と分類精度の関係

4.3.1 実験方法

次に，使用する文書数が分類精度にどのような影響を与えるかについて実験を行った．図8は本提案手法において実験に用いる文書数を変化させた場合の提案手法の実験結果である．抽出キーワード数は1000単語に固定した．

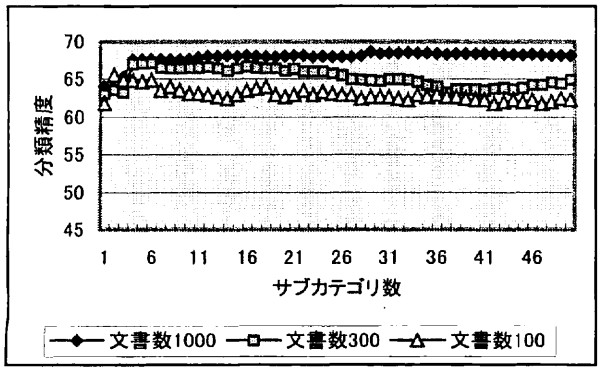

図8　提案手法における使用文書数と分類精度の関係

さらに，使用するキーワード数についても分類精度に対する影響を調べた．図9はキーワード数を変化させた場合の提案手法の実験結果である．抽出文書数は学習データを各カテゴリ1000文書，テストデータを各カテゴリ500文書に固定した．

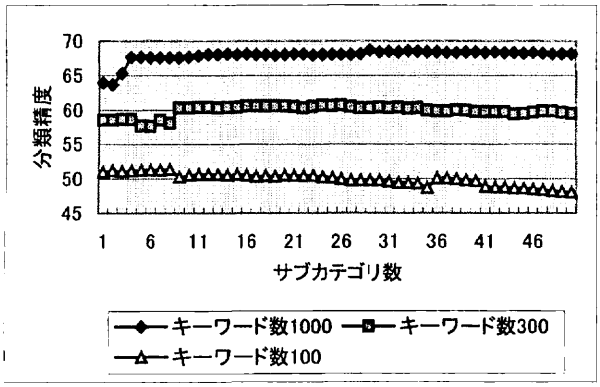

図9　提案手法における使用キーワード数と分類精度の関係

4.3.2 実験結果と考察

図8からは使用文書数が多いほど分類精度が上がることが，また図9からはやはり使用キーワード数が多いほど分類精度が上がることがわかった．

さらに，文書数の分類精度に対する影響に比べると，キーワード数が分類精度に与える影響のほうが大きいことがわかった．つまりこのような文書自動分類システムにおいては，サンプルデータ数を多く集めることよりも，キーワードを大量に抽出することの方が重要であると言える．これは，特に短い文書の場合に抽出キーワード数が少ないと，文書中にほとんどキーワードが出現せず，十分有効な文書ベクトルが作成できないためだと考えられる．実際にキーワード数が100単語の場合などは，1文書中にキーワードが一つも出てこない文書も存在した．

5 むすび

今回のシミュレーション結果により，従来手法に対するサブカテゴリを生成する提案手法の有効性が示された．また，それに付随して改良した新規文書とカテゴリとの類似度計算手法の有効性も明らかになった．

最適なサブカテゴリ数に関してはデータに依存するものであり，探索的にかつ全てのカテゴリについて同様に求めていた．しかし，データの構造を解析することにより，自動的にかつ独立的に求めることも可能であると思われ，今後も研究の余地があると言える．

今回の従来手法としてはごく単純なベクトル空間モデルに基づく自動分類手法を使用した．しかし，単語の共起情報をした手法[7]，意味解析を利用した手法[8]，辞書を用いた手法[9]など様々な研究がベクトル空間モデルに基づく自動分類に対してなされ有効性が示されてきている．本手法はそれら他手法と相反することなく並行して利用することが可能であり，本手法と組み合わせて使用することにより更なる精度の向上が望まれる．

さらに，今後の方針としては本手法の有効性を示すために他の実データに適用し，一般性があるかどうか，どのような性質を持つ実データに適用できるかなどを検証することが考えられる．

＜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 Processing 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.

[7]藤井洋一，鈴木克志，辻秀一，段落内共起情報を利用した文書自動分類方式，情報処理学会ジャーナル Vol.42 No.03，2001.

[8]河合敦夫，意味属性の学習結果に基づく文書自動分類方式，情報処理学会ジャーナル Vol.33 No.09，1992.

[9]呉勇，山田祥，岸本陽次郎，文書自動分類のための分野関連後辞書の構成，情報処理学会 研究報告 情報学基礎 No.057，1999.