自然言語処理 106－2
（1995．3．9）

分類体系相互の関係を利用したテキストの自動分類

山本 和英＊　　増山 繁＊　　内藤 昭三＊＊
{yamamoto,masuyama}@smlab.tutkie.tut.ac.jp　naito@slab.ntt.jp

＊豊橋技術科学大学 知識情報工学系
＊＊NTT ソフトウェア研究所

概要

本稿では，分類体系相互の関係を利用した日本語テキストの自動分類手法を提案する．従来の分類手法は，表記の統計情報を用いた手法と分類体系に依存した情報を用いた手法の二つに大別できるが，本手法ではシソーラスによる統計情報から分類体系相互の関係を自動学習するという両者の中間的な手法を用いることで，前者の分類精度の低さと後者の汎用性の低さの問題点を同時に解決した．本手法を用いて合計 1260 記事の新聞コラムを対象にした 10 カテゴリーへの分類を行った結果，2 カテゴリーに対して 100% の再現率，別の 2 カテゴリーで 100% の適合率となり，全体の平均でも約 95% の正解率となった．

An Automatic Classification Method for Japanese Texts using Mutual Category Relations

Kazuhide YAMAMOTO*　Shigeru MASUYAMA*　Shozo NAITO**

* Dept. of Knowledge-based Info. Eng., Toyohashi Univ. of Tech.
** NTT Software Laboratories

Abstract

This paper proposes an automatic text classification method using mutual category relations. Conventional texts classification methods can be roughly divided into two approaches; one depending statistics and the other depending on some category system. We adopt an intermediate method of the above two approaches. Our new method automatically learns mutual category relations from statistical information obtained by a thesaurus. Our method overcomes the following two defects that former methods have; the insufficient accuracy of the former approaches and the insufficient generality of the latter. As an experiment we classify 1260 Japanese newspaper columns into ten categories using our method. The results show total accuracy of about 95%, while two categories out of ten attain recall rate of 100% and another two categories attain precision rate of 100%.

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．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きている．

テキストの自動分類手法は，あらかじめ分類すべき分野(以下カテゴリーと呼ぶ)を設定し，各テキストにこのうちのいずれかの分野を割り当てる手法[Kaw92, Kes93, Yua93]と，カテゴリーをあらかじめ設定しない手法(クラスタリング)[Tsu94]の二つに分けられる．本稿では，前者のタイプの自動分類を対象とする．

従来提案されてきた日本語のテキストをカテゴリーに分類する手法は，以下の2種類に大別できる[Kaw92, Wat94]．

1. 表記の統計情報を用いた手法[Tam88]

2. 分類体系に依存した情報を用いた手法[Kam87]

前者は処理が簡単で汎用性が高いが，分類精度が低くなるとされる．一方，後者の方法は高い分類精度を達成する反面，分野の拡張，変更が困難と考えられていた．

本稿で提案する手法は，テキストに出現した語をシソーラスで分類し，その頻度情報をカテゴリー相互の関係によって加工した特徴ベクトルを用いてテキストの分類を行う．本手法は，シソーラスと統計情報のみを使用し，処理は簡単であり，またカテゴリーに依存した情報は統計情報から自動作成できるので，汎用性が高い．このことは，従来の2手法のうちの前者の手法の長所を保ちながら，後者の手法の欠点を解消していることを意味し，両者の中間的な手法に位置付けることができる．

芥子らは文脈ベクトルを用いた連想検索手法を提案している[Kes93]．この手法は，数百の出現頻度の高い語について手作業で文脈ベクトルを作成して，重要単語の文脈ベクトルを機械学習するというものである．この手法では，人手を介することで労力がかかり，文脈ベクトル作成の際に個人差による揺れが生じる可能性がある．また，ベクトルの要素となる語の選択という問題が生じる．

統計的な情報を用いた分類としては，漢字を単位にした方法[Wat94]，名詞の共起関係を使用した方法[Yua93]，キーワードのχ^2値を利用した方法[Tam88]などが提案されている．また，シソーラスを用いた日本語テキストの自動分類については[Kaw92]の手法がある．河合は，各カテゴリーごとに偏って出現する意味属性をあらかじめ自動学習し，その結果を用いてテキストの自動分類を行っている．本稿で提案する手法はシソーラスを用いる点では河合の手法と共通するが，以下に示す相違点を持つ．

- カテゴリー相互の相対的な関係によって特徴ベクトルを作成する．
- シソーラスの分類項目別に集計した情報だけを用いて単語別の統計情報は使用しないため，記憶容量の消費が少ない[1]．
- シソーラスが階層型である必要がないため，ネットワーク型などその他の形状であっても対応でき，汎用性が高く，拡張性にも優れている．この性質により，シソーラスに専門用語，固有名詞などの分類項目を自由に追加できる．

以下では，2章で本手法の内容について述べる．3章では本稿で行った実験の内容，及び結果を示す．4章では実験で正しく分類できなかったテキストについて，5章では本手法を定性的に，それぞれ考察を行う．最後に6章でまとめを行う．

2 分類手法

ここでは，本研究で提案する手法について説明する．本手法は大きく以下の三つのプロセスに分かれる．

1. カテゴリーの特徴ベクトルの作成

2. テキストの特徴ベクトルの作成

3. テキストと各カテゴリーとの類似度計算

図1に，本手法の大まかな処理の流れを示す．

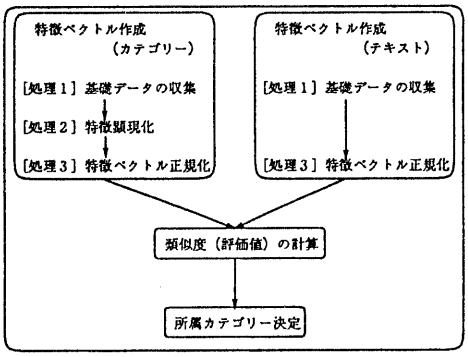

図1: 本手法全体の処理の流れ

2.1 カテゴリーの特徴ベクトルの作成

基礎データの収集 まずカテゴリーごとにそのカテゴリーの基礎データを収集する．基礎データの対象となるテキストは，そのカテゴリーの特徴を最も典型的に表しているテキスト1記事を選ぶ方法も考えられるが，典型的な1記事を選ぶことは容易ではない．そこで，本研究ではカテゴリー対象のテキスト中から無作為に10記事を抽出し，それらの平均を基礎データとする方法を採用した．

[1][Kaw92]では，単語別に集計した情報も併用して用いたほうが高い精度が得られている．

まず選択した10テキストに対して形態素解析を行って，単語をシソーラスの分類項目別に集計する．ここで，同表記で複数の意味を持つ語(多義語)，および同一の意味であっても複数の分類項目に所属している語については，分類項目を自動特定することが現状では困難であるという背景から，その語が所属するすべての分類項目について，単一分類項目の時と同一の頻度を追加した．

ただし，テキストの長短の差が大きい時は，長いテキストはテキスト全体を対象にはせず，適度な位置まで(最も短いカテゴリーのテキスト長×2)を基礎データの収集対象にした．これは，頻度よりも出現そのものを重要視する本手法の特性のために，テキスト長にあまり差がある場合には本手法が有効に機能しないからである．

[処理1] 各カテゴリー別に，無作為抽出したテキストを形態素解析し，シソーラスの分類項目別に集計する． ■

以下では，予め定められたカテゴリーを$\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_M$ (M: 分類するカテゴリー数)，カテゴリー$\mathcal{C}_i$ $(0 \leq i \leq M)$に属するテキストに含まれている全単語をシソーラスによって分類，集計した結果を$\mathcal{C}_i = (e_{i1}, e_{i2}, \cdots, e_{iN})$と表記する($N$: シソーラスの分類項目数)．

特徴顕現化 テキスト中に出現する語の中には，分類に有効な語とそうでない語が含まれている．テキストを有効に分類するためには，これらの語のうち，分類に有効な語，あるいは分類項目をより目立たせる必要がある．本処理では，分類体系相互の関係を使用して特徴の顕現化処理を施す．

ここでは，前処理と同様に分類項目iの出現したカテゴリー数に着目する．例えば，ある分類項目iが半分以上のカテゴリーに出現した場合と，分類項目iがある一部のカテゴリーにしか出現しなかった場合を比較すると，後者の場合はテキストを分類する際の非常に重要な情報となり得る．このように，分類項目iの出現したカテゴリー数が少なくなるほど，その分類項目は重要な役割を果たし，唯一のカテゴリーの場合にその重要性が最大になる．ここでは，このような効果を与えるために以下の処理を行う．

[処理2] 分類項目jについて，$e_{ij} > 0$であるカテゴリー数を$m(0 \leq m \leq M)$[2]としたとき，すべての$\mathcal{C}_i(1 \leq i \leq M)$のすべての要素$e_{ij}(1 \leq j \leq N)$について，以下の計算を行う．

$$e_{ij} \leftarrow e_{ij}(M - m) \tag{1}$$

以上の処理をすべての分類項目$j(1 \leq j \leq N)$に対して行う． ■

[処理2]のうち，$m = M$の特別な場合は以下の[処理2-1]に該当する．

[処理2-1] 分類項目jについて，もしすべての$\mathcal{C}_i(1 \leq i \leq M)$の要素で$e_{ij} > 0$であるならば，すべての$\mathcal{C}_i(1 \leq i \leq M)$の要素$e_{ij}$について，以下の処理を行う．

$$e_{ij} \leftarrow 0 \tag{2}$$

以上の処理をすべての分類項目$j(1 \leq j \leq N)$に対して行う． ■

この処理は，全カテゴリーに共通して出現が見られる分類項目に対してその値を0にする処理であり，カテゴリーの弁別効果を持たない分類項目の，特徴ベクトル全体への影響を排除する効果を持っている．

全カテゴリーに出現の見られる分類項目に対して，どのカテゴリーにどれだけ出現したかという頻度情報は特徴ベクトル作成の有力な情報とはなり得ず，むしろこれらの分類項目は一般に高頻度であるために悪影響を与える可能性が強いと考えられる．このため，これらの情報は無視する方が適当と考えられ，本稿ではこれらの語の情報は特徴ベクトルの作成には反映させなかった．

特徴ベクトルの正規化 最後に，以上のようにして得られた特徴ベクトルの長さが同一になるように，以下の[処理3]に示す正規化を行う．

[処理3] すべての$\mathcal{C}_i(1 \leq i \leq M)$のすべての要素$e_{ij}(1 \leq j \leq N)$について，以下の計算を行う．

$$e_{ij} \leftarrow \frac{e_{ij}}{l_i} \tag{3}$$

ただし，$l_i = \sqrt{\sum_{k=1}^{N} e_{ik}^2}$とする． ■

2.2 テキストの特徴ベクトルの作成

これから分類すべきテキストについて，カテゴリーと同様に特徴ベクトルを作成する．ただし，図1に示すように，カテゴリーの特徴ベクトル作成における[処理2](カテゴリー間比較のための処理)は必要ないので，[処理1]と[処理3]を行う[3]．

2.3 類似度の計算

前節に示す処理で得られたカテゴリーの特徴ベクトルとテキストの特徴ベクトルを比較し，その類似度(評価値)を見ることで，テキストが属するカテゴリーを決定する．本研究で使用する類似度を，以下で定義する．

[定義1] カテゴリー$\mathcal{C}_i$とテキスト$\mathcal{T}$の特徴ベクトルがそれぞれ，$\mathcal{C}_i = (e_{i1}, e_{i2}, \cdots, e_{iN})$，$\mathcal{T} = (t_1, t_2, \cdots, t_N)$であるとする．カテゴリー$\mathcal{C}_i$とテキスト$\mathcal{T}$の類似度$eval(\mathcal{C}_i, \mathcal{T})$を以下の式(内積)で定義する．

$$eval(\mathcal{C}_i, \mathcal{T}) = \sum_{k=1}^{N} e_{ik} \cdot t_k \tag{4}$$

■

[2] 形式的に，$m = 0$の場合(すべてのカテゴリーで$e_{ij} = 0$の場合)も含めて[処理2]とした．

[3] テキストの自動分類を行うだけなら[処理3]も必要ないが，異なるテキストで評価値の比較を行うために，本研究の実験では[処理3]も行った．

2.4 未知語の取扱い

本研究では，解析対象をシソーラスに掲載されている語のみに限定した．シソーラスを使用することで，シソーラスにない未知語，すなわち新語，特殊な専門用語，多くの固有名詞などを取り扱うことが不可能になる．

一般的に，専門用語，固有名詞などはしばしばテキストの特徴を反映し，そのテキストの分野を特定する上で重要度の高い要素である可能性がある．しかしこれらの語は非常に多くあるため，汎用性を持ったテキストの自動分類を行うためには，これらの語に依存しない手法が望ましいと考えた．また，固有名詞は分類の有力な手がかりにはならないという指摘 [Kaw92] もあるため，本研究ではあえてこれらの語を使用せずに分類を行うことを試みた．

3 実験

本手法の有効性を確認するために，テキストのカテゴリーへの分類を自動的に行う実験を行った．ここでは，この実験について述べる．

3.1 実験内容

実験に使用したテキストはいずれも日本経済新聞の 1990 年，および 1992 年に掲載されたコラムである．実験に使用したコラムとその記事数を表 1に示す．ただし表の「記事数」には，カテゴリーの特徴ベクトル作成に使用した 10 記事を含む．

カテゴリー	コラム名	記事数
つり	「つり」	59
証券	「まちかど」	93
医学	「医フロンティア」「くすり百科」	94
政治	「92 選挙駆ける」	98
税金	「税金相談」	103
家庭	「家族はいま」	116
音楽	「音楽」	154
グルメ	「味」「味力」	195
新製品	「ニューフェース」	200
経済学	「やさしい経済学」	248

表 1: 実験に使用したコラム

実験では，コラム名に応じて表 1に示す 10 個のカテゴリーを設定した．コラムによって記事数が異なるのは，前述した年度に掲載された同タイトルのコラムをすべて使用しているためである．ただし，コラム「ニューフェース」は 1 記事のテキスト長が極端に短く記事数が多いので，1 日分をまとめて 1 記事とし，テキスト長の長い上位 200 記事を使用した．

シソーラスには角川類語新辞典 [Oon81] を使用した．同辞典は日本語の語彙を 10 種類の大分類に分類し，さらに各大分類を 10 種類の中分類，各中分類を 10 種類の小分類へと階層的に分類している．本実験では，この辞典の 1000(10^3) の小分類を「分類項目」として設定した．よって，各カテゴリー，及びテキストの特徴ベクトルは 1000 次元のベクトルとなる．ただし，実際にはどのカテゴリーにも出現しない分類項目は除いているので，実際の次元は 1000 未満となる．

また形態素解析は，独自に作成して文献 [Yam95] で使用したシステムをそのまま使用した．

3.2 実験結果

100 記事のコラム (各カテゴリー 10 記事×10 カテゴリー) について，2.1節に示した処理を行った結果，実際に使用した (語の出現があった) 分類項目数は 380 となった．これは各特徴ベクトルが 380 次元のベクトルであったことを意味する．

表 2に 2.1節 [処理 2-1] の対象となった分類項目を示す．表 2に列挙されている分類項目はいずれも全カテゴリーに対して出現が予想される分類項目ばかりであり，またここに挙げられていない分類項目でも，多くのカテゴリーに共通する分類項目は出現したカテゴリー数が多かった (つまり，[処理 2] で m が大きかった) ことから，カテゴリーの弁別効果を持たない分類項目の除去には成功していることが確認できる．

こそあど (101)	内外 (103)	数 (120)	多少 (126)
時機 (151)	先後 (156)	今昔 (158)	同一 (188)
等級 (192)	限度 (194)	大変 (195)	こんな (199)
思考 (411)	世界 (709)	番号 (823)	単位 (828)
助数詞 (829)	接辞 (834)		

表 2: 全カテゴリーに出現した分類項目名 (分類番号)

表 3には，[処理 1] の結果出現したカテゴリーが唯一だった分類項目のうち，カテゴリー別に出現頻度の高かった上位 3 項目をそれぞれ列挙する．表 3には，そのカテゴリーに非常に関係の深い分類項目が必ずしも列挙されているわけではなく，相対的に見てそのカテゴリーのみの出現が予想される分類項目が多い．このことから，本手法はこのような複数の分類項目によってカテゴリーの特徴が捉えられていることがわかり，カテゴリー間の相対的な関係を抽出する本手法が有効に機能していることがわかる．

カテゴリー	分類項目名	(分類番号)	
つり	欺瞞 (456)	狩猟 (398)	川 (036)
証券	札 (974)	行為 (360)	証明 (418)
医学	薬剤 (910)	内臓 (067)	光学器械 (993)
政治	推挙 (778)	党派 (715)	選択 (378)
税金	取捨 (373)	従業 (364)	用地 (042)
家庭	家庭 (717)	親族 (528)	学校 (722)
音楽	楽曲 (874)	音楽 (870)	演奏 (871)
グルメ	風味 (144)	野菜 (927)	料理 (923)
新製品	機械 (990)	球技用語 (899)	写真 (864)
経済学	村落 (707)	奉仕 (795)	応対 (787)

表 3: 単独のカテゴリーに出現した主な分類項目

カテゴリー	正解	2	3	4	5	6	7	8	9	10	r_i	p_i
新製品	190	0	0	0	0	0	0	0	0	0	100.00%	94.06%
税金	93	0	0	0	0	0	0	0	0	0	100.00%	76.23%
家庭	105	1	0	0	0	0	0	0	0	0	99.06%	87.50%
政治	87	1	0	0	0	0	0	0	0	0	98.86%	95.60%
つり	48	1	0	0	0	0	0	0	0	0	97.96%	97.96%
証券	80	2	1	0	0	0	0	0	0	0	96.39%	98.77%
音楽	138	4	0	0	1	1	0	0	0	0	95.83%	100.00%
医学	77	5	0	0	1	1	0	0	0	0	91.67%	98.72%
グルメ	169	6	4	0	2	1	2	0	1	0	91.35%	100.00%
経済学	204	25	4	3	1	1	0	0	0	0	85.71%	97.14%
1260 記事	1191	45	9	3	5	4	2	0	1	0	94.52%	94.52%

表 4: 実験結果

次に実験結果を表4に示す．ただし表中で，「正解」とは「正しいカテゴリーに分類した (評価値で 1 番目になった) 記事数」，「$n(2 \leq n \leq 10)$」は「評価値で n 番目になった記事数」を示す．また，各カテゴリーの再現率と適合率を以下の式で定義する．

[定義 2] カテゴリー C_i の再現率 r_i，適合率 p_i を以下の式で定義する．

$$r_i = \frac{correct_i}{original_i}, \quad p_i = \frac{correct_i}{result_i} \tag{5}$$

ただし，$correct_i$: カテゴリー C_i のテキストの内，カテゴリー C_i に分類された記事数，$original_i$: カテゴリー C_i の原記事数，$result_i$: 実験でカテゴリー C_i に分類された記事数とする． ■

また，本実験では最も高い評価値となったテキストの場合のみを正解としているので，全カテゴリーの再現率と適合率は一致する．以下では，この全カテゴリーの再現率 (=全カテゴリーの適合率) を，正解率と呼ぶ．

本実験では，表4に示す通り約 95% の正解率となった．個々のカテゴリーについても，9 カテゴリーの再現率，8 カテゴリーの適合率が 90% 以上という結果となった．また，正解とならなかった 69 記事のうちの 45 記事 (65%) は評価値 2 位であるので，仮に上位 2 位までの出力を正解とすると全体で 98.1% の再現率であることを意味する．以上の実験結果は，他の文献とは実験環境 (対象とするテキスト，分類するカテゴリー数など) が異なるため単純に比較することはできないが，非常に精度が高く，実用面でも十分耐え得るものであるといえる．

4 考察

ここでは，本研究の実験で正しく分類されなかったテキストの計 69 記事について検討する．

カテゴリー別の分類では，カテゴリー「経済学」を「税金」と誤って判断したものが 18 記事と，最も多かった．個々の誤ったテキストをみると，テキストの内容そのものが，正しいカテゴリーと誤って判断したカテゴリーの両者の内容を共に含んだ内容のものが多く，誤って判断されたテキストの多くはこのことが原因と考えられる．

一方で，これとは異なる原因で誤ったテキストも存在する．例えば「経済学」を「医学」と誤って特定したテキストは「道路混雑の政治経済学 (3) 創価大学教授岡野行秀氏」[1990 年 9 月 19 日][4]である．このテキストは道路の交通量について数学的に述べられたテキストであり，「医学」とは関係のないテキストである．またこのテキストには (交通，密度，速度) などの単語が特に多く使用されていた．このテキストが「医学」と判断された原因は，テキストに多用された前述の語のうち「交通」という語は特徴ベクトル作成の際のどの学習テキストにも出現せず，「密度」と「速度」が属する分類項目 (分類番号 122) は，「医学」のカテゴリーの基礎データのみに出現していたためである．このため，カテゴリー「医学」の類似度が高くなったと考えられる．

次に，本研究の実験で評価値順位の最も低かった (9 位) テキストを検証する (表4下線部分)．このテキスト「川波，大阪」[1990 年 7 月 18 日] はカテゴリー「グルメ」のテキストであり，内容はウナギの店の紹介である．このテキストが「グルメ」で高い評価値が得られなかったのは，シソーラスには漢字表記「鰻」しかない，という点と，テキスト中に「グルメ」に主に影響する語が (ウナギ以外に) 少なかった点の二つが主な理由と考えられる．またこのテキストは，他テキストよりも全カテゴリーで低い評価値が与えられており，いわば「特徴のないテキスト」であったと位置付けることができる．このため，このテキストに頻出した「〜円 (特に新製品に影響，以下同様)」「〜時 (税金，医学)」「〜分 (新製品)」「横綱・大関・関脇・小結[5](証券)」などの語が他のカテゴリーに影響し，結果として「グルメ」のカテゴリーの評価値が相対的に低いものとなったことが原因と考えられる．

[4] タイトルは日本経済新聞 CD - ROM 版に付されていたタイトル，日付は原テキストの新聞掲載日を示す．以下同様．

[5] テキストでは，いずれも店のメニューとして使用されている．また，これらの語が属する分類項目は「地位 (682)」である．

5 議論

文献[Kaw92]での手法において，正しいカテゴリーに分類できない要因として，河合は以下の3点を指摘している．

1. 使用したシソーラスの分類項目では異なる分類項目に分散して出現したことにより，傾向が捉えられなかったため
2. 多義語に対して意味属性を絞り込んでいないため
3. テキスト中の比喩，慣用表現の使用のため

以下では，これらの要素について順に検討する．

まず最初の要素であるが，[Kaw92]の例では，(出土品，発掘，文化財)という単語に対してそれぞれ(物品，採取，財産)の分類項目となるため，これらの単語から分野(考古学)を導き出すのは困難であるとしている．一方本手法では，他のカテゴリーで「採取」や「財産」などの分類項目が出現していなければ，これらの分類項目はカテゴリー特定の有力な判断材料となる．また，仮に出現していてもその頻度が他のカテゴリーでまれである場合も判断材料となる．このように，複数の分類項目に分散して出現することは，本手法においてはカテゴリーの特徴が複数存在することに相当する．従って，仮にこのうちいくつかの分類項目で他カテゴリーでの出現が見られその分類項目が特徴でなくなったとしても，他の特徴が存在するため，全体としてカテゴリーを的確にとらえている可能性が高い．

このように，複数の分類項目に分散出現することは，むしろ本手法に対して有効に働く．一般のテキストは，必ずしも出現する分類項目に偏りがあるとは限らないので，本手法は有効であると考えられる．

次に語の多義性の問題であるが，多義語の意味特定は困難であるので，本手法においても意味の特定は行っていない．しかし，カテゴリーの特徴ベクトル作成時に例えば「スポーツ」のカテゴリーで「アンカー」という語が使用された場合，もう一つの意味である「いかり」の属する分類項目に他のカテゴリーで出現がなければ，カテゴリー特定の妨げにはならない．このことから，多義性の問題は本手法においても依然として存在するが，悪影響を及ぼす可能性は比較的低い．

また第三の要因である慣用表現については，その表現が一般的に多用されるものであればそれがどのカテゴリーの特徴ともならないため問題なく，まれに出現するものであればその語が全体に及ぼす影響は低いと考えられるためこれも問題ない．また，ある表現がある特定のカテゴリーのみで頻出した場合は，その表現がどのような表現であってもカテゴリー特定の判断材料となり得る．一方で，テキスト中で多種多様な慣用表現を使用している場合には本手法が有効に機能しない可能性があるが，そのようなテキストは特殊なものであると考えられる．以上より，慣用表現についても悪影響を及ぼす可能性は比較的低い．

6 おわりに

本稿では，分類体系相互の関係を利用した日本語テキストの自動分類手法を提案した．この手法を用いて合計1260記事の新聞コラムを対象にした10カテゴリーへの分類を行った結果，全体の平均で約95%が正しく分類できた．なお，実験は日本語のテキストを対象にして行ったが，本手法は対象とする言語のシソーラスさえあれば任意の言語に適用可能であり，本手法の一般性は高い．

あるテキストを特徴づけることは他テキストとの比較によってはじめて可能となる．本手法はこの「相対性」の精神に基づいた手法であると位置付けできる．キーワードの自動抽出や文章の抄録・要約の作業も，文章分類と同様にこの相対性の要素を持っていることから，今後は本手法のこれらへの適応が課題である．

謝辞

本研究で，シソーラスに使用した「角川類語新辞典」[Oon81]を機械可読辞書の形で提供いただき，その使用許可をいただいた(株)角川書店に深謝する．

参考文献

[Kam87] 亀田弘之，藤崎博也：テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム，情報処理学会論文誌，Vol. 28, No. 11, pp. 1103–1111 (1987).

[Kaw92] 河合敦夫：意味属性の学習結果にもとづく文書自動分類方式，情報処理学会論文誌，Vol. 33, No. 9, pp. 1114–1122 (1992).

[Kes93] 芥子育雄，乾隆夫，石鞍謙一郎：大規模文書データベースからの連想検索，技術研究報告 AI92-99，電子情報通信学会 (1993).

[Oon81] 大野晋，浜西正人：角川類語新辞典，角川書店 (1981).

[Tam88] 田村淳，渡辺道枝，原良憲，笠原裕：統計的手法による文書自動分類，全国大会論文集 36-6U5，情報処理学会 (1988).

[Tsu94] 津田宏治，仙田修司，美濃導彦，池田克夫：共起関係の固有ベクトルを用いる単語クラスタリング法，研究会資料 NL103-6，情報処理学会 (1994).

[Wat94] 渡辺靖彦，竹内雅人，村田真樹，長尾眞：χ^2 法を用いた重要漢字の自動抽出と文献の自動分類，技術研究報告 NLC94-25，電子情報通信学会 (1994).

[Yam95] 山本和英，増山繁，内藤昭三：文章内構造を複合的に利用した論説文要約システム *GREEN*，自然言語処理，Vol. 2, No. 1, pp. 39–55 (1995).

[Yua93] 湯浅夏樹，上田徹，外川文雄：大量の文書データから自動抽出した名詞間共起関係による文書の自動分類，研究会資料 NL98-11，情報処理学会 (1993).