# 段落内共起情報を利用した文書自動分類方式

藤井洋一[†] 鈴木克志[†] 辻 秀一[††]

従来，文書の自動分類の精度向上において問題となっていた多義語の曖昧性を解消し，分類精度を向上するために，以下の 3 つの点を特徴とする文書自動分類方式を提案する．1. 単語出現頻度と文書分類カテゴリとの間の $\chi^2$ 統計による重み付けにおいて複数の分類カテゴリで重要度が高い単語を分類多義語と定義する．2. 重要度が高い分類カテゴリを分類多義語に付け加えた単語を分類多義拡張単語とし，出現頻度を補正する．3. 分類多義語と同一段落内で共起する単語のみから構成する共起単語の共起ベクトルと入力文書の共起単語の共起ベクトルとの類似度を計算することで，入力文書の分類多義語の出現頻度を分類多義拡張単語の出現頻度に補正する．新聞記事 65,078 記事における詳細な 734 分類カテゴリへの分類実験の結果，分類精度が従来方式による場合の 48.1%から 3.5 ポイント改善された 51.6%になった．また，再現率と適合率単独で見ると，再現率が従来方式による場合の 45.5%から最大で 7.4 ポイント改善された 52.9%になり，適合率が 43.5%から 7.1 ポイント改善された 50.6%になった．

# An Automatic Document Classification Method Using Lexical Co-occurrences in the Paragraph

Youichi Fujii,[†] Katsushi Suzuki[†] and Hidekazu Tsuji[††]

This paper describes an automatic document classification using lexical co-occurrences. In the conventional documents classification methods, there is a problem of miss-classification caused by ambiguous words. In order to resolve the ambiguity depending on the class categories, we use the lexical co-occurrences with the ambiguous word. The method has three main ideas as follows: 1. We define 'the ambiguous word in classification' as the word that characterizes in plural classes in $\chi^2$ method between the word frequency and document classes. 2. We define 'the expanded ambiguous word in classification' as the word that joins 'the ambiguous word in classification' to calacterized class. 3. We adjust a word frequency for the ambiguity in classification using the words co-occurred with the ambiguous word in the same paragraphs. We classify 65,078 newspaper articles into 734 classes in the experiment. The result shows the performance improvement of 3.5 points from 48.1% of the conventional method to 51.6%. And the best result shows the recall improvement of 7.4 points from 45.5% to 52.9%, and the precision improvement of 7.1 points from 43.5% to 50.6%.

## 1. はじめに

近年，ワープロによる文書作成が当たり前になるとともに，WWW のホームページのようにネットワーク上での情報交換を目的として大量の電子化された文書が作成されている．これら文書を管理するため，従来から文書管理の重要性が指摘され，様々な文書管理の方法が試みられてきた．従来の文書管理では，文書を組織などの明確な軸に基づいて格納し，文書作成者や管理者が付与したキーワードと書誌情報で管理する方法がとられ[1]，利用者は，そのキーワードと書誌情報を使って所望の文書を検索していた．

そこで，管理者の手を煩わすことなく文書管理を行い，かつ有効な検索結果を得たいという要求があり，文書データベースを構築するための技術として，重要キーワードの自動抽出[2]，文書の内容での自動分類[3)~9),15)]，文書の検索結果を利用するための技術として，検索結果のランキング[10]，文書の抄録の作成[11)]などの研究が行われてきた．

このうち，文書自動分類の方式としては，既存の分類に新たな文書を割り当てる方式[3)~7),15)] と，分類を持たない文書集合から分類構造を生成するクラスタリングと呼ばれる方式[8),9)] がある．既存の分類に新たな文書を割り当てる方式としては，知識ベースを用いた

† 三菱電機株式会社情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation
†† 東海大学工学部電子工学科
Department of Electronics, School of Engineering, Tokai University

分類木による方式[3]と統計的な情報を用いる方式[4)~7)]がある．さらに，一般には単語を単位として処理するのに対して，英語では句（phrase）を用いた試み[15]もある．また，クラスタリングの方式でも統計的な情報を用いる方式[8]がある．知識ベースを用いる方式は細かい設定が可能であり，分野限定すれば分類精度を向上できる．しかし，定期的に知識ベースをメンテナンスする必要があるとともに，分野が異なる場合には，新たに知識ベースを構築しなおす必要がある．一方，統計的な方式を用いる場合は，分類先が既知の文書データが存在すれば分類の特徴を自動的に学習できるため柔軟なシステム構築が可能であるが，分類精度が低いという問題がある．統計的な情報を用いる方式にはベクトル空間モデル[13]を使ったもの[4)~8)]と，確率モデル[9),15)]を使ったものがある．以下では，ベクトル空間モデルを使った文書の自動分類方式に関して述べる．

従来のベクトル空間モデルによる自動分類方式は，tf·idf（term frequency times inverse document frequency）による方式[14]と $\chi^2$ 統計を応用した方式[4]があげられる．tf·idfは，文書中に現れた単語の頻度をtf，単語がどれくらい特定の文書で現れるかをidfで表現している．一方，$\chi^2$ 統計を応用した方式では，各分類中に現れるであろう理論頻度と実際の出現頻度との差から，単語の各分類への重みを表現する．ともに，高頻度で，特定の文書にのみ特徴的に現れる単語が分類に有効であるという考え方である．

これら，単語を利用した自動分類方式に対して分類精度を向上させるための様々な取組みがある．たとえば，単語の頻度情報を単純に利用した上記の方式に対して，シソーラスの意味属性を併用することによって分類精度を向上させる試み[4]がある．この方式は，同じ意味属性を持つ単語を同一視して頻度集計することによって分類精度を向上させようとするものである．シソーラスの意味属性と併用する方式に関しては，シソーラスの分解能に大きく左右されるため，シソーラスを的確に作成する必要がある．さらに，単語を対応する意味属性に割り当てるための高度な言語処理を必要としている．

また，単語の共起情報を用いて単語共起ベクトルを生成し，その単語共起ベクトルから文書ベクトルを生成することで自動分類し精度を向上させる試み[6]や，オントロジーと融合させて分類精度を向上させる試み[7]などがある．文献6)の単語共起を用いた方式では，分類カテゴリに基づかず共起情報を収集しており，分類が詳細になり単語の分類に対する特徴が小さくなった場合には精度が下がることが考えられる．

さらに，クラスタリングにおいて辞書の語義文を利用して多義解消したうえで，異なる単語の同じ意味の語義を同一視することでクラスタリング精度を向上させる試み[8]がある．この方式は，辞書の語義文との分類対象の文書とで多義解消を行っているため辞書の内容に大きく依存してしまう．また，データスパースネスにより処理が有効に働かなかったり，類似度が高い異なる記事が誤ってクラスタリングされる可能性が考えられる．

また，ベクトル空間モデルによる重み付けの方法を出現頻度の高い単語との比率で定義し，新聞記事の特徴を考慮してより前に出現した単語に重い重みを与える試み[12]がある．

以上のように，様々な自動分類の試みが行われているが，上記の提案方式によるベクトル空間モデルでの評価は分類カテゴリの数が10程度の場合が多く，自動分類の応用システムを考えた場合，十分な分類カテゴリ数とはいえない．また，分類カテゴリ数が多くなると，分類精度が低下することは想像されるが，それがどの程度であるかも明らかになっていない．

そのため本論文では，詳細な分類カテゴリでも自動分類できることを目指して，複数の分類カテゴリで高頻度に出現する単語に着目する．本方式は，文書内でその単語が出現した状況を段落内の共起情報を利用して自動学習し，分類精度を向上させる方式である．まず，2章で従来の方式について説明し，3章で本提案の方式について説明する．4章では，新聞記事を対象として分類実験を行い，評価した．5章で誤分類する分類カテゴリの分析を行い，課題について考察した．

## 2. 従来の自動分類方式

従来の自動分類方式として $\chi^2$ 統計を応用した文献4)の自動分類方式を説明する．ただし，以下では，文献4)の自動分類方式のうち意味属性を使わない単語のみでの自動分類方式を従来方式と記す．

**図1**は，従来方式の図である．処理はすでに分類先が分かっている文書を使って分類カテゴリ特徴ベクトルを計算する学習フェーズと，分類先が未知である文書に対して文書ベクトルを作成し，分類カテゴリ特徴ベクトルと文書ベクトルとの類似度によって分類先を決定する自動分類フェーズの2つに分けられる．

学習フェーズでは，すべての学習文書に対して形態素解析などにより単語を抽出し，文書ベクトルを作成する．さらに，各学習文書の正解分類先を利用して各分類カテゴリごとに単語の出現頻度を集計したあとで，

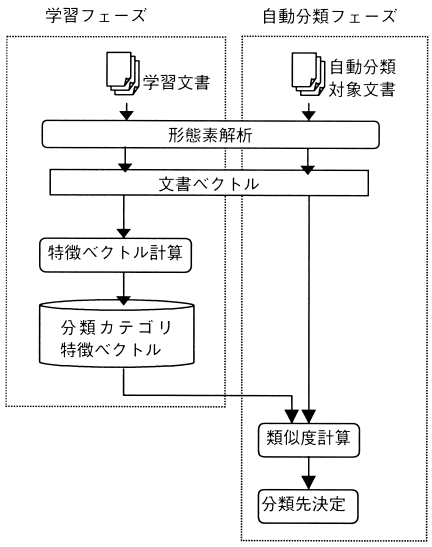


**図 1** 従来の自動分類方式
Fig. 1 Conventional automatic classification method.

分類カテゴリの特徴ベクトルを作成する．

まず，分類カテゴリを $C_i$ $(i = 1, 2, \cdots, N)$，単語を $w_j$ $(j = 1, 2, \cdots, L)$ としたとき，分類カテゴリ $C_i$ での単語 $w_j$ の頻度を $F_{ij}$ とする．次に，分類カテゴリ $C_i$ での単語 $w_j$ の理論頻度 $M_{ij}$ を式 (1) によって計算する．

$$M_{ij} = \sum_{i=1}^{N} F_{ij} \cdot \sum_{j=1}^{L} F_{ij} \Bigg/ \sum_{j=1}^{L} \left( \sum_{i=1}^{N} F_{ij} \right). \quad (1)$$

最後に，分類カテゴリ $C_i$ の特徴ベクトル $\mathbf{Y_i} = (Y_{i_1}, Y_{i_2}, \cdots, Y_{iL})$ の各要素 $Y_{ij}$ を式 (2) とする．

$$Y_{ij} = (F_{ij} - M_{ij}) \cdot |F_{ij} - M_{ij}| / M_{ij}. \quad (2)$$

一方，自動分類フェーズでは，自動分類対象文書から形態素解析などにより同様に単語を抽出し，自動分類対象文書中の単語 $w_j$ の出現頻度 $d_j$ から，文書ベクトル $\mathbf{D} = (d_1, d_2, \cdots, d_L)$ を求める．分類カテゴリ特徴ベクトル $\mathbf{Y_i}$ に対して文書ベクトル $\mathbf{D}$ との類似度 $s_i$ を式 (3) で計算する．

$$s_i = \sum_{j=1}^{L} (Y_{ij} \cdot d_j). \quad (3)$$

最後に，$s_i$ を正規化する $S_i$ を式 (4) で求め，$S_i$ の値が大きい分類カテゴリを分類先とする．

$$S_i = s_i \Bigg/ \left( \sum_{i=1}^{N} s_i \right) \quad (4)$$

一方，文献 6) の方式では，あらかじめ特徴ベクトルの基準となる単語を $n$ 個選定し，記事 $i$ に出現した単語 $w_j$ の出現頻度 $v_{ij}$ をもとに出現頻度ベクトル $V_i = (v_{i_1}, v_{i_2}, \cdots, v_{in})$ を計算する．次に単語 $w_i$ の特徴ベクトルを $W_i = \sum_{j=1}^{m} v_{ji} \cdot V_j / |V_j|$ と定義する（$m$ は学習記事数）．さらに，記事の特徴ベクトルを $A_i = \sum_{j=1}^{n} \log(m/m_j) \cdot v_{ij} \cdot W_j / |W_j|$ と定義する．分類カテゴリ $i$ の特徴ベクトル $C_i$ は分類カテゴリ $i$ を正解とする記事の特徴ベクトルの平均とする．自動分類対象の記事 $u$ に対しても出現頻度ベクトル $V_u$ をもとに同様に記事の特徴ベクトル $A_u$ を求め，$A_u \cdot C_i / |A_u| \cdot |C_i|$ が大きい分類 $i$ を分類先とする．

## 3. 分類多義語に基づく自動分類方式

今回，自動分類の方式として 2 章で述べた文献 4) の方式をもとに新たな方式を提案する．

従来の方式としてはほかにも tf·idf を利用した方式がある．tf·idf と $\chi^2$ 統計を応用した方式の違いは，tf·idf が関連しない分類先に対するベクトルの値が 0 であるのに対して，$\chi^2$ 統計を応用した方式では，負の値が与えられる点にある．我々は，関連しない分類先に対して関連しないということを表現できる $\chi^2$ 統計を応用した方式を利用することで精度良く分類できると考え，$\chi^2$ 統計を応用した方式を採用する．

提案方式は，以下で定義する分類多義語を用いて自動分類を行うものである．

### 3.1 分類多義語とは

本方式では，単語出現頻度と分類カテゴリとの間の従来の $\chi^2$ 統計による重み付け結果に対して複数の分類カテゴリで重要度が高い単語を分類多義語と定義する．

たとえば，単語「大統領」が分類カテゴリ〈政治〉にのみ多く現れたとすると，「大統領」は，分類カテゴリ〈政治〉へ分類するための有効な単語となり，分類多義語ではない．ところが，分類カテゴリ〈首相〉や〈大統領選挙〉，〈地方行政一般〉，〈外交関係〉といった複数の分類カテゴリで単語「大統領」が頻繁に出現したとすると，単語「大統領」はどの分類カテゴリに対しても重要度が高く，分類多義語である．

我々は，分類多義語が分類カテゴリによって異なる単語とともに使用される可能性が高いと考えた．たとえば，単語「大統領」は，分類カテゴリ〈首相〉では，「最高責任者」，「発言」といった単語とともに現れ，分類カテゴリ〈大統領選挙〉では，「選挙」，「当選」といった単語とともに現れ，分類カテゴリ〈外交関係〉では，「来日」，「対談」といった単語とともに現れる可能性が高くなる．

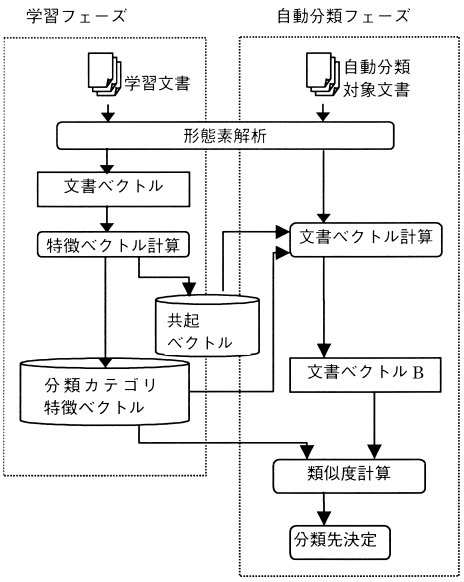


**図 2** 本提案の自動分類方式
Fig. 2 This automatic classification method.

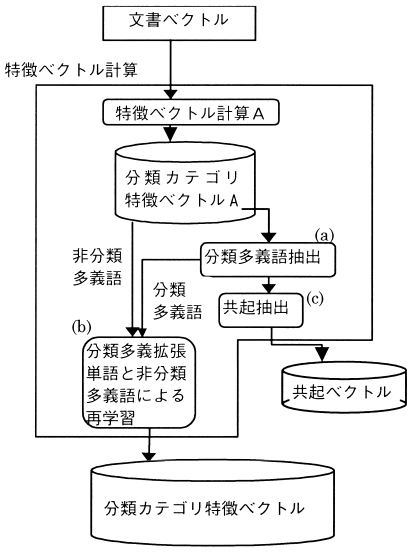


**図 3** 特徴ベクトル計算の流れ
Fig. 3 Flow of calculating characterized vectors.

### 3.2 分類多義語による自動分類方式

分類多義語による自動分類方式では，分類多義語に重要度が高い分類カテゴリをつけたものをあたかも単語であるかのように扱う．すなわち，分類カテゴリ〈首相〉,〈大統領選挙〉,〈地方行政一般〉，および〈外交関係〉で頻繁に現れる単語「大統領」に対しては「大統領 〈首相〉」と「大統領 〈大統領選挙〉」,「大統領 〈地方行政一般〉」,「大統領 〈外交関係〉」という拡張単語（以下，分類多義拡張単語と呼ぶ）を定義し，これをベクトル空間モデルでの類似度計算の基底となる単語とする．

自動分類対象である入力文書中の分類多義語の多義性を解消して分類多義拡張単語に変換し，分類多義拡張単語を頻度計算の単位とすれば，分類多義語を 1 単語のまま扱うよりも分類精度が向上すると考えられる．

**図 2** は，本方式の図である．学習フェーズでは，分類多義語拡張単語によって類似度計算を行うために，特徴ベクトル計算の中で分類多義語を抽出する．さらに，分類対象文書の分類多義語の頻度を分類多義拡張単語の頻度に分配するための共起ベクトルを作成する．一方，自動分類フェーズでは，分類対象文書の分類多義語の出現頻度を，分類多義拡張単語の出現頻度に分配し，文書ベクトル B を作成する．最後に，分類カテゴリ特徴ベクトルと文書ベクトル B によって類似度を計算して，従来方式と同じく分類先を決定する．

以下では，特徴ベクトル計算の方法，および文書ベクトルの計算方法について述べる．

#### 3.2.1 特徴ベクトル計算の方法

本方式での特徴ベクトル計算の方法を**図 3** に従って説明する．

ここで，特徴ベクトル計算 A は，式 (1)，(2) による従来の特徴ベクトル計算に対応する．さらに，分類カテゴリ特徴ベクトル A も，式 (1)，(2) により計算した従来の分類カテゴリ特徴ベクトルに対応する．

(a) 分類カテゴリ特徴ベクトル A から分類多義語を抽出する．

(b) 分類多義語を分割した分類多義拡張単語と分類多義語とならなかった非分類多義語を使って，分類カテゴリ特徴ベクトルを再計算する．

(c) 各学習文書中に出現する分類多義語に対して，分類多義語と段落内共起する単語を抽出し，抽出された段落内共起する単語の頻度を加算したものを共起頻度ベクトルとし，重み付け変換して共起ベクトルとする．

図 3 の (a) では，まず，各単語 $w_j$ に対して重要分類カテゴリ集合 $T_j$ を式 (5) とおく．

$$T_j = \{C_i | \max_{1 \le k \le N} (Y_{kj}) \cdot V_1 \le Y_{ij} (i = 1, 2, \cdots, N)\} \tag{5}$$

式 (5) の閾値 $V_1$ は，記事 1 年分の分類結果に対して値を変化させ，最も良い分類精度を示した値とした．$V_1$ として，0.0 とすると理論頻度よりも多く現れたものを単純に集めることなり，統計的な誤差によるノイズが大きくなる可能性が高く，逆に 0.9 のように値を大きくすると分類多義拡張単語の数が少なくなり本方

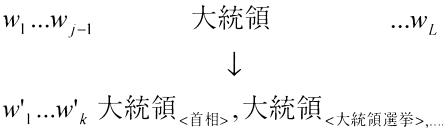


**図 4** 分類多義語の添え字対応づけ
Fig. 4 Mapping of 'the ambiguous word in classification'.

式の効果が薄くなると考えられる．そして，重要分類カテゴリ数 $|T_j|$（$|X|$ は集合の要素数）が 2 以上の単語 $w_j$ を分類多義語とする．

(b) では，分類多義語を分類多義拡張単語に分割する．分類多義拡張単語は，重要分類カテゴリ数 $|T_j|$ が 2 以上の単語 $w_j$ に，式 (5) を満足する分類カテゴリ $C_i$ を連結した形で記述する．したがって，分類多義語 $w_j$ に対して，$|T_j|$ 個の分類多義拡張単語ができる．そこで，各単語を**図 4** のように置き換える．すなわち各単語に対して，非分類多義語の場合はそのままの単語とし，分類多義語の場合はその分類多義語拡張単語で置き換えたものに対して番号を順番に振り直し，それを $w'_k$ $(k = 1, 2, \cdots, L')$ とおく．

図 4 は，大統領が分類多義語と判断された場合の添え字の対応づけの例を示している．

次に，分類多義拡張単語と非分類多義語による出現頻度 $F'_{ik}$ を求める．まず，分類多義拡張単語に対しては対応する分類カテゴリを返す関数 (6) を定義する．

さらに，$F'_{ik}$ を単語 $w_j$ の分類カテゴリ $C_i$ での出現頻度 $F_{ij}$ から計算する．$w'_k$ が $w_j$ の分類多義拡張単語の場合は式 (7) で計算する．

$$\varphi(k) = w'_k \text{を作成するときに付けた分類カテゴリ.} \tag{6}$$

$$F'_{ik} = \begin{cases} F_{ij} & (\varphi(k) = C_i) \\ F_{ij}\,/|T_j| & (\varphi(k) \neq C_i \text{ かつ } \varphi(k) \notin T_j) \\ 0 & (\varphi(k) \neq C_i \text{ かつ } \varphi(k) \in T_j) \end{cases} \tag{7}$$

すなわち，$w'_k$ のカテゴリが $C_i$ の場合は $F'_{ik}$ に $F_{ij}$ の頻度をそのまま割り当て，$w_j$ から拡張されたそれ以外の分類多義語への $C_i$ への頻度を 0 とする．一方，$w'_k$ の重要分類カテゴリに含まれないカテゴリに対しては頻度を等分する．

さらに，$w'_k$ が非分類多義語で $w_j$ に等しい場合は，式 (8) で計算する．

$$F'_{ik} = F_{ij} \tag{8}$$

**図 5** は分類多義拡張単語の頻度計算の方法を具体的に示したもので，各分類カテゴリに対して対応する分類カテゴリが付いた分類多義拡張単語がある場合には，対応する分類多義拡張単語にすべての頻度を割り当てる．一方，対応する分類カテゴリが付いた分類多義拡張単語がない場合には，すべての分類多義拡張単語に均等に頻度を割り当てる．

分類カテゴリ頻度（分類多義語）

| 分類項目＼単語 | 大統領 |
|---|---|
| <首相> | 120 |
| <大統領選挙> | 50 |
| <裁判> | 5 |
| <地方行政一般> | 70 |
| <外交関係> | 90 |
| : | 4 |

<首相>は「大統領<首相>」が分類多義拡張単語なので頻度 120 をそのまま「大統領<首相>」に与える．<裁判>は大統領の重要分類カテゴリでないので，値を 4 つの分類多義拡張単語に割り振り，1.25 ずつ与える．

分類多義語による分類カテゴリ頻度

| 分類項目＼分類多義拡張単語 | 大統領<首相> | 大統領<大統領選挙> | 大統領<地方行政一般> | 大統領<外交関係> |
|---|---|---|---|---|
| <首相> | 120 | 0 | 0 | 0 |
| <大統領選挙> | 0 | 50 | 0 | 0 |
| <裁判> | 1.25 | 1.25 | 1.25 | 1.25 |
| <地方行政一般> | 0 | 0 | 70 | 0 |
| <外交関係> | 0 | 0 | 0 | 90 |
| : | 1 | 1 | 1 | 1 |

**図 5** 分類多義語の頻度分配例（学習フェーズ）
Fig. 5 Example of the frequency division of 'the ambiguous word in classification' (Leaning phase).

最後に，$F'_{ik}$ に対する，理論頻度を計算し，分類カテゴリ特徴ベクトルを計算する．

次に，共起ベクトルの作成方法について述べる．

(c) では，学習記事を $D_l$ $(l = 1, 2, \cdots, P)$ とし，学習記事 $D_l$ の文書ベクトルを $\mathbf{DL}_l$ とおく．また，分類カテゴリ $C_i$ が，文書 $D_l$ の正解分類先に含まれるかどうかで 0，1 を返す関数 (9) を定義する．

次に，学習記事 $D_l$ における単語 $w_j$ と同一段落内に出現する単語の頻度を集計した段落内共起ベクトルを $\mathbf{O}_{lj}$ とおき，分類多義語 $w'_k$ の共起頻度ベクトル $\mathbf{O}'_k$ を式 (10) で計算する．

$$\delta(C_i, D_l) = \begin{cases} 0 & (C_i \text{が } D_l \text{の分類先でない}) \\ 1 & (C_i \text{が } D_l \text{の分類先}) \end{cases} \tag{9}$$

$$\mathbf{O}'_k = \sum_{l=1}^{p} \mathbf{O}_{l\phi(k)} \cdot \delta(\varphi(k), D_l)$$

$$(\phi(k) \text{ は } w'_k \text{ の分類多義拡張前の単語 } w_j \text{ の } j) \tag{10}$$

これにより，各学習記事 $D_l$ の正解分類カテゴリに対してのみ，段落内共起する単語の出現頻度を計算したことになる．

共起頻度ベクトル $\mathbf{O}'_k$ に対して，次の (i)，(ii) の変換を行って共起ベクトル $\mathbf{O}^*_k$ とする．

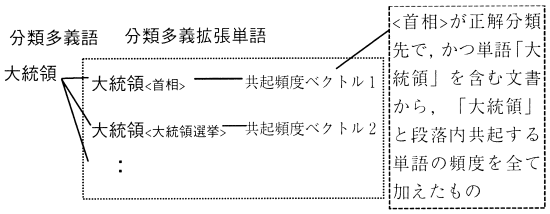


**図 6** 共起頻度ベクトル計算の例
Fig. 6 Example of calculating co-occurence vector.

(i) 分類多義拡張単語ごとに長さ 1 に正規化する．$\mathbf{O}'_k = (o_{1k}, o_{2k}, \cdots, o_{Lk})$ とおき，式 (11) を計算する．

(ii) ベクトルを共起単語が出現した分類数で割る．$\mathbf{O}''_k = (o'_{1k}, o'_{2k}, \cdots, o'_{Lk})$ とおく．また，出現分類多義数 $n(i,k)$ を式 (12) と定義する．

式 (12) では，各分類多義語での共起する単語の個数が求まる．

また，各単語の共起重みを式 (13) で定義する．

分類多義語拡張単語 $w'_k$ の共起ベクトル $\mathbf{O}^*_k$ を式 (14) とおく．

$$\mathbf{O}''_k = \mathbf{O}'_k \Big/ \sqrt{\sum_{j=1}^{L} (o_{jk}^2)} \tag{11}$$

$$n(i,k) = |\{o'_{im} | o'_{im} \neq 0 (m \in T_{\phi(k)})\}| \tag{12}$$

$$\mu(i,k) = \begin{cases} o'_{ik}/n(i,k) & (\text{if } n(i,k) \neq 0) \\ 0 & (\text{if } n(i,k) = 0) \end{cases} \tag{13}$$

$$\mathbf{O}^*_k = (\mu(1,k), \mu(2,k), \cdots, \mu(L,k)) \tag{14}$$

(i) は頻度を学習量に依存せず比較する目的で行い，(ii) は重要度が高い分類カテゴリで，より多く出現した共起単語が分類多義語解消に影響する力を小さくする目的で行う．

**図 6** は，共起頻度ベクトル作成の具体的イメージを示したもので，抽出された各分類多義語に対して，それぞれの分類多義拡張単語に付いている分類カテゴリと正解分類先が一致し，かつ分類多義語を含む文書すべてから，分類多義語と段落内共起する単語の頻度合計値を値とする共起頻度ベクトルである．

### 3.2.2 文書ベクトル計算の方法

本方式での文書ベクトル計算の方法を**図 7** に従って説明する．

(A) 分類多義語に対しては分類対象文書中の分類多義語と段落内共起するすべての単語の頻度を抽出し，分類多義語の文書共起ベクトルとする．

(B) 各分類多義語に対して学習結果の共起ベクトルと，分類対象文書の文書共起ベクトルとから分類多義の多義性を解消して，分類多義拡張単語の頻度に分配する．

(C) 分類多義拡張単語と非分類多義語の頻度から文書ベクトルを作成する．

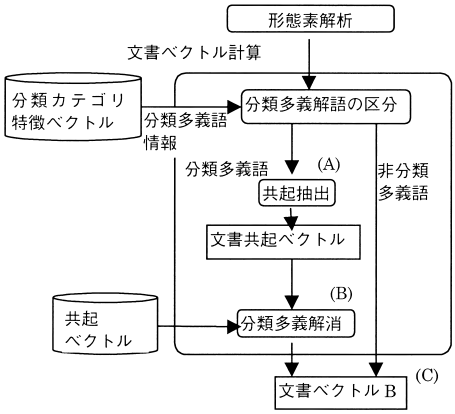


**図 7** 文章ベクトル計算の流れ
Fig. 7 Flow of calculating document vectors.

図 7 の (A) では，分類対象文書の文書ベクトルを $\mathbf{D} = (d_1, d_2, \cdots, d_L)$ とおく．また，分類対象文書中の単語 $w_j$ と段落内共起する単語の文書共起頻度ベクトルを $\mathbf{OD}_j$ とする．

(B) では，文書共起頻度ベクトル $\mathbf{OD}_j$ と共起ベクトル $\mathbf{O}^*_k$ との内積を計算することで，$w_j$ の頻度 $d_j$ から $w'_k$ の頻度 $d'_k$ を決定する．まず，頻度分配の閾値 $b_j$ を式 (15) で求める．

ここで $V_2$ という値を掛けたのは，単純に平均以上かどうかで閾値 $b_j$ を設定せず，統計的な誤差を吸収できるようにするためである．$V_2$ は実験によって決定する．

次に，各分類多義拡張単語 $w'_k$ の頻度 $d'_k$ を式 (16) で計算する．

$$b_j = \left( \sum_{k \in \{k | \phi(k) = j\}} \mathbf{OD}_j \cdot \mathbf{O}^*_k \right) \cdot V_2 / |T_j| \tag{15}$$

$$d'_k = \begin{cases} 0 & (\text{if } \mathbf{OD}_j \cdot \mathbf{O}^*_k < b_j) \\ d_j \cdot \dfrac{\mathbf{OD}_j \cdot \mathbf{O}^*_k}{\sum_{k \in \Psi(j)} \mathbf{OD}_j \cdot \mathbf{O}^*_k} & \\ & (\text{if } \mathbf{OD}_j \cdot \mathbf{O}^*_k < b_j) \\ & (\Psi(j) = \{k | \mathbf{OD}_j \cdot \mathbf{O}^*_k \geq b_j, \phi(k) = j\}) \end{cases} \tag{16}$$

式 (16) は，共起ベクトルとの内積値が閾値 $b_j$ 以下の分類多義拡張単語には頻度を割り当てず，閾値以上の分類多義拡張単語に対して内積値に比例して頻度を配分することを示している．

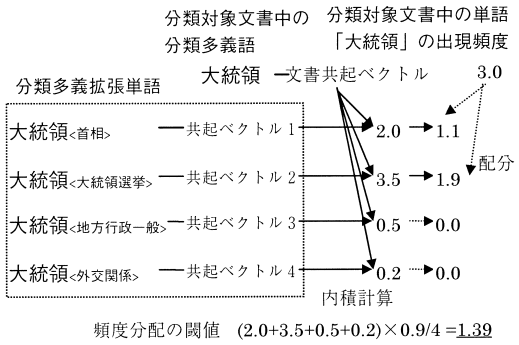


**図 8** 分類多義語の頻度分配例（自動分類フェーズ）
Fig. 8 Example of the frequency division of 'the ambiguous word in ckassification' (Classification phase).

**図 8** は分類多義語の多義解消を具体的に示したもので，文書共起ベクトルと，分類多義語の共起ベクトルとの内積を計算し，分類対象文書中の分類多義語の頻度を一定値以上の内積値を持つ分類多義拡張単語の頻度に分配する．

(C) では，式 (16) の計算結果から，分類多義拡張単語をベクトルの基底とする文書ベクトル B $\mathbf{D}' = (d'_1, d'_2, \cdots, d'_{L'})$ を求める．

なお，提案方式の特徴は以下の点にある．文献 4) の方式での意味属性の扱いは，異なる単語を同一視する処理方式であり，提案方式は 1 つの単語を使用された状況で分割することが特徴である．また，共起情報を用いる従来方式[6] では，正解分類先に関係なく共起を抽出するのに対して，本方式では正解分類先を基により詳細な共起情報を抽出する方式をとっている．さらに，多義解消の従来方式[8] では，多義語を辞書の語義文との共起で解消するのに対して，本方式では正解分類先ごとの共起情報で解消する．

## 4. 新聞記事を使った実験評価

今回，新聞記事（朝日新聞記事 1 年分：朝日新聞社提供）[16] を対象として分類実験を行った．新聞記事の特徴としては，内容別に分類されている，階層的に分類されている，かつ用語が比較的統一されており頻度による情報が有効に働く，といったことがあげられる．したがって，分類が細かくなった場合の分類精度を段階的に評価できるとともに，表記のゆれによる影響を少なくして純粋な精度を評価できると判断した．実際の文書管理システムを構築する場合には，異表記や同義語の辞書を用意することで，専門分野などの文書管理も可能となる．

形態素解析には JUMAN 2.0[17]+EDR 日本語単語辞書 1.5 版[18] を利用した．

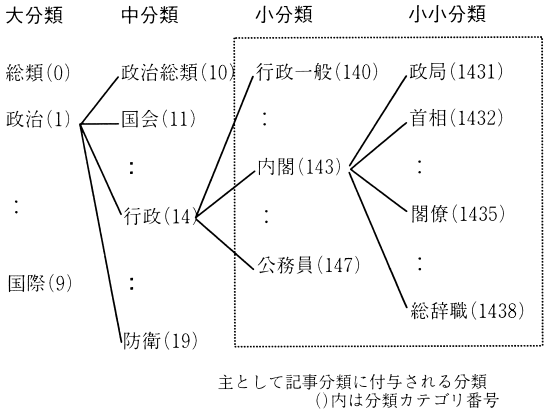


**図 9** 新聞記事分類
Fig. 9 Classes of news articles.

### 4.1 実験内容

今回の実験に用いた新聞記事の「主題分類」は**図 9** に示すように 4 階層の分類を持っている．大分類，中分類，小分類，および小小分類のうち，各記事には主として小分類と小小分類が複数個付与されており，これを正解として自動分類の実験と評価を行った．データはランダムサンプリングにより学習セットと評価セットに分けた．

従来方式による自動分類の実験のため，(a) 大分類，(b) 中分類，(c) 小分類と小小分類（以下，小分類と呼ぶ）の 3 種類の分類を考えた．1 記事に付与された分類カテゴリは，(c) の階層で複数個付与されており，複数個付与されている記事は，複数の分類カテゴリの内容が含まれていると考え，学習時の 1 記事中の頻度を**表 1** のように分配した．たとえば，〈図書〉，〈読書〉，および〈日本文学〉という正解分類カテゴリが付与されている記事を考える．(c) の分類では，1 記事中の頻度を正解分類カテゴリ数 3 で割った頻度を各分類カテゴリで学習した．一方，(a) の分類では，上位の大分類カテゴリ〈総類〉と〈文化〉を正解分類カテゴリとし，学習時には，(c) の割合をそのまま加えた大分類カテゴリでの「頻度にかける割合」が出現したものとして学習した．(b) の分類でも (a) と同様の処理を行った．

実験は以下に示す 2 つを行った．

［**実験 1**］従来方式による自動分類の精度を確認するため，単語のみの自動分類を，200，1,000，および 10,000 記事を利用して大分類，中分類，および小分類の 3 種類で行った．実験 1 では，従来方式の実験結果をも検証するため，記事サイズを従来方式と同様の条件とするため 150 文字から 500 文字の記事を古い順番に取り出し，学習記事と実験記事がそれぞれ 75%と

**表 1** 学習時の正解分類先対応付け方法の例
Table 1 Example of weight dividing at learning.

| 小分類項目 番号 | 項目名 | 頻度にかける割合 | 大分類項目 番号 | 項目名 | 頻度にかける割合 |
|---|---|---|---|---|---|
| 052 | 図書 | 0.33 | 0 | 総類 | 0.66 |
| 054 | 読書 | 0.33 | | | |
| 412 | 日本文学 | 0.33 | 4 | 文化 | 0.33 |

各分類項目の番号は分類コードを示す

**表 2** 閾値 $V_1$ による分類結果
Table 2 Classification results (parameter $V_1$).

| 閾値 $V_1$ | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 |
|---|---|---|---|---|---|
| 分類精度 | 48.1% | 51.4% | 48.8% | 48.4% | 48.3% |

25%となるようランダムサンプリングした．

［**実験 2**］提案方式の有効性を確認するため，新聞記事 1 年分を利用して，単語のみを利用した自動分類と共起情報を用いた自動分類を，大分類，中分類，および小分類で行った．学習した記事は 61,500 記事で，自動分類した記事は 3,578 記事である．データは実験 1 同様にランダムサンプリングにより学習記事と自動分類記事に分けた．記事サイズに制限はつけなかった．また，このとき，式 (5) の閾値 $V_1$ を決定するために新聞記事 1 年分を使って分類精度を評価した．閾値 $V_2$ は仮に 0.9 とした．結果は**表 2** のとおりである．この閾値 $V_1$ は，共起によって分類多義語を分割する単語を決定するもので，学習記事数に大きく依存しない値と考えられる．以後 $V_1 = 0.1$ とする．ただし，式 (5) の閾値 $V_1$ を小さくすると分類多義語数が多くなり，分類カテゴリ特徴ベクトルおよび，共起ベクトルのサイズが大きくなる．そのため，実際に分類多義語として選択したのは式 (17) の値が大きく，分類多義数が 10 以下の上位 10,000 語に制限した．

$$\sum_{i=1}^{N}(F_{ij} - M_{ij})^2 \Big/ M \qquad (17)$$

式 (17) は分類に重要な単語ほど値が大きくなるため，単語を制限することによって分類多義語の多義性が非常に多く，分類時にノイズとなる可能性が高い単語を省くことになる．

今回の実験で利用した単語は，名詞，固有名詞，未知語，およびサ変名詞であり，1 文字のみからなる単語と記号列のみの未知語は取り除いた．

さらに，文献 6) との比較のため，文献 6) の方式に基づき，新聞記事 1 年分を利用した分類実験を行った．今回は文献 6) で最良と述べられている単語選択方式として特定の分類カテゴリに出現する単語を 10,000 語選択した．選択方法は全出現回数に対して特定の分類カテゴリに出現した出現割合が高いものから 10,000 語とした．なお，文献 6) では，典型的な数記事を人手で選び分類カテゴリベクトルとしていたが，今回の新聞記事では，ほとんどの記事に複数の正解分類先が与えられており，人手での選択も困難である．そこで，本提案方式と同様に表 1 の方法で加算した結果の平均を分類カテゴリベクトルとした．

**表 3** 実験 1 の分類結果
Table 3 Classification results of Experiment 1.

| 分類 (分類項目数) | 記事数 200 | 1000 | 10000 |
|---|---|---|---|
| 大分類(10) | 48.4% | 63.4% | 68.8% |
| 中分類(92) | — | 50.0% | 58.5% |
| 小分類(734) | — | 37.8% | 48.1% |

**表 4** 分類多義語数
Table 4 Num. of 'the ambiguous word in classification'.

| | 10000 記事 | 1 年分 |
|---|---|---|
| 総単語数 | 30032 | 91575 |
| 大分類多義語数 | 13709 | 45461 |
| 小分類多義語数 | 24637 | 77460 |

### 4.2 実験結果と評価

記事には複数の正解分類先が付与されているので，すべての自動分類記事の各分類カテゴリへの評価値（式 (4) の $S_i$）の閾値を変化させて，閾値以上の分類先を自動分類結果とした．以下に示す分類精度は，適合率と再現率が等しくなるときの正解率（%）を示している．再現率と適合率はそれぞれ式 (18)，(19) で定義される．

$$\text{再現率} = \frac{\text{自動分類結果正解数}}{\text{記事に付与された分類数}} \cdot 100 \qquad (18)$$

$$\text{適合率} = \frac{\text{自動分類結果正解数}}{\text{総自動分類数}} \cdot 100 \qquad (19)$$

#### 4.2.1 実験結果

実験 1 の自動分類を行った結果を**表 3** に示す．

200 記事では従来方式の実験とほぼ同様の約 50%の精度を得ることができ，大分類に限れば記事の数を 10,000 記事にまで増やすことで約 70%まで精度向上した．しかし，いずれの場合も小分類では大分類よりも約 20 ポイント精度が悪く，分類が細かくなったときに問題が大きいことが分かる．

次に，提案方式で仮定した分類多義語が，分類が細かくなった場合に多くなっているかどうかを実際に集計した結果を**表 4** に示す．

**表 5** 実験 2 の分類結果
Table 5 Classification results of Experiment 2.

| | 文献 6) | 従来方式 | 本方式 | 従来方式との差 |
|---|---|---|---|---|
| 大分類(10) | 20.9% | 70.3% | 70.6% | 0.3% |
| 中分類(92) | — | 57.3% | 58.8% | 1.5% |
| 小分類(734) | 8.5% | 48.1% | 51.6% | 3.5% |

**表 6** 閾値 $V_2$ による分類結果
Table 6 Classification results (parameter $V_2$).

| 閾値 $V_2$ | 0.80 | 0.85 | 0.90 | 0. 95 | 1.0 |
|---|---|---|---|---|---|
| 分類精度 | 51.2% | 51.3% | 51.4% | 51.6% | 50.7% |

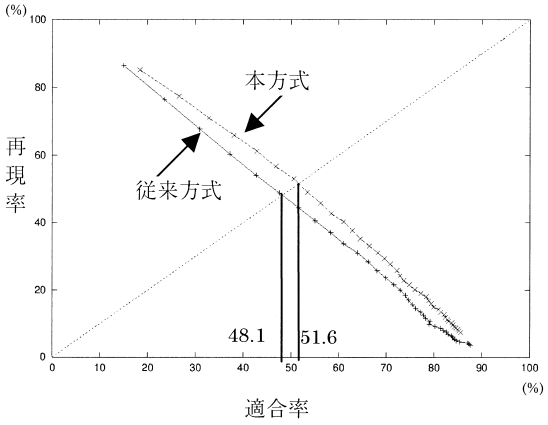


**図 10** 適合率–再現率曲線
Fig. 10 Recall-precision curve.

上記結果より，分類多義語は分類が細かくなった場合に 1.7 から 1.8 倍程度増えていることが確認された．このことから，分類精度を下げている原因の 1 つとして分類多義語と定義される単語の分類先の曖昧性が増加したことが考えられる．

次に実験 2 の分類結果を**表 5** に示す．

分類が中分類，小分類と詳細になるにつれて，精度向上の差が大きくなっていることからも，本方式の効果が結果として現れている．

表 5 においての分類精度は式 (15) の $V_2$ の値を 0.05 ずつ変化させて最も分類精度が良かった 0.95 の場合である．$V_2$ と分類精度の関係を**表 6** に示す．

さらに，実験 2 の小分類での適合率–再現率の曲線を**図 10** に示す．

従来方式と比較して，小分類で 3.5 ポイント向上している．これを再現率と適合率で単独に評価すると，それぞれ最大で 7.4%（適合率 50.6%時に再現率が 45.5%から 52.9%へ），7.1%（再現率 52.9%時に適合率が 43.5%から 50.6%へ）向上した．適合率–再現率曲線においても全体的に従来方式よりも良い分類精度であった．

**表 7** 分類多義解消結果
Table 7 Results of 'the ambiguous word in classification' disambiguation.

| | 正解分類多義語 | 不正解分類多義語 |
|---|---|---|
| 上位 1〜1000 | 0.73 | 0.14 |
| 1001〜2000 | 0.72 | 0.13 |
| 2001〜3000 | 0.71 | 0.12 |
| 3001〜4000 | 0.70 | 0.11 |
| 4001〜5000 | 0.68 | 0.12 |
| 5001〜6000 | 0.69 | 0.12 |
| 6001〜7000 | 0.69 | 0.12 |
| 7001〜8000 | 0.70 | 0.11 |
| 8001〜9000 | 0.70 | 0.12 |
| 9001〜10000 | 0.66 | 0.13 |

さらに文献 6) の方式と比較しても，表 5 に示したとおり，本提案方式による結果のほうが良い結果となっている．

また，予想していたように，分類精度の向上は小分類における方がより大きかった．表 4 の分類多義語の増加は小分類の方がより精度向上した裏付けと判断できる．

#### 4.2.2 分類多義語による精度向上

分類多義語を用いた本方式によって誤分類が減少し，従来の方式と比較して分類精度が向上した例がある．

〈大統領選挙〉という分類には,「大統領」という単語の影響で，従来方式では 156 記事が誤分類されていた．それに対して本方式は「大統領」を分類多義語として扱うことで，誤分類が 109 記事減った．ほかにも〈金融一般〉,〈五輪冬〉といった分類で誤分類が減少することで，全体として分類精度が向上している．

#### 4.2.3 分類多義語による多義解消の有効性評価

分類多義語の有効性を確認するため，分類多義語の多義解消結果を**表 7** に示す．表 7 は分類多義語を式 (17) の値が大きい順に 1,000 語ずつにまとめ，実際の分配頻度と，理想的な分配頻度の比を平均したものである．理想的分配は，分類対象記事の正解分類先を元に，各分類多義語の分配ごとに正解分類カテゴリがついた分類多義拡張単語（以下，正解分類多義拡張単語）に実際の出現頻度を等分し，それ以外の分類多義拡張単語（以下，不正解分類多義拡張単語）には頻度 0 を割り当てた場合とする．したがって，表 7 の値は，正解分類多義拡張単語に対しては 1.0，非正解分類多義拡張単語に対しては，0.0 が理想である．

正解分類多義拡張単語と，不正解分類多義拡張単語に対して，理想値まではいかないが，期待どおりに多義解消されていることが分かる．すなわち，不正解分類多義拡張単語の頻度がおさえこまれたことになる．

**表 8** 分類結果の傾向
Table 8 Tendency of classification results.

| | | 正解分類先数 | | |
|---|---|---|---|---|
| | | 増加 | 0 | 減少 |
| 誤分類数 | 増加 | 6.7* / 2.6 | 2.0 | 1.9 |
| | 0 | 1.7 | 0 | 1.4 |
| | 減少 | 5.7 | 5.5 | 1.0 / 12.9 |

*正解増加数＞不正解増加数

**表 9** 大分類での比較
Table 9 Comparation at high classes.

| 分類レベル | 従来方式 | 本方式 |
|---|---|---|
| 大分類で直接分類 | 70.3% | 70.6% |
| 小分類での結果を大分類に利用 | 73.3% | 74.7% |

さらに，**表 8** は正解，不正解の増減を分類カテゴリごとに計算し，同じ傾向の分類カテゴリをまとめ，変化した記事数を分類カテゴリ数で割ったものである．誤分類数が減少した分類カテゴリで変化した記事が多い特徴がある．

このことから，4.2.2 項の評価内容と同様，分類多義語の多義解消で，特定単語が高頻度で出現したため誤分類した記事が減少し，精度向上したと推定される．

#### 4.2.4 小分類を利用した大分類での精度向上

小分類での自動分類結果について大分類レベルでの評価を行った結果を**表 9** に示す．

小分類での分類結果を使って大分類での分類先を決定するすることで，大分類で直接学習した場合よりも従来方式で 3.0 ポイント，本方式で 4.1 ポイント分類精度が向上した．この原因としては大分類間の境界に位置する記事を小分類で分類した結果から推定することで，より細かい類似性が判定できたためと考えられる．

## 5. 考　　察

4 章の実験結果における誤分類の状況，および自動分類の意味属性の取扱いについて整理し，これら誤分類について今後の可能な対処方法について考察した．これらの対処方法を実現することでさらなる精度向上が可能と考えられる．

(1) 特徴量の弱い分類カテゴリの場合

今回利用した分類カテゴリでは，中分類での分類先は明らかであるが小分類に具体的に対応する分類カテゴリがない場合に分類するための〈○○一般〉という分類が用意されている．たとえば，図 9 では中分類〈行政〉の小分類として〈行政一般〉が用意されている．これらの分類カテゴリは他の分類カテゴリと比較して出現する単語の特徴が弱いため，従来方式でも本方式でも精度が低い．

**表 10** 誤分類例 1
Table 10 Example of miss-classification 1.

| | 再現率 | 適合率 |
|---|---|---|
| 従来方式 | 35.4% | 37.4% |
| 本方式 | 38.4% | 40.3% |

**表 11** 誤分類例 2
Table 11 Example of miss-classification 2.

| 分類項目 | ＜五輪一般＞ | | ＜五輪夏＞ | | ＜五輪冬＞ | |
|---|---|---|---|---|---|---|
| | 再現率 | 適合率 | 再現率 | 適合率 | 再現率 | 適合率 |
| 従来方式 | 20.0% | 50.0% | 97.7% | 16.9% | 100.0% | 41.1% |
| 本方式 | 80.0% | 22.7% | 97.7% | 53.6% | 100.0% | 60.0% |

**表 12** 誤分類例 3
Table 12 Example of miss-classification 3.

| 分類項目 | ＜日本文学＞ | | ＜郷土芸能＞ | | ＜回顧＞ | |
|---|---|---|---|---|---|---|
| | 再現率 | 適合率 | 再現率 | 適合率 | 再現率 | 適合率 |
| 従来方式 | 25.0% | 44.4% | 0.0% | 0.0% | 0.0% | 0.0% |
| 本方式 | 27.0% | 52.0% | 0.0% | 0.0% | 0.0% | 0.0% |

実際に，実験 2 の小分類の結果に対して〈○○一般〉という名前の分類カテゴリに限定して再現率と適合率を求めた結果は**表 10** のように再現率，適合率ともに全体の分類精度よりも 10 ポイント以上分類精度が下がっていることが分かる．

同様に，〈五輪一般〉，〈五輪夏〉，および〈五輪冬〉という分類を見ると，一般的な〈五輪一般〉を正解とする記事の分類精度が，より具体的な〈五輪夏〉や〈五輪冬〉と比較すると悪いことが**表 11** から分かる．

これらの分類カテゴリに対しては，分類の階層性を意識した学習と分類方式の検討が必要と考える．

(2) 分類カテゴリを示す単語が存在しない場合分類カテゴリの中には，記事を分類するための特徴的な単語が存在しないものがある．たとえば，分類カテゴリ〈日本文学〉や，〈郷土芸能〉，〈回顧〉などは，それぞれ記事の内容ではなく，著者が日本人であるとか，特定の地域での伝統芸能であるとか，戦争を回顧的に書いたといった観点で付与される．

実際に，分類カテゴリ〈日本文学〉，〈郷土芸能〉，および〈回顧〉の分類結果を**表 12** に示す．いずれも特徴的な単語が少なく統計的な手法では自動分類が困難であることが分かる．このようなものはほかにも，〈言論の自由〉，〈世相〉，〈話題〉など数多くの分類カテゴリがあげられる．

これらは，たとえば，〈日本文学〉であれば，著者

名を抽出して判断するといったように分類カテゴリの個々に依存した特別な処理方式を考える方法が考えられる．

(3) 唯一の分類カテゴリのみ特徴的に現れた単語の場合

本方式は，特定の分類カテゴリにのみ特徴的に現れる単語が処理対象とならないため，誤分類を修正することができない．たとえば分類カテゴリ〈写真〉は単語「写真」が分類に有効な単語となっている．しかし，逆に，〈写真〉に誤分類したほとんどの記事は単語「写真」が出現している．

これらの記事に対しては，文書構造の解析などにより重要な文に出現する単語の重みを変える処理方式の検討が必要と考える．具体的には文献 12) の単語の出現位置による重みの変更などが可能である．

(4) 意味属性の取扱い

誤分類の課題とは異なるが，本方式では，文献 4) の意味属性を扱った処理との比較をここでは行っていない．この意味属性を扱った方式は異なる単語を同一視して分類精度を向上させることが主であるのに対して，本手法は多義語を新たな単語と見なすことで分類精度を向上させようとしている．したがって，この方式と本提案方式を組み合わせて利用することにより，さらに精度向上することが期待される．

## 6. おわりに

本論文では，統計的な手法により詳細な分類に対して分類精度を向上させる方式を提案した．本方式は，複数の分類カテゴリで分類に有効と判断された単語を分類多義語と定義したのち，抽出した分類多義語が特徴的に現れた分類カテゴリを分類多義語に付与した分類多義拡張単語を生成し，分類多義語の頻度を段落内の共起情報を用いることで分類多義拡張単語の頻度に分配して，類似度計算することで分類先を決定するものである．

本手法を使った新聞記事の分類実験の結果，詳細な分類に対して本手法による自動分類の有効性と効果を評価することができた．

さらに本方式の今後の課題である誤分類の状況について整理し，可能な限り対処方法について考察した．

**謝辞** なお，本研究にあたり，朝日新聞社電子電波メディア局の関係者の方々に新聞記事の利用を了解いただいた．

## 参考文献

1) 伊藤哲郎：情報検索（ソフトウェア講座 19），昭晃堂 (1986).
2) 木本晴夫：日本語新聞記事からのキーワード自動抽出と重要度評価，電子情報通信学会論文誌（D-I），Vol.J74-D-I, No.8, pp.556–566 (1991).
3) 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.1103–1112 (1987).
4) 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，情報処理学会論文誌，Vol.33, No.9, pp.1112–1122 (1992).
5) 徳永健伸，岩山 真：重み付き IDF を用いた文書の自動分類について，情報処理学会自然言語処理研究会資料，NL-100-5, pp.33–40 (1994).
6) 湯浅夏樹，上田 徹，外川文雄：大量文書データ中の単語共起を利用した文書分類，情報処理学会論文誌，Vol.36, No.8, pp.1819–1827 (1995).
7) 岩爪道昭，白神謙吾，畑谷和右，武田英明，西田豊明：オントロジーに基づく広域ネットワークからの情報収集・分類・統合化，情報処理学会論文誌，Vol.38, No.3, pp.606–615 (1997).
8) 福本文代，鈴木良弥，福本淳一：辞書の語義文を用いた文書の自動分類，情報処理学会論文誌，Vol.37, No.10, pp.1789–1799 (1996).
9) Iwayama, M. and Tokunaga, T.: Hierarchical Bayesian Clustering for Automatic Text Classification, *Proc. IJCAI-95*, pp.1322–1327 (1995).
10) 野本晶子，野口直彦：文書構造と共起表現を用いた文書ランキング，情報処理学会第 52 回全国大会予稿集，5P-6, 第 4 分冊 pp.203–204 (1995).
11) 住田一男，知野哲朗，小野顕司，三池誠司：文書構造解析に基づく自動抄録生成と検索提示機能としての評価，電子情報通信学会論文誌（D-II），Vol.J78-D-II, No.3, pp.511–519 (1995).
12) 新谷 研，角田達彦，大石 巧，長尾 真：単語の共起情報と出現位置による新聞の関連記事の検索手法，情報処理学会論文誌，Vol.38, No.4, pp.855–862 (1997).
13) Salton, G.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).
14) Buckley, C., Allen, J. and Salton, G.: Automatic Routing and Retrieval Using SMART: TREC-2, *Information Processing & Management*, Vol.31, No.3, pp.315–326 (1995).
15) Lewis, D.D.: An Evaluation of Phrasal and Culusterd Representations on the Text Categorization Task, *15th Annual International ACM SIGIR Conference on Research and development in Information Retrieval*, pp.37–50 (1992).
16) 朝日新聞記事データベース（1991 年 9 月〜1992

年 8 月).
17) 松本裕治：日本語形態素解析システム JUMAN 使用説明書 version 2.0 (1994).
18) EDR 電子化辞書 日本語単語辞書 1.5 版,(株) 日本電子化辞書研究所 (1995).

(平成 12 年 2 月 29 日受付)
(平成 13 年 1 月 11 日採録)

**藤井　洋一**（正会員）

1963 年生．1986 年大阪大学理学部数学科卒業．同年三菱電機（株）入社．1990〜1992 年（株）日本電子化辞書研究所出向．自然言語処理，音声言語処理の研究開発に従事．現在，三菱電機（株）情報技術総合研究所音声・言語インタフェース技術部所属．日本音響学会会員．

**鈴木　克志**（正会員）

1956 年生．1979 年東京大学工学部計数工学科卒業．同年三菱電機（株）入社．自然言語処理，文書処理の研究開発に従事．現在，同社情報技術総合研究所音声・言語インタフェース技術部所属．

**辻　　秀一**（正会員）

1946 年生．1969 年大阪大学基礎工学部電気工学科卒業．1974 年大阪大学大学院基礎工学研究科博士課程修了．工学博士．同年三菱電機（株）入社．中央研究所，コンピュータシステム製作所，情報電子研究所，情報技術総合研究所にて，画像処理システム，マンマシンシステム，知識情報処理システム等の研究開発に従事．1997 年電子商取引実証推進協議会（ECOM）へ出向し電子商取引基盤の研究に従事．2000 年東海大学工学部電子工学科教授．現在に至る．電子情報通信学会，人工知能学会，IEEE 各会員．