# 大量文書データ中の単語間共起を利用した文書分類

湯浅夏樹† 上田徹† 外川文雄†

本稿では，特徴ベクトルを用いて自動的に文書分類を行う二つの手法を提案する．一つは，大量の文書データを用いて，同一記事中の単語間共起関係から分野の特徴を表す単語出現頻度分布の近似値を求め，この値を要素とする特徴ベクトルを用いて文書を分類する手法である．もう一つは，EDRの辞書をシソーラスとして用い，単語間の類似度を求め，この単語類似度を要素とする特徴ベクトルを用いて文書を分類する手法である．これらの手法を人手による分類と比較したところ，単語間共起を用いた手法では83.5%の記事が正しく分類され，易しい記事だけに限定すれば98.0%の記事が正しく分類されることが確認できた．また，シソーラスを用いた手法では，63.75%の記事が正しく分類されることが確認できた．

# Classifying Articles Using Lexical Co-occurrence in Large Document Databases

NATSUKI YUASA † TORU UEDA † and FUMIO TOGAWA †

This paper describes two methods for classifying articles using feature vectors. The first method, using lexical co-occurrences within large document databases, generates a feature vector in which the elements are generated by lexical co-occurrences and are similar to word frequency distribution. The second method, using the EDR electronic dictionary as a thesaurus, generates a feature vector in which the elements represent the similarity in meaning between words. Classifications using these methods were compared to manual classifications. In the method using lexical co-occurrences, a success rate of 98.0% was achieved for articles that were considered easy to classify manually, and 83.5% was achieved for all articles. In the method using the thesaurus, a success rate of 63.75% was achieved.

## 1. はじめに

最近ではパソコンやワープロの普及とコンピュータネットワークの発達により，電子化された文書が大量に流通するようになってきている．さらにCD-ROMによる辞書や新聞記事などの大規模文書データも普及しつつある．しかし，コンピュータネットワークやCD-ROM等から得られる文書データは，膨大すぎて人手で整理するには手に負えなくなってきている．大量の文書データを人手を介することなく自動的に整理することができれば，大量の情報の中から有益な情報を取り出しやすくなる．逆に言えばある程度自動的に文書データを処理できなければ，このような大量の文書のほとんどは情報洪水の中に埋もれてしまうことになる．

自動的に文書をフィルタリングするシステムとしてMaloneはInformation Lens[1),2)]を提案している．これはテンプレートを用いたsemistructuredな文書を利用するもので，用途によっては便利であるが，すでに膨大に存在している決められた構造を持たない普通の文書には利用できない．PollockのIscreen[3)]も同様の情報フィルタリングシステムである．

構造化されていない普通の文書に対して，Pollack[4)]やGallant[5)]は，文脈ベクトル(単語の意味の分散表現)を用いて文書の文脈をつかむ手法を提案し，芥子[6)]はGallantの文脈ベクトルをベースにした連想検索手法を提案している．芥子の手法は人手で数百の出現頻度の高い単語（コア単語）にのみ文脈ベクトルを作成すれば大量の文書データ中の統計情報をもとに全重要単語の文脈ベクトルを機械学習できるものである．しかし，この手法ではコア単語の文脈ベクトルは人手で作成する必要があり，この値は作成する人の個人差によって揺れが生じる恐れがある．

そこで，できるだけ人手を介さずに安定した特徴量を得て，これをもとに構造化されていない普通の文書をその内容に応じて自動的に選別する手法を開発する必要があると考えられる．

このような手法の実現のためには主に次の二種類の

† シャープ（株）応用システム研究所
Integrated Media Laboratories, Sharp Corporation

方法が考えられる．一つは「大量の文書データ中の統計情報を用いる手法」，もう一つは「既に人手で構築されている辞書の情報を用いる手法」である．

従来から，統計情報を用いた文書データ処理に関しては様々な研究が行われている．例えば，長尾は $\chi^2$ 値を用いて日本語文献中の重要語を自動抽出することができることを示している[7]．また，分類に関しては，梅田の，漢字一文字とカタカナ列を同等に頻度で評価しこれに基づいて文献を分類する手法[8]，亀田の，人手で作成したキー概念を用いて新聞記事を分類する手法[9]，田村の，分野ごとにキーワードを自動抽出しそのキーワードの出現頻度の偏りを用いて文書を分類する手法[10]等が提案されている．しかしこれらはいずれも十分な精度での分類はできていなかった．

最近はコンピュータ処理能力の増大にともない，再び統計情報を用いた文書データ処理に関した研究が盛んに行われるようになってきている．例えば津高はニューラルネットワークの一種である自己組織化マップを用いてカテゴリ分類やキーワード付けを自動的に行う手法[11]を提案している．

もう一つの「既に人手で構築されている辞書の情報を用いる手法」に関しては，最近はCD-ROMに収められた辞書も出てきているが，これらのほとんどは人間が閲覧することを目的として作成された辞書であり，コンピュータで扱うのには不便なものが多かった．しかし1986年から9年間の国家プロジェクトとして企画された電子化辞書プロジェクトは「EDR電子化辞書」と呼ぶ大規模で本格的な機械処理用の辞書[12]を開発しており，崔らはこのEDRの辞書を用いて単語類似度を計算する手法を提案している[13]．

最近では，大量の文書データから単語間の共起関係に基づいて得られるベクトルと，辞書の語意定義から定まる単語間距離に基づいて得られるベクトルとで能力を比較した結果も報告されている[14]．

本稿では，「大量の文書データ中の統計情報を用いる手法」として大量の文書データを用いて同一記事中の単語間共起関係から分野の特徴を表す単語出現頻度分布の近似値を求め，この値を要素とする特徴ベクトルを用いて文書を分類する手法を提案する．また，「既に人手で構築されている辞書の情報を用いる手法」としてEDRの辞書をシソーラスとして用い，単語間の類似度を求め，この単語類似度を要素とする特徴ベクトルを用いて文書を分類する手法を提案する．そしてこの二種類の分類手法について人手での分類との比較を行う．最後に単語間共起を用いた手法を改良した結果について述べる．

## 2. 単語間共起を用いた分類手法

ある分野に属する文書中に出現する単語は分野ごとに特徴があると考えられる．したがってある文書が与えられたら，その文書の属する分野の単語出現頻度分布を得ることができれば，その文書がどのような分野に属しているのかを推定することができる．そのための方法の一つとして，各単語に特徴ベクトルとして単語出現頻度分布を持たせておき，文書中に出現する単語の特徴ベクトルからその文書の特徴ベクトル（＝その文書の属する分野の単語出現頻度分布の近似値）を生成する方法が考えられる．

ここでは新聞記事一つ一つを分野ごとに分類するタスクにおいて，この特徴ベクトルをいかに構成すべきかを考えてみる．

一つの記事に注目した時，その記事はある一つの分野に属するとみなせる．そして特徴ベクトルの類似度で分野を判定するのであるから，同じ記事に属す単語は類似した特徴ベクトルを持つべきである．また，ある一つの記事の単語出現頻度分布はその記事の属する分野の単語出現頻度分布の一部を構成しているのであるから，その記事の属する分野の単語出現頻度分布を近似していると仮定することができる．すると，単語出現頻度分布を学習させるための記事を多数用意しておき，学習用の記事を読み込むたびにその記事中に出現している単語の特徴ベクトルに，その記事の単語出現頻度分布を加算するようにしておけば，多数の学習用記事を読んだ後には各単語の特徴ベクトルは，その単語が含まれていた記事の属する分野の単語出現頻度分布に類似したものになる．このようにして単語の特徴ベクトルが得られたら，ある記事が与えられたらその記事中に出現する単語の特徴ベクトルをすべて加算したものをその記事の特徴ベクトルとすれば，それはその記事の属する分野の単語出現頻度分布を近似したものになると仮定できる．

各分野の単語出現頻度分布と，各単語の特徴ベクトルのイメージを**図1**に示す．この図は「政治分野の記

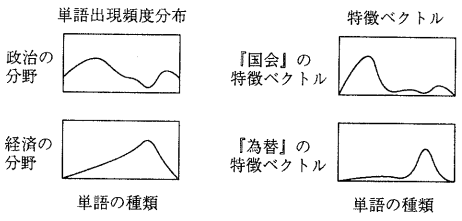



**図1** 単語出現頻度分布と特徴ベクトル
Fig. 1 Frequency distribution and feature vector.

事群に含まれている単語の出現頻度分布」,「経済分野の記事群に含まれている単語の出現頻度分布」,「学習後に『国会』という単語に付けられた特徴ベクトル」,「学習後に『為替』という単語に付けられた特徴ベクトル」のイメージを示している.

### 2.1 単語間共起を用いた特徴ベクトル

同一記事中の単語間共起を用いて特徴ベクトルを生成する方法を説明する.

単語出現頻度分布を調査する単語を $word_1, word_2, \cdots, word_n$ の $n$ 個とし,記事は $m$ 個あるとする.記事 $i$ に含まれる単語の出現頻度ベクトル $\boldsymbol{V}_i$ を

$$\boldsymbol{V}_i=(v_{i1}, v_{i2}, \cdots, v_{in}) \quad (1)$$

$v_{ij}$:記事 $i$ 中に出現する $word_j$ の個数

で表すと,単語 $word_i$ の特徴ベクトル $\boldsymbol{W}_i$ は,以下の式で表される.

$$\boldsymbol{W}_i=(w_{i1}, w_{i2}, \cdots, w_{in})=\sum_{j=1}^{m} v_{ji}\cdot\frac{\boldsymbol{V}_j}{|\boldsymbol{V}_j|} \quad (2)$$

この式からわかるように,全記事について単語の出現頻度ベクトル $\boldsymbol{V}_j$ をその記事中での出現頻度分の重み付きで加算していくため,単語の特徴ベクトル $\boldsymbol{W}_i$ は単語 $word_i$ が頻繁に含まれる記事の分野の単語出現頻度分布に類似した値を持つことになる.

記事の特徴ベクトル $\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_m$ は,単語の特徴ベクトルから以下の式で算出される*.

$$\boldsymbol{A}_i=\sum_{j=1}^{n}\log\left(\frac{m}{m_j}\right)\cdot v_{ij}\cdot\frac{\boldsymbol{W}_j}{|\boldsymbol{W}_j|} \quad (3)$$

$m_j$:$word_j$ が含まれている記事の個数

未知の記事 $u$(単語の特徴ベクトルを算出するために与えられた $m$ 個の記事以外の記事)に対する特徴ベクトル $\boldsymbol{A}_u$ も,その記事の単語の出現頻度をベクトル $\boldsymbol{V}_u=(v_{u1}, v_{u2}, \cdots, v_{un})$ で表せば,以下の式で算出される.

$$\boldsymbol{A}_u=\sum_{j=1}^{n}\log\left(\frac{m}{m_j}\right)\cdot v_{uj}\cdot\frac{\boldsymbol{W}_j}{|\boldsymbol{W}_j|} \quad (4)$$

このようにして求められた記事の特徴ベクトル(以後記事ベクトル)の値を一般に用いられている分類手法で分野に分ければ,類似した内容を持つ記事が同じ分野に分類される.しかし,この場合は各分野に分類されたものがどういう意味を持つのかを判断するのが難しくなる.そこで,実験では人手で各分野の典型的な文書を選出し,その文書との類似度によって分類を行わせることにした.

記事 $x$ と記事 $y$ との類似度は,記事ベクトル $\boldsymbol{A}_x$, $\boldsymbol{A}_y$ のそれぞれの絶対値を1に正規化してから両者の内積を求めることで得られる.分野 $i$ の典型的な記事の記事ベクトルを $\boldsymbol{C}_i$ で表すと,記事 $u$ がどの分野に属するかを判定するには,全 $C_i$ について

$$S_i=\frac{\boldsymbol{A}_u\cdot\boldsymbol{C}_i}{|\boldsymbol{A}_u|\cdot|\boldsymbol{C}_i|} \quad (5)$$

を計算し,最大類似度 $S_m=\max_i S_i$ となる $m$ を求める.すると記事 $u$ は分野 $m$ に分類される.

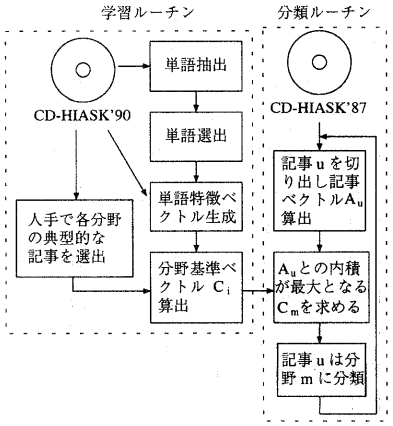



**図2** 単語間共起を用いた手法の処理フロー
Fig. 2 Flowchart of the method using lexical co-occurrence.

### 2.2 学習と分類の手順

本手法による特徴ベクトルの学習手順と新聞記事の分類手順を以下に示す.これを図示したものが**図2**である.

**学習手順**

(1) 特徴ベクトルを生成するための大量の新聞記事データを用意.
(2) このデータから単語を抽出.
(3) 抽出された単語の中から,特徴ベクトルを生成する際に使用する単語を選出.
(4) 大量の新聞記事データから単語の特徴ベクトル生成.
(5) 人手で新聞記事から各分野の典型的な記事を三つずつ選び出し,これらの記事の特徴ベクトルを平均することで各分野の特徴ベクトルを算出.(分野基準ベクトル)

**分類手順**

(1) 分類したい記事から,その記事の特徴ベクトルを算出.(記事ベクトル)
(2) 記事ベクトルと分野基準ベクトルとの類似度を求め,その記事が属している分野を決定.

## 3. シソーラスを用いた分類手法

人手で作成されたシソーラスとして,EDR の辞

* $\log(m/m_j)$ を掛けているのは,出現確率が低い単語ほど,一回の出現の重要度は高いと考えられるため.

書[15),16)]を使用した．EDRの辞書は，(株)日本電子化辞書研究所（略称EDR）において，大規模で汎用の機械用辞書を目指して開発されている電子化辞書であり，単語辞書，概念辞書，対訳辞書，共起辞書等から構成されている．単語辞書には各単語の語義情報として概念見出し（その単語がどういう概念を表しているかの情報）が記述されており，概念辞書には概念間の関係が記述されている．そして，これらの情報を用いて単語間の類似度を計算する手法が提案されている[13)]．

### 3.1 シソーラスを用いた特徴ベクトル

シソーラスを用いて単語の特徴ベクトルを生成する方法を説明する．

前章の単語間共起を用いた特徴ベクトルの生成方法では，仮に $word_i$ と $word_j$ とが同一記事中に頻繁に共起しているとすると単語の特徴ベクトルの $w_{ij}$ や $w_{ji}$ の値が大きくなる．同じ記事中に共起するということは，なんらかの点で関連していて，類似度の高い単語である場合が多いと考えられる．そこで，シソーラスを用いて単語間の類似度を算出すれば，大量のデータから共起関係を抽出しなくても単語の特徴ベクトルが得られると考えられる．

類似度の計算法は崔進らの方法[13)]を利用した．これは，まずEDRの単語辞書の情報をもとに同義関係(単語間に共通な概念が一つ以上存在する）での類似度 $\alpha$ を算出し，次に概念辞書の情報をもとに類似関係（単語間には共通な概念が存在しないが，ある上位概念の階層に共通な概念が存在する）での類似度 $\beta$ を算出する．そして，最後に $\alpha$ と $\beta$ をもとに，単語間の類似度 $\delta$ を算出する方法である．

それぞれの算出方法を説明する．

類似度 $\alpha$ は，同義関係においての類似度であり，二つの単語間に存在する共通な概念の数である．つまり単語 $word_i$ が持つ概念の集合を $C_i$ で表し，集合 $X$ に存在する要素の個数を $\|X\|$ で表すと，単語 $word_x$ と単語 $word_y$ の類似度 $\alpha$ は，

$$\alpha=\|C_x\cap C_y\| \tag{6}$$

となる．

類似度 $\beta$ は，類似関係における類似度であり，単語の持つ概念の上位概念に共通な概念がどの程度存在するかを表す．概念集合 $C_x, C_y$ が与えられたら，上位概念を $N$ 段目まで検索し，各段階での上位概念集合 $S_k^x$ および $S_k^y\{k=1,2,\cdots,N\}$ を算出する．次に以下の式を用いて階層 $k$ での類似度 $\beta_k$ を算出する．

$$\beta_k=(1+K_{\beta 1}\cdot CS_k)\times\left(1+K_{\beta 2}\cdot\left(\frac{CS_k}{N_{kx}}+\frac{CS_k}{N_{ky}}\right)\right)-1 \tag{7}$$

この式で，$N_{ki}\{i=x,y\}$ は $word_i$ の $k$ 段目の上位概念集合内の異なる概念数，$CS_k$ は $S_k^x$ と $S_k^y$ の間に共通な上位概念の個数，$K_{\beta 1}, K_{\beta 2}$ は定数である．最後に以下の式で類似度 $\beta$ を算出する．($K_i$ は定数)

$$\beta=\sum_{i=1}^{N}K_i\cdot\beta_i \tag{8}$$

以上より $\alpha,\beta$ を求めたら，単語 $word_x$ と単語 $word_y$ の類似度 $\delta$ を，以下の式で算出する．($K_\alpha, K_\beta$ は定数)

$$\delta=1-e^{-(K_\alpha\cdot\alpha+K_\beta\cdot\beta)} \tag{9}$$

$\delta$ は，0から1の範囲内を変動し，類似度が高いほど1に近付く．

各式は定数（重み値）を変化させることで類似度の値を調整することができるが，今回は崔進らの論文[13)]中に $N=2$ の時の重み値のサンプルとして掲載されている値を用いた．すなわち，$N=2, K_\alpha=0.45, K_\beta=0.028, K_{\beta 1}=2.75, K_{\beta 2}=8.25, K_1=1, K_2=0.05$ として類似度を算出した．

特徴ベクトルを作成する単語を $word_1, word_2, \cdots, word_n$ の $n$ 個とし，式（9）で算出される単語 $word_x$ と単語 $word_y$ の類似度を $\delta_{xy}$ で表すと，単語 $word_i$ の特徴ベクトル $\boldsymbol{W}_i$ は，以下の式で表される．

$$\boldsymbol{W}_i=(\delta_{i1},\delta_{i2},\cdots,\delta_{in}) \tag{10}$$

式（2）では $\boldsymbol{W}_i$ の各要素は，出現頻度に応じた値になっていたので．例えば仮に $word_i$ と $word_j$ とが同一記事中に頻繁に共起しているとすると $w_{ij}$ の値が大きくなった．一方，式（10）の場合は $word_i$ と $word_j$ の類似度が高いと $w_{ij}$ の値が大きくなる．

単語の特徴ベクトルから記事の特徴ベクトルを求める方法は，前章の方法と同様であり，記事 $i$ に含まれる単語の出現頻度ベクトル $\boldsymbol{V}_i$ を

$$\boldsymbol{V}_i=(v_{i1},v_{i2},\cdots,v_{in}) \tag{11}$$

$v_{ij}$：記事 $i$ 中に出現する $word_j$ の個数

で表すと，記事の特徴ベクトル $\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_m$ は，

$$\boldsymbol{A}_i=\sum_{j=1}^{n}v_{ij}\cdot\frac{\boldsymbol{W}_j}{|\boldsymbol{W}_j|} \tag{12}$$

として求められる．

記事の特徴ベクトルを用いて記事を分類する方法も前章の方法とまったく同じである．

## 4. 実　　験

以上の二つの手法の有効性を確認するために，これら二つの手法による分類と人手による分類とを比較する以下の二つの実験を行った．

●実験1

朝日新聞1987年の記事中で人間にとって分類が易しい20記事と難しい20記事を，本手法で分

類した結果と20人が分類した結果との比較.

●実験2

朝日新聞1987年の400記事を1人が分類した結果との比較.

これらの実験に使用したデータ等は以下のとおりである.

(1) CD-HIASK (朝日新聞のCD-ROM) 1990年版[17] (約150Mバイト, 101966記事)

単語間共起を用いた手法において, 単語の特徴ベクトルを生成するためのデータとして使用した. また各分野の典型的な記事もここから抜き出し, 分野基準ベクトルを生成する時にも使用した.

(2) CD-HIASK (朝日新聞のCD-ROM) 1987年版[18]から抜き出した記事

人間の分類と本手法の分類の比較に使用した. これの一例を以下に示す.

---

物価と為替の安定維持が最大の課題　澄田日銀総裁が語る

澄田日銀総裁は31日, 朝日新聞とのインタビューで, 新年の金融政策について, 物価の安定維持が最大の課題であることを強調しつつ, 内需拡大, 対外不均衡の是正に取り組む姿勢を明らかにした. その一方で, 日本経済が国際的に影響力を増していることを踏まえ, 国際協調がますます重要になっていることを指摘しながらも, 金融政策が外圧や国内政治からの独立性と自主性を確保することが一層……

---

(3) EDR電子化辞書評価版第2版

単語データの抽出にEDR電子化辞書の日本語単語辞書評価版第2版[15]を使用し, 単語間の類似度の計算に上記日本語単語辞書とEDR電子化辞書の概念辞書評価版第2版[16]を使用した. 評価版第2版の日本語単語辞書の登録語数は基本語約15.5万語, 専門用語約4.2万語であり, 評価版第2版の概念辞書の収録概念数は約30万概念である.

(4) 単語データ

EDR電子化辞書の日本語単語辞書評価版第2版[15]中から「平仮名だけからなる三文字以下の単語」と「漢字以外の一文字単語」を除いた全単語 (約17万語) を使用した. 文書から単語を抽出する方法は, 長さの長いものを優先して選択するパターンマッチング (最長一致法) によるが, 連続する二単語を複数の組み合わせで抽出できる場合には, その二単語の合計の長さが最長になる組み合わせの最初の単語を選択する手法 (二文節最長一致法) を用いた. ただし誤抽出をできるだけ減らすため, 漢字一文字の単語の場合は前後が非漢字の場合のみ抽出した. 朝日新聞1990/1/1朝刊の最初から抽出された500単語について調査した結果, この方法で95%程度正しく抽出されることを確認した.

単語出現頻度分布は理想的には全単語の分布を調べるべきだが, コンピュータの記憶容量等の関係で, 4096個以下の単語に制限して実験した. この単語の選出方法は, 単純に朝日新聞1990年版の中で出現頻度の高いものから順番に選出した (抽出単語数は4096, 2048, 1024, 512, 256, 128, 64の7種類で実験した).

(5) 特徴ベクトル

前述の各方法で特徴ベクトルを算出するが, ベクトルの要素の精度は32 bit浮動小数点数で計算した.

### 4.1 実験1

「政治」「経済」「事件」「国際」のどれかに属すると考えられる記事を, 朝日新聞のCD-ROMの1987年版から40記事, 人手で抜き出した. ただし抜き出す時には, その中の20記事は比較的簡単に分類できるものを選び, 残りの20記事は分野をまたがっているなど分類が難しい記事を選んだ.

この40記事を情報関係のエンジニア20人 (男性14人, 女性6人) に分類してもらったが, その際に分類の目安として以下のキーワードを提示した.

- 「政治」政治, 国会, 首相
- 「経済」経済, 為替, 金利
- 「事件」事件, 犯罪, 裁判
- 「国際」国際, 軍事, 戦争

分類に迷った場合でも必ず一つの分野を選択するようにしてもらった.

各記事について, 最も選択した人が多かった分野を正解の分野とみなして, 20人による分類結果と本手法での分類結果とで正解率を比較した.

### 4.2 実験2

朝日新聞1987年の先頭から400記事を抜き出し, 1人が各記事を「政治」「経済」「国際」「社会 (犯罪, 事件)」「社会 (教育, 人間)」の五つの分野のどれかに分類した. 分類する際にどうしても一つの分野にしぼり切れない記事については二つの分野に分類することを許した. 二つの分野に分類された記事は107記事である. この1人の分類結果を正解とみなして本手法での

分類結果の正解率を求めた．

## 5. 結　　果

### 5.1 実験1の人手による分類結果

評価用のデータ40記事を20人の人によって，四つの分野に分類してもらった．

各記事がどの分野として選ばれたかを表にすると**表1**のようになった．易しい記事の記事番号は1～20，難しい記事の記事番号は21～40である．

この結果から各記事は最も多く選ばれた分野に属するとして，人手による分類の正解率を計算すると，

易しい記事の正解率98.5%

難しい記事の正解率82.75%

となる．

人手による分類で誤って分類された記事はすべて，複数の分野に属しているとみなすことができるものばかりであった．その複数の分野の中のどの分野に分類するかが人によって異なるということが誤分類の原因になっていると考えられる．

### 5.2 本手法による分類結果

実験1の結果を**表2**に，実験2の結果を**表3**に示す．

**表2**中の「易」「難」「全体」は以下のとおりである．

「易」：人間にとって分類が易しい20記事での正解率

「難」：人間にとって分類が難しい20記事での正解率

「全体」：40記事全体での正解率

表3中の「易」「難」「全体」は以下のとおりである．

「易」：分類選択時の一位候補の類似度（記事ベクトルと一位候補の分野基準ベクトルとの内積）と二位候補の類似度（記事ベクトルと二位候補の分野基準ベクトルとの内積）の比が大きい200記事での正解率

「難」：上記の「易」以外の200記事での正解率

「全体」：400記事全体での正解率

### 5.3 学習や分類に要する時間

単語間共起を利用した手法で朝日新聞1年分の記事から単語の特徴ベクトルを学習するのに必要な時間と，実験1における人手による分類に要した時間とベクトルの次元数が2048の場合に本手法による分類に要した時間を**表4**に示す．なお，本手法はSUN SPARCstation 10上にて実行し，分類時間には特徴ベクトルを外部記憶装置から読み込むための時間等の初期設定時間は含まれていない．機械による分類時間は，単語間共起を利用した場合とシソーラスを利用した場合とで違いはなかった．

なお，今回の実験では分類精度を向上させるのを主目的にしていたため，速度を向上させるための処理は何もしていない．したがって，この値はあくまでも参

**表1** 人手による分類結果[人]
Table 1 Result of the manual classification.

| 記事番号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 政治に分類 | 0 | 20 | 0 | 0 | 18 | 0 | 1 | 0 | 0 | 20 |
| 経済に分類 | 20 | 0 | 20 | 0 | 2 | 0 | 0 | 20 | 0 | 0 |
| 事件に分類 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 20 | 0 |
| 国際に分類 | 0 | 0 | 0 | 20 | 0 | 0 | 19 | 0 | 0 | 0 |
| 記事番号 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 政治に分類 | 0 | 0 | 1 | 0 | 19 | 0 | 1 | 20 | 0 | 0 |
| 経済に分類 | 20 | 0 | 0 | 0 | 1 | 0 | 19 | 0 | 0 | 0 |
| 事件に分類 | 0 | 20 | 0 | 0 | 0 | 20 | 0 | 0 | 20 | 0 |
| 国際に分類 | 0 | 0 | 19 | 20 | 0 | 0 | 0 | 0 | 0 | 20 |
| 記事番号 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| 政治に分類 | 16 | 0 | 16 | 10 | 18 | 2 | 1 | 0 | 2 | 0 |
| 経済に分類 | 1 | 0 | 4 | 1 | 0 | 0 | 11 | 17 | 17 | 0 |
| 事件に分類 | 3 | 20 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 20 |
| 国際に分類 | 0 | 0 | 0 | 9 | 2 | 18 | 8 | 2 | 1 | 0 |
| 記事番号 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
| 政治に分類 | 0 | 18 | 0 | 0 | 15 | 3 | 14 | 0 | 0 | 0 |
| 経済に分類 | 0 | 2 | 17 | 0 | 3 | 0 | 5 | 0 | 13 | 17 |
| 事件に分類 | 20 | 0 | 1 | 3 | 0 | 17 | 0 | 20 | 0 | 0 |
| 国際に分類 | 0 | 0 | 2 | 17 | 2 | 0 | 1 | 0 | 7 | 3 |

**表2** 実験1の結果[%]
Table 2 Result of the experiment 1.

| 次元数 | | 4096 | 2048 | 1024 | 512 | 256 | 128 | 64 |
|---|---|---|---|---|---|---|---|---|
| 単語間 | 易 | 90.0 | 95.0 | 95.0 | 95.0 | 90.0 | 85.0 | 80.0 |
| 共起を | 難 | 50.0 | 50.0 | 50.0 | 50.0 | 40.0 | 60.0 | 65.0 |
| 利用 | 全体 | 70.0 | 72.5 | 72.5 | 72.5 | 65.0 | 72.5 | 72.5 |
| シソー | 易 | 80.0 | 70.0 | 75.0 | 80.0 | 70.0 | 65.0 | 60.0 |
| ラスを | 難 | 40.0 | 40.0 | 40.0 | 45.0 | 40.0 | 30.0 | 40.0 |
| 利用 | 全体 | 60.0 | 55.0 | 57.5 | 62.5 | 55.0 | 47.5 | 50.0 |

**表3** 実験2の結果[%]
Table 3 Result of the experiment 2.

| 次元数 | | 4096 | 2048 | 1024 | 512 | 256 | 128 | 64 |
|---|---|---|---|---|---|---|---|---|
| 単語間 | 易 | 98.0 | 98.0 | 96.0 | 96.5 | 92.0 | 84.0 | 80.5 |
| 共起を | 難 | 67.5 | 69.0 | 66.0 | 57.5 | 57.5 | 48.5 | 44.0 |
| 利用 | 全体 | 82.75 | 83.5 | 81.0 | 77.0 | 74.75 | 65.75 | 62.25 |
| シソー | 易 | 75.0 | 80.5 | 75.5 | 75.5 | 74.5 | 71.0 | 67.5 |
| ラスを | 難 | 47.5 | 43.5 | 52.0 | 48.5 | 49.0 | 46.0 | 39.0 |
| 利用 | 全体 | 61.25 | 62.0 | 63.75 | 62.00 | 61.75 | 58.5 | 53.25 |

表4 所要時間
Table 4 The necessary time.

| | 人手による分類 | 機械による分類 | 倍率 |
|---|---|---|---|
| 学習時間 | 0 | 約7時間 | — |
| 初期設定時間 | 0 | 60秒 | — |
| 分類時間 | 平均約30分 (20分～1時間) | 16秒 | 約100倍 |

考値としてとらえて頂きたい．

また，ベクトルの次元数を64～4096に変化させても学習時間にはほとんど変化はなく，分類時間の変化は0.625倍～1.5倍の範囲に収まっていた．これは処理時間のうちのかなりの割合を単語抽出のためにとられているためと考えられる*．分類のための初期設定時間は，ベクトルの次元数が64～512の場合は22～25秒，ベクトルの次元数が1024の場合は33秒，ベクトルの次元数が4096の場合は160秒であった．

### 5.4 考　　察

以上の結果より，以下のことがわかった．

●単語間共起を用いた場合とシソーラスを用いた場合との比較

表2や表3を見ると，単語間共起を用いた場合に比べて，シソーラスを用いた場合はかなり分類の正解率が低いことがわかる．表3の「全体」に注目すると，単語間共起を用いた場合は，ベクトルの次元数が2048の時に正解率は83.5%と最高になったが，シソーラスを用いた場合の正解率は最高でもベクトルの次元数が1024の時の63.75%であった．また，表3の「易」に注目すると，ベクトルの次元数が2048の時，単語間共起を用いた場合には98.0%という高い正解率が得られたのに対し，シソーラスを用いた場合には正解率は80.5%にとどまった．

単語間共起を用いた手法は，その特徴ベクトルの生成法から考えて，学習用のデータと同じような傾向の文書を分類するのに有利な手法である．今回の実験では学習データも評価データも朝日新聞の記事を用いたが，もし学習データと評価データとで全く種類の異なる文書データを用いた場合（例えば朝日新聞を学習データとし，評価は研究論文をその研究テーマで分類する等）には，このような高い分類正解率は得られないと考えられる．

逆にシソーラスを用いた手法は学習用のデータを必要としないので，分類する文書の性質によらずに安定した分類を行える可能性がある．

また，本論文のシソーラスを用いた手法は，EDRの概念辞書の上位概念をたどる回数$N$を2として求めた単語類似度の値を特徴ベクトルの要素とする単純な手法であり，EDRの辞書が持つ概念間の複雑な関係の情報が失われてしまっている．また，単語類似度を求める際の重み値（定数）の値も最適なものを調査したわけではない．これらの点もシソーラスを用いた場合の正解率が低い原因になっていると考えられる．

●分類が簡単な記事について

表2から，実験1において単語間共起を用いた場合は，ベクトルの次元数が256以上であれば，人間にとって分類が易しい記事については90%以上の正解率が得られることがわかる．

表3から，単語間共起を用いた場合は記事の難易度は分野の一位候補の類似度と二位候補の類似度との比の大小によってある程度判定できることがわかる．これによって難易を判定した場合，ベクトルの次元数が4096の時と2048の時には易しい記事での分類正解率は98.0%となり，実験1の人間の易しい記事の分類の正解率である98.5%とほぼ同じ値が得られた．したがって，易しい記事だけを選択して自動分類すれば，それは人間と同程度の正確さで自動分類ができることがわかる．

●分類が難しい記事について

表2より，実験1において単語間共起を用いた場合は，分類の難しい記事については，ベクトルの次元数を大きくしても正解率の向上にはつながらないことがわかる．分類の難しい記事は分野をまたがっている記事を多く含んでいるため，記事中に使われている単語で分野を判定すると間違えるものが多い．したがって本手法のように記事中に使われている単語だけで分野を判定するのでは限界があることがわかる．

分類の難しい記事の分類の正解率を向上させるためには，文や文章の構造等も分析するような手法を取り入れる必要があると考えられる．

## 6. 単語間共起を用いた手法の改良

前章までの結果から，シソーラスを用いた手法より単語間共起を用いた手法の方が分類の正解率が高いことが確認できたので，単語間共起を用いた手法の分類

* 単語抽出はベクトルの次元数によらず，EDR電子化辞書の日本語単語辞書評価版第2版中から「平仮名だけからなる三文字以下の単語」と「漢字以外の一文字単語」を除いた全単語（約17万語）を常に使用している．

表5 各単語選別方法を用いた時の実験2の結果[%]
Table 5 Result of the experiment 2 when the each method of word filtering was used.

| 次元数 | 4096 | 2048 | 1024 | 512 | 256 | 128 | 64 |
|---|---|---|---|---|---|---|---|
| A:出現頻度順 | 82.75 | 83.5 | 81.0 | 77.0 | 74.75 | 65.75 | 62.25 |
| B:分野毎分散大 | 86.75 | 86.0 | 83.75 | 80.0 | 77.0 | 72.5 | 70.75 |
| C:特定分野多出 | 89.25 | 86.25 | 86.75 | 82.75 | 82.25 | 71.25 | 64.5 |

正解率をさらに向上させることを試みた.

特徴ベクトルを生成する際に,単語を選別する必要があったが,先の実験ではこれは単純に出現頻度の多い順に選択していた.この選別方法を用いると,出現頻度の高い単語が特徴ベクトルを持っていることになるので,比較的短い文書にも特徴ベクトルを持っている単語が含まれる可能性が高くなり,特徴ベクトルを計算できない記事の割合は小さくなる.その反面,どの分野の文書にも含まれている単語は,本来は分類には不要な単語であるのに,特徴ベクトルを持たせる単語に選出されてしまう.このような単語が分類の正解率に悪影響を与えている可能性がある.

そこで,各単語が分野ごとにどのような出現頻度を持っているかを調査し,分野ごとの出現頻度が偏っている単語だけを選出して特徴ベクトルを付けるようにしてみるとどうなるかを実験した.

単語の選出方法としては以下の方法B,方法Cの二通りを行なった.(方法Aは第2章で用いた方法)

方法A:出現頻度の高い方から順番に選出

方法B:分野ごとの出現頻度の分散が大きい単語を選出

方法C:特定の分野にのみ多く出現する単語を選出

これらについての実験2の結果を**表5**に示す.

表5を見ると,方法Aのように単純に出現頻度の多い順に単語を選出するより,方法Bや方法Cのように分類に有効と考えられる単語だけを使用するようにしたほうが分類の正解率が上がることが確認できる.

特に方法Cは,ベクトルの次元数が4096の時に89.25%という今回の各方法の中では最も高い正解率を得ることができた.方法Cと方法Bとを比較すると,ベクトルの次元数が大きい場合には方法Cの方が正解率が高いが,ベクトルの次元数が小さい場合には方法Bの方が正解率が高くなった.方法Cは特定の分野にのみ多く出現する単語を選出しているため,ある程度以上単語数を減らすと記事中に特徴ベクトルを持つ単語が少なくなってしまい,正解率が下がる.逆に単語数を多くとれば,記事中に特徴ベクトルを持つ単語が多くなり,しかもこれらの単語の中には複数の分野に出現するようなあいまいな単語は少ないので,高い正解率が得られたと考えられる.

## 7. 終わりに

本稿では,「大量の文書データ中の統計情報を用いる手法」として,大量の文書データから得られる単語間共起関係を用いて単語の特徴ベクトルを生成し,この特徴ベクトルを用いて記事を分類する手法を提案した.そして,朝日新聞の1年分の文書データで特徴ベクトルを学習した後に,新聞記事を4種類あるいは5種類の分野に分類する実験を行ったところ,特徴ベクトルの次元数を2048とすると83.5%の記事が正しく分類され,そのうち人手で容易に分類が可能な記事については95%以上正しく分類されることが確認できた.さらに,単語の特徴ベクトルを作成する時に使用する単語を選別すれば分類の正解率が向上することが確認できた.

また,「既に人手で構築されている辞書の情報を用いる手法」として,シソーラスを用いて単語の特徴ベクトルを生成し,この特徴ベクトルを用いて記事を分類する手法を提案した.こちらは特徴ベクトルの生成に大量の文書データを用意する必要はないが,分類の正解率は最高でも63.75%と,単語間共起を用いた分類手法に比べてかなり低いことがわかった.この理由としては以下のことが考えられる.

- シソーラスを用いた特徴ベクトルの生成では,崔進らによる単語類似度計算法[13]で上位概念をたどる回数$N$を2とした場合の類似度の値をそのまま利用したが,これによってEDRの辞書が持つ概念間の複雑な関係の情報が失われてしまっている.
- 単語間共起を用いた分類手法は学習データと良く似た文書データを分類するのに適していると考えられるのに対し,シソーラスを用いた分類手法は文書データの性質によらずに安定した分類を行える可能性がある.

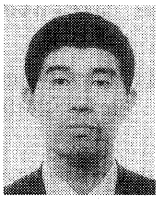
今後の課題としては,大量の文書データ中の統計情報と,人手で構築されている辞書の情報とを両方用いることで,より分類の正解率を高めた手法が得られないかということ等を検討していきたい.

**謝辞** 本研究にあたり,CD-HIASKの使用を了解いただいた朝日新聞社ニューメディア本部の関係者の方々に感謝いたします.また本研究の機会を与えて下さった応用システム研究所所長中島隆之氏に感謝いたします.

## 参考文献


1) Malone, T. W. et al.: Semistructured Messages Are Surprisingly Useful for Computer-Supported Coordination, *ACM Trans. Office Information Systems,* Vol. 5, No. 2, pp. 115-131 (1987).
2) Malone, T. W., Grant, K. R., Turbak, F. A., Brobest, S. A. and Cohen, M. D.: Intelligent Information-Sharing Systems, *Comm. ACM,* Vol. 30, No. 5, pp. 380-402 (1987).
3) Pollock, S.: A Rule-Based Message Filtering System, *ACM Trans. Office Information Systems,* Vol. 6, No. 3, pp. 232-254 (1988).
4) Waltz, D. L. and Pollack, J. B.: Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation, *Cognitive Science,* Vol. 9, pp. 51-74 (1985).
5) Gallant, S. I.: A Practical Approach for Representing Context and for Performing Word Sense Disambiguation Using Neural Networks, *Neural Computation,* Vol. 3, pp. 293-309 (1991).
6) 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, 電子情報通信学会, AI 92-99, pp. 73-80 (1993).
7) 長尾 真, 水谷幹男, 池田浩之: 日本語文献における重要語の自動抽出, 情報処理, Vol. 17, No. 2, pp. 110-117 (1976).
8) 梅田茂樹, 細野公男ほか: 漢字カタカナ列の頻度情報に基づいた日本語文献の自動分類, 第32回情報処理学会全国大会論文集, 4 T-10, pp. 1687-1688 (1986).
9) 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).
10) 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, 第36回情報処理学会全国大会論文集, 6 U-5, pp. 1305-1306 (1988).
11) 津高新一郎: 自己組織化マップを用いたテキスト自動分類の試み, 第46回情報処理学会全国大会論文集, 5 G-1, 分冊4, pp. 187-188 (1993).
12) EDR電子化辞書©株式会社日本電子化辞書研究所.
13) 崔 進, 小松英二, 安原 宏: EDR電子化辞書を用いた単語類似度計算法, 情報処理学会自然言語処理研究会報告, NL 93-1, pp. 1-6 (1993).
14) Niwa, Y. and Nitta, Y.: Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries, *Proc. COLING 94,* Vol. 1, pp. 304-309 (1994).
15) EDR電子化辞書 日本語単語辞書評価版第2版, ©株式会社日本電子化辞書研究所.
16) EDR電子化辞書 概念辞書評価版第2版, ©株式会社日本電子化辞書研究所.
17) CD-HIASK 朝日新聞全文記事情報1990年版, 紀伊国屋書店, 日外アソシエーツ.
18) CD-HIASK 朝日新聞全文記事情報1987年版, 紀伊国屋書店, 日外アソシエーツ.


(平成7年10月21日受付)
(平成7年 4 月14日採録)

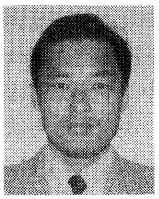


**湯浅 夏樹** (正会員)

1967年生. 1990年東京工業大学工学部情報工学科卒業. 1992年同大学院理工学研究科修士課程(情報工学専攻)修了. 同年シャープ株式会社入社. 応用システム研究所に勤務. 自然言語処理, ヒューマンインタフェース等の研究に従事.

**上田 徹** (正会員)

1959年生. 1981年大阪大学工学部通信工学科卒業. 同年シャープ株式会社入社. 現在同社応用システム研究所に勤務. 音声認識, ニューラルネット, 統計的な言語処理等の研究に従事. 電子情報通信学会, 音響学会各会員.



**外川 文雄**

1953年生. 1976年金沢大学工学部電子工学科卒業. 同年シャープ株式会社入社. 1989-1991年米国オレゴン州でニューロコンピュータ技術共同開発. 現在同社応用システム研究所に勤務. 音声認識, 文字認識, ニューラルネット, ヒューマンインタフェース等の研究に従事.