

概念識別子の頻度分布を利用した文書分類

湯浅夏樹 外川文雄

シャープ(株)応用システム研究所

〒261 千葉市美浜区中瀬 1-9-2

概要

本稿では、大量新聞記事テキストデータを用いて概念識別子の頻度分布を学習し、文書を分類する手法について述べる。概念識別子とは EDR 単語辞書に記述されている「単語の概念を表す数値」のことである。

朝日新聞一年分のデータ中の概念識別子の出現頻度分布から各概念識別子に対応する特徴ベクトルを生成し、新聞の記事を分類する手法を開発した。この手法を用いた分類を人手による分類と比較したところ、特徴ベクトルの次元数を 2048 とした場合、本手法で 85% 程度の記事が正しく分類され、そのうち分類が易しいと考えられる記事だけに限定すれば 98% の記事が正しく分類されることが確認できた。

Classifying Articles Using Frequency Distribution of the Concept ID

Natsuki Yuasa Fumio Togawa

Integrated Media Laboratories, Sharp Corporation

1-9-2, Nakase, Mihama-ku, Chiba-shi, Chiba, 261 Japan

Abstract

This paper describes a method for classifying articles using frequency distribution of the concept ID. The concept ID is a number which shows a concept of a word, and it is described in the EDR word dictionary.

We developed a method for generating a feature vector automatically using statistical information derived from the concept IDs in a large document database consisting of one year's content of the Asahi Shimbun newspaper. Classification using this feature vector of 2048 dimensions was compared to manual classification, and a success rate of 98% was achieved for articles that were considered easy to classify manually, and 85% for all articles.

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、膨大過ぎて人手で整理するには手に負えなくなっている。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だろう。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中に埋もれてしまうことになる。

そこで、近年大規模テキストデータベースの統計情報を用いて文書を自動分類する研究が盛んに行なわれている[1]。

我々はこれまでに、この自動分類の手法の一つとして、同一記事内の名詞間共起関係から名詞の特徴ベクトルを生成し、この特徴ベクトルを用いて文書を分類する手法を提案した[2]。

しかし、この手法では、単純に表層の名詞の共起関係だけを調べているので、表記の揺れや、同義語は異なる名詞として扱われる。大量の文書データの統計情報を利用することで、表記の揺れや同義語の影響は少なくなるが、意味の同じものは同じものとして扱った方が良いと考えられる。

最近日本電子化辞書研究所(略称EDR)からEDR電子化辞書[3]が発表されたが、このEDR電子化辞書の単語辞書には各単語にその単語の意味(概念)を区別するための概念識別子(concept ID)が付加されている。これは表記が異なっても概念の同じものは同じ番号が割り当てられている。また概念の同じ単語に関しては日本語の単語も英語の単語も同じ番号が割り当てられている。

そこで、この概念識別子について大量の文書データから統計情報を調べれば、単純に名詞を用いた場合よりも良い結果が得られることが期待できる。

本稿では大量の文書データを用いて概念識別子の出現頻度分布を抽出し、これを用いて文書の分類を行なう手法について述べる。そして、この手法での分類と、人手による分類とを比較した結果について述べる。

2 出現頻度分布の抽出と分類手法

ある分野に属する文章中に出現している単語は分野毎に特徴があると考えられる。つまりある文章が与えられた時にその文章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、その文章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。

これと同様に、ある分野に属する文章中に出現している単語の概念識別子は分野毎に特徴があると考えられ、ある文章が与えられた時にその文章の属する分野の概念識別子出現頻度分布を得ることができれば、その文章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。

この概念識別子出現頻度分布を得るための方法の一つとして、各概念識別子に特徴ベクトルとして概念識別子出現頻度分布の推定値を持たせておき、文章中に出現する概念識別子の特徴ベクトルからその文章の特徴ベクトルを生成する方法が考えられる。

ここで問題なのは各概念識別子に持たせる特徴ベクトルの生成方法である。

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とする。その場合、特徴ベクトルの類似度で分野を判定するのであるから、同じ記事に属す概念識別子は類似した特徴ベクトルを持つべきである。また、ある一つの記事の概念識別子出現頻度分布はその記事の属する分野の概念識別子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概念識別子出現頻度分布を近似していると仮定する。すると、概念識別子の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概念識別子の特徴ベクトルに、その記事の概念識別子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概念識別子の特徴ベクトルは、その概念識別子が含まれていた記事の属する分野の概念識別子出現頻度分布に近付く。このようにして概念識別子の特徴ベクトルが得られたら、今度はある記事が与えられたらその記事中に出現する概念識別子の特徴ベクトルを全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概念識別子出現頻度分布を近似したものになると仮定できる。

ただし、一般に一つの単語には複数の概念識別子が対応していることが多く、一対一で「単語→概念識

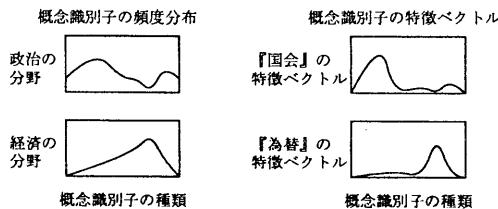


図 1: 概念識別子出現頻度分布と特徴ベクトル

別子」の変換を行なうことはできない。また、EDR 電子化辞書の単語辞書の各概念識別子には EDR コーパス内でのその概念の頻度の情報が付加されているので、この頻度情報も用いた方が良いと考えられる。以上を考慮して概念識別子の特徴ベクトルを生成することにする。

各分野の概念識別子出現頻度分布と、各概念識別子の特徴ベクトルのイメージを図 1 に示す。この図は「政治分野の記事群に含まれている概念識別子の出現頻度分布」、「経済分野の記事群に含まれている概念識別子の出現頻度分布」、「学習後に『国会』という概念識別子に付けられた特徴ベクトル」、「学習後に『為替』という概念識別子に付けられた特徴ベクトル」のイメージを示している。

以上の考え方をもとに特徴ベクトルを生成し、この特徴ベクトルの値によって実際に新聞記事を分類し、人間の分類結果と比較を行なう。

2.1 概念識別子出現頻度分布を利用した特徴ベクトル

特徴ベクトルの生成方法を説明する。

文書を形態素解析する単語を $word_1, word_2, \dots, word_p$ の p 個とし、特徴ベクトルを持たせる概念識別子を $cid_1, cid_2, \dots, cid_n$ の n 個とし、概念識別子の特徴ベクトルを学習するために用意された記事は 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 \mathbf{V}_i を

$$\mathbf{V}_i = (v_{i1}, v_{i2}, \dots, v_{ip}) \quad (1)$$

v_{ij} : 記事 i 中に出現する単語 $word_j$ の個数

で表す。

EDR 単語辞書の各単語には対応する概念識別子と、その（単語、概念識別子）のペアの EDR コー

パス内の出現頻度が記述されている。同じ表記の単語でも複数の概念を持つものが多いが、そのような単語は複数の概念識別子とそれぞれの概念においての出現頻度が記述されている。

理想的にはある単語が出現した時、その文脈での最適な概念識別子を決定し、その概念識別子がその文章中で出現したと考えるべきだが、最適な概念識別子の決定を自動的に行なうのは困難である。そこで、今回の実験では以下の二種類の方法を試みた。

1. 方法 1

その単語に対応する全ての概念識別子がその文章中で出現したことにする。ただし頻度情報の比に応じて、各概念識別子の出現頻度を重み付けする。

2. 方法 2

その単語に対応する全ての概念識別子のうち、頻度情報が最大のもの一つだけがその文章中で出現したことにする。

頻度情報はあくまでも EDR コーパス内でのものであり、比較的多くの単語（全体の約 80%）において、頻度情報=0 となっているので、単語に対応する概念識別子の出現頻度が全て 0 であるなら、方法 1 においては、全ての概念識別子が出現したことにして、方法 2 においては、その単語に最初に対応付けられている概念識別子が出現したこととした。

単語 $word_i$ と概念識別子 cid_j との関連の強さを返す関数を $r(word_i, cid_j)$ とする。これは頻度情報によって決定される。なお $(word_i, cid_j)$ の頻度情報を f_{ij} で表すことにして、 $word_i$ に対応する cid_j は、 $cid_{j_1} \dots cid_{j_q}$ の q 個であるとする。

方法 1 の場合は

$$r(word_i, cid_j) = \frac{f_{ij}}{\sum_{k=j_1}^{j_q} f_{ik}} \quad (2)$$

となり、方法 2 の場合は

$$r(word_i, cid_j) = \begin{cases} 1 & f_{ij} = \max_{k=j_1}^{j_q} f_{ik} \\ 0 & f_{ij} \neq \max_{k=j_1}^{j_q} f_{ik} \end{cases} \quad (3)$$

となる。

ただし、全ての $f_{ik} = 0$ ($k = j_1 \dots j_q$) の場合は、方法 1 の場合は

$$r(word_i, cid_j) = \frac{1}{q} \quad (4)$$

となり、方法 2 の場合は

$$r(word_i, cid_j) = \begin{cases} 1 & j = j_1 \\ 0 & j \neq j_1 \end{cases} \quad (5)$$

となる。

以上より、記事 i の概念識別子の出現頻度ベクトル \mathbf{U}_i を

$$\begin{aligned} \mathbf{U}_i &= (u_{i1}, u_{i2}, \dots, u_{in}) \\ u_{ij} &= \sum_{k=1}^p v_{ik} \cdot r(word_k, cid_j) \end{aligned} \quad (6)$$

と定義する。

概念識別子 cid_j の特徴ベクトル \mathbf{W}_j は、 \mathbf{U}_i を用いて以下の式で表される。

$$\mathbf{W}_j = (w_{j1}, w_{j2}, \dots, w_{jn}) = \sum_{i=1}^m u_{ij} \cdot \frac{\mathbf{U}_i}{|\mathbf{U}_i|} \quad (7)$$

この式からわかるように、全記事について概念識別子の出現頻度ベクトル \mathbf{U}_i 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、概念識別子 cid_j の特徴ベクトル \mathbf{W}_j は概念識別子 cid_j が頻繁に含まれる記事の分野の概念識別子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 $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_m$ は、概念識別子の特徴ベクトルから以下の式で算出される。

$$\mathbf{A}_i = \sum_{j=1}^n u_{ij} \cdot \frac{\mathbf{W}_j}{|\mathbf{W}_j|} \quad (8)$$

概念識別子の特徴ベクトルは以下のように生成される。この処理フローを図 2 に示す。

こうして求められた概念識別子の特徴ベクトルは、その概念識別子がよく使われている記事群全体の概念識別子出現分布を表す。従って、任意の文章が与えられたら、その文章中の概念識別子の特徴ベクトルの和を取ることで、その文章が属している分野の概念識別子出現分布の推定値が得られる。この処理フローを図 3 に示す。

これらの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 2 つの分野に分ければ記事の分類が行なえる。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章を一つずつ選出し、そ

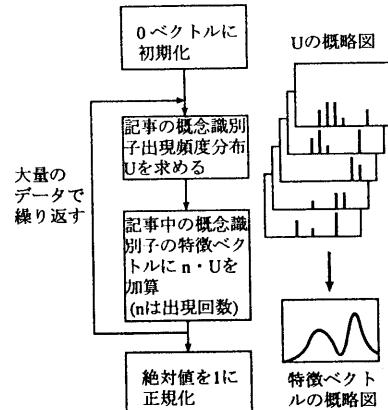


図 2: 概念識別子の特徴ベクトルを求める処理フロー

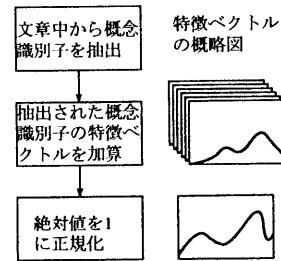


図 3: 記事の特徴ベクトルを求める処理フロー

の文章との類似度によって分類を行なわせることにした。

記事間の類似度は各記事ベクトルの絶対値を 1 に正規化してから内積を求ることで得られる。例えば分野 1 の典型的な文章の記事ベクトルを C_1 、分野 2 の典型的な文章の記事ベクトルを C_2 とすると、記事 u がどちらの分野に属するかを判定するには、記事 u の記事ベクトル A_u と C_i との内積を取る。 $C_1 \cdot A_u > C_2 \cdot A_u$ なら記事 u は分野 1 に分類され、 $C_1 \cdot A_u < C_2 \cdot A_u$ なら分野 2 に分類される。

この概念識別子出現頻度分布は理想的には全概念識別子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 個以下の概念識別子に制限して実験した。この概念識別子の選出方法は、単純に出現頻度の高いものから順番に選出した。

2.2 分類手順

本手法による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4である。

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. このデータから概念識別子を抽出。
3. 抽出された概念識別子の中から、特徴ベクトルを生成する際に使用する概念識別子を選出。
4. 大量の新聞記事データから特徴ベクトル生成。
5. 人手で新聞記事から各分野の典型的な記事を3つずつ選び出し、各分野の特徴ベクトルを生成。(分野基準ベクトル)
6. 分類したい記事について、その記事中の概念識別子の特徴ベクトルから記事の特徴ベクトルを計算。(記事ベクトル)
7. 記事ベクトルと分野基準ベクトルとを比較してその記事が属している分野を決定。これは記事ベクトル、分野基準ベクトルとともに絶対値を1に正規化してから両ベクトル間の内積を計算し、内積が最大になる分野基準ベクトルに対応する分野を、その記事の分野とみなす。

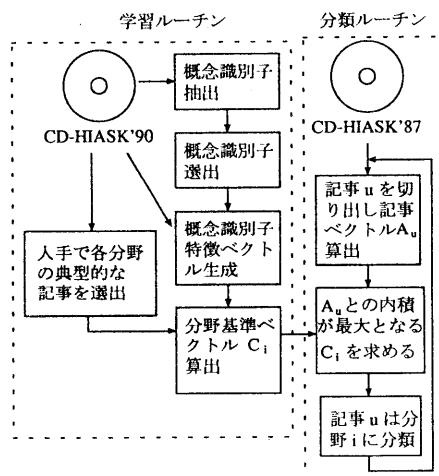


図4: 本手法の処理フロー

3 実験

本手法の有効性を確認するために、人手による分類と本手法による分類とを比較する以下の2つの実験を行なった。

• 実験1

朝日新聞1987年の記事中で人間にとて分類が易しい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は以下の通り。

1. CD-HIASK(朝日新聞のCD-ROM)1990年版 [4](約150Mバイト、101966記事)
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 [5]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。これは以下のようなものである。

債券相場が急落 東京市場

東京債券市場は18日午前、取引開始時から売り一色の展開となり、現物、先物相場とも急落した。東京証券取引所に上場されている現物国債の指標銘柄、89回債は売り物に押され、2円49銭安の114円18銭に下がり、利回りは一時3.10%まで上昇した。89回債の売買途中の値下がり幅としては最大…

3. EDR電子化辞書

単語データの抽出にEDR電子化辞書の日本語単語辞書[6]と、専門用語辞書(情報処理)[7]の日本語専門用語辞書を使用した。辞書の登録語数は基本語約23万語、専門用語約12万語である。辞書の情報のうち(単語、概念識別子、頻度情報)の三組を使用した。頻度情報はEDRコーパス内のその(単語、概念識別子)

子) の頻度であり、全(単語、概念識別子)の約 20%にこの頻度情報が付加されている。残りの約 80%の頻度情報は 0 になっている。

4. 単語データ

EDR 電子化辞書の日本語単語辞書と日本語専門用語辞書中の全単語を使用した。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)を用了。ただし誤抽出が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 年 1 月 1 日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。

5. 概念識別子データ

EDR 電子化辞書の日本語単語辞書と日本語専門用語辞書中の全単語について、関連する概念識別子を調査し、使用頻度の高いものを採用した。各単語に関連する概念識別子を全て使用する場合と、一つだけ使用する場合との二通りで実験を行なった。

6. 特徴ベクトル

特徴ベクトルの次元数は 4096, 2048, 1024, 512, 256, 128, 64 の 7 種類で実験した。

この次元数個分の概念識別子の選出の方法であるが、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。

3.1 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。そのうち 20 記事については比較的簡単に分類できるものを選び、残りの 20 記事について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。

3.2 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 5 分野に分類した。どうしても一つの分野にしぶり切れない記事については 2 つの分野に分類することを許した。2 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。

4 結果

4.1 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 名の人によって、4 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 のようになつた。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。

4.2 本手法による分類結果

各方法による分類結果を表 2~表 4 に示す。

表 2

特徴ベクトルの次元数を 2048 とした時の実験 1 の人手による分類で最も多く選ばれた分野を正解とした場合の正解率。

- 易：分類が易しい 20 記事での正解率。

表 1: 人手による分類結果

	1	2	3	4	5	6	7	8	9	10
政	0	20	0	0	18	0	1	0	0	20
経	20	0	20	0	2	0	0	20	0	0
事	0	0	0	0	0	20	0	0	20	0
国	0	0	0	20	0	0	19	0	0	0
	11	12	13	14	15	16	17	18	19	20
政	0	0	1	0	19	0	1	20	0	0
経	20	0	0	0	1	0	19	0	0	0
事	0	20	0	0	0	20	0	0	20	0
国	0	0	19	20	0	0	0	0	0	20
	21	22	23	24	25	26	27	28	29	30
政	16	0	16	10	18	2	1	0	2	0
経	1	0	4	1	0	0	11	17	17	0
事	3	20	0	0	0	0	0	1	0	20
国	0	0	0	9	2	18	8	2	1	0
	31	32	33	34	35	36	37	38	39	40
政	0	18	0	0	15	3	14	0	0	0
経	0	2	17	0	3	0	5	0	13	17
事	20	0	1	3	0	17	0	20	0	0
国	0	0	2	17	2	0	1	0	7	3

- 難：分類が難しい 20 記事での正解率。
- 全体：実験 1 に使用した 40 記事全体での正解率。

表 3

特徴ベクトルの次元数を 2048 とした時の実験 2 での正解率。

- 易：分類選択時的一位候補の内積と二位候補の内積の比が大きい 200 記事での正解率。
- 難：「易」以外の 200 記事での正解率。
- 全体：実験 2 に使用した 400 記事全体での正解率。

表 4

実験 2 で、特徴ベクトルの次元数を変化させた時の各方法での正解率。「易」「難」「全体」は表 3 と同じ。

なお、参考までに「3 実験」で提示した記事例（見出しは「債券相場が急落 東京市場」）の特徴ベクトルと各分野基準ベクトルとの内積の値を以下に示す。

政治 0.874, 経済 0.949, 国際 0.869,
社会（犯罪、事件）0.888, 社会（教育、人間）0.884
従ってこの記事は「経済」に分類される。

5 考察

以上の結果より、以下のことがわかった。

表 2: 次元数 2048 での実験 1 の結果 [%]

使用概念識別子	易	難	全体
人手による分類	98.5	82.75	90.6
方法 1(全て)	95.0	50.0	72.5
方法 2(一つ)	90.0	50.0	70.0

表 3: 次元数 2048 での実験 2 の結果 [%]

使用概念識別子	易	難	全体
方法 1(全て)	98.0	73.5	85.8
方法 2(一つ)	98.0	71.0	84.5

• 分類が簡単な記事について

表 2 より、実験 1において、特徴ベクトルの次元数を 2048 とすると、分類が簡単な記事については 90% 以上の正解率が得られた。表 3 から、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。この時、易しいと判定した記事の分類正解率は 98% となり、人手による易しい記事の分類正解率の 98.5% とほぼ同じ値が得られた。したがって、簡単な記事は自動分類し、難しい記事は人に分類候補を選択させるという処理を行なわせることができると考えられる。

• 分類が難しい記事について

表 2 より、実験 1において、分類が難しい記事について特徴ベクトルの次元数を 2048 とすると、50% の正解率しか得られなかった。これに対し人間の分類の正解率は 82.75% だった。分類が難しい記事は表面を眺めるだけでなく、内容をじっくり読まないとどの分野に分類すべきかわからない記事が多い。しかし、本手法では文の構造に依存せず、その文中に現れる単語の概念識別子だけをもとに分類を行なっている。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。

• 概念識別子の用い方について

表 4: 次元数を変化させた時の実験 2 の結果 [%]

次元	方法 1(全て使用)			方法 2(一つ使用)		
	易	難	全体	易	難	全体
4096	98.5	69.5	84.0	98.0	70.0	84.0
2048	98.0	73.5	85.8	98.0	71.0	84.5
1024	97.0	65.0	81.0	97.0	61.0	79.0
512	97.0	62.0	79.5	95.0	59.0	77.0
256	90.5	63.0	76.8	94.0	59.0	76.5
128	86.0	60.0	73.0	88.5	56.5	72.5
64	86.5	51.5	69.0	86.5	53.0	69.8

表 4 の「全体」に注目すると、各単語に関連している全ての概念識別子を用いた方が、一つの概念識別子だけを用いるより、やや分類正解率が高いことがわかる。

頻度情報が記述されている概念識別子の割合が少ない(約 20%)ことと、単語辞書に最初に記述されている概念識別子が必ずしもその単語を代表する概念識別子ではないことが、一つの概念識別子だけを用いた場合に正解率が低かった原因ではないかと考えられる。

しかし、正解率の差は小さく、今回の実験データでは、たまたま全ての概念識別子を用いた方が正解率が高かったということも考えられる。

6 終わりに

本稿では、大量のデータから得られる概念識別子の出現頻度を利用して、概念識別子の特徴ベクトルを自動的に生成し、この概念識別子の特徴ベクトルから文書の特徴ベクトルを生成することで、文書を自動的に分類することができる手法について説明した。そして、朝日新聞の 1 年分の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なったところ、人手で容易に分類が可能な文書については特徴ベクトルの次元数を 2048 とすると 90% 以上正確に分類することができることが確認できた。また人手で分類が困難な文書については 50%~70% 程度の正解率しか得られないことがわかった。

分類が易しい記事を人手で分類する時の分類正解率は実験 1 より 98.5% という値が得られているが、実験 2 のベクトルの次元数 2048 の場合の易しい記事の分類の正解率は 98% であり、易しい記事に関し

ては人手とほぼ同等の精度で分類できることが確認できた。

今後の課題としては、各単語に対応している複数の概念識別子から最適な概念識別子を選択する方法の開発、より良い評価法の開発、朝日新聞以外の文書データでも本手法が有効かどうかの検証、より分類精度を高めるための手法の開発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致します。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致します。

参考文献

- [1] 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, 情報処理学会第 46 回(平成 5 年前) 全国大会講演論文集 5G-1, 分冊 4, pp.187-188 (1993).
- [2] 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, 情報処理学会自然言語処理研究会研究報告, 93-NL-98, pp.81-88 (1993).
- [3] EDR 電子化辞書 ⓒ株式会社日本電子化辞書研究所.
- [4] 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
- [5] 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
- [6] EDR 電子化辞書 日本語単語辞書 ⓒ株式会社日本電子化辞書研究所.
- [7] EDR 電子化辞書 専門用語辞書(情報処理) ⓒ株式会社日本電子化辞書研究所.