# キーワードの統計分布情報を用いた文書の多重分類システム

3N－1

森田貴子　松田芳樹　橋本哲也　東野純一
(株) 日立製作所 システム開発研究所

## 1. はじめに

計算機ネットワークを介して情報を収集する環境の普及と、通信速度の高速化にともない、今後さらに情報が氾濫して行くと予想される。そこで、収集した情報を有効活用するために、文書情報を自動的に分類、整理するシステムに関する研究を行ってきた。文書の自己組織的クラスタリング手法はこれまでにも多く提案されているが[1, 2, 3]、本研究ではパーソナルユースを前提とした多重分類システムの開発を目的とする。

文書情報は、(1) 時々刻々変化する、(2) 時、場合、受け手ごとに内容の受け取り方が異なる、という性質がある。そこで本システムでは、「一文書を多面的（多重）に分類し、分類結果を階層的に整理すること」を基本方針とした。本稿では、多重分類手法による新聞記事の自動分類実験について報告する。

## 2. 多重分類手法

多重分類手法は次に示す二つの分類処理の組合せによって実現する（図1参照）。

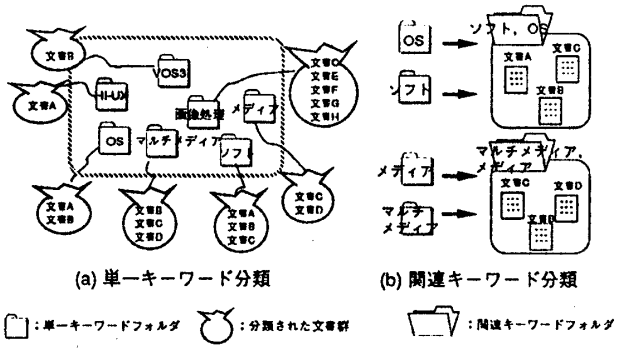


図 1: 多重分類を実現する二つのキーワード分類

### 2.1 単一キーワード分類

情報は多面的な意味を持ち、例えば、「総理大臣が訪米し、経済問題について米大統領と会談した」という文書は、「政治」、「国際」、「経済」のいずれの分類にも解釈できる。そこで、情報の多面性を考慮し一文書を複数の分類カテゴリに多重分類する。そのために、まず分類対象の文書群を各キーワードを含む文書ごとにまとめる。これを単一キーワード分類と呼ぶ。

### 2.2 関連キーワード分類方式

単一キーワード分類だけでは分類カテゴリ内に関連度の低い文書が含まれる可能性がある。そのため共通の文書を多く含む単一キーワードフォルダの組を統合する処理を次に行う。これを関連キーワード分類と呼ぶ。関連キーワード分類を $n$ 回繰り返すと、$2^n$ 個のキーワードからなる分類カテゴリが生成される。すなわち、統合されたキーワード群のうち一つ以上のキーワードを含む文書がまとめられた分類カテゴリが生成される。

### 2.3 分類カテゴリの階層化

関連キーワード分類によって生成された分類カテゴリ内の文書群について、上述した二つの分類を再帰的に適用し、分類カテゴリを階層的に整理する。

## 3. キーワードの統計分布情報の利用

本手法では生成される分類カテゴリ内の文書の類似性を確保するため、キーワードの統計分布情報を用いた次の処理を施す。

### 3.1 フォルダの統合可否判断

関連キーワード分類を繰り返し行いフォルダの統合をし過ぎて、分類カテゴリ内の文書の関連性が弱くなるのを防ぐため、次の統合可否判断を行う。

まず、フォルダ内の文書は「キーワードの出現頻度とキーワードの重要度の積の並び（ワードベクトルと呼ぶ）$\mathbf{W_i}$」として表現する。重要度には「文献から重要語を抽出するためのカイ２乗式」[4]を用いた。

$$\mathbf{W_i} = (F_1V_1, F_2V_2, \cdots, F_jV_j, \cdots, F_pV_p) \qquad (1)$$

ただし、i は文書番号、$j$ はキーワード番号、$1 \leq j \leq p$（$p$ はキーワード数）、$F_j$ はキーワード $j$ の出現頻度、$V_j$ はキーワード $j$ の重要度。

フォルダ内の全文書のワードベクトルの平均 $\mathbf{W_a}$ を求め、各文書のワードベクトル $\mathbf{W_i}$（$1 \leq \mathbf{i} \leq$ 文書数 $n$）との距離を計算する。ワードベクトル間の距離は文書の類似度を表すもので、(2) 式で定義する。

$$d(D_i, D_j) = 1 - \frac{\mathbf{W_i} \bullet \mathbf{W_j}}{|\mathbf{W_i}||\mathbf{W_j}|} = 1 - \cos\theta \qquad (2)$$

ただし、$\bullet$ は内積、$|\mathbf{W_i}|$ は $\mathbf{W_i}$ の大きさ、文書 $D_i$、$D_j$ のワードベクトルを $\mathbf{W_i}$、$\mathbf{W_j}$、$\mathbf{W_i}$ と $\mathbf{W_j}$ のなす角度を $\theta$、$D_i$、$D_j$ 間の距離を $d(D_i, D_j)$ とする。

さらに、すべての「平均ベクトル $\mathbf{W_a}$ と各文書との距離」から平均距離 $d$ と分散 $\sigma$ を求める。

---

Hierarchical Text Categorization System using Statistics of Keywords.
Takako MORITA, Yoshiki MATSUDA
Tetsuya HASHIMOTO and Junichi HIGASHINO
Systems Development Laboratory, Hitachi, Ltd.

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} \quad (3), \qquad \sigma = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1} \quad (4)$$

ただし、$1 \leq i \leq n$ ($i$ は文書番号、$n$ は文書数)、$d_i$ は文書番号 $i$ の文書と平均ベクトルとの距離。

統合前の二つのフォルダ $A$、$B$ の平均距離 $\bar{d}_A$、$\bar{d}_B$ および分散 $\sigma_A$、$\sigma_B$ から、平均距離の平均値 $\bar{d}_{AB}$ と、分散の平均値 $\sigma_{AB}$ を求め、統合後のフォルダの平均距離 $\bar{d}_{merge}$、分散 $\sigma_{merge}$ と比較して、(5) 式または (6) 式の条件が満たされるとき統合不可と判断する。

$$|\bar{d}_{merge} - \bar{d}_{AB}| \geq T_d \quad (5), \qquad \frac{\sigma_{merge}}{\sigma_{AB}} \geq T_\sigma \quad (6)$$

ただし、$|d|$ は $d$ の絶対値、$T_d$、$T_\sigma$ はしきい値。

### 3.2 ノイズ文書の除去

分類カテゴリには関連性の低い文書がノイズとして含まれている可能性がある。そこで分類カテゴリ内の各文書のワードベクトル $\mathbf{W_i}$ と平均ベクトル $\mathbf{W_a}$ との距離を計算し、しきい値 $T_d$ 以上の文書はフォルダから除去する。これによってノイズを削減する。

## 4. 新聞記事の多重分類実験

本手法の有効性を検証するために、新聞記事に対する多重分類実験を試みた。なお、本実験は HI-UX/WE2 を OS とする日立 3050RX/340 (128MB, 131MIPS) 上で行った。

### 4.1 実験内容

実験は日経産業新聞紙面の一日分のサンプル記事の中から、「産業」関連の 9 分野に属する 96 件の記事を用いた。各記事から文字種の情報だけを利用してキーワードを抽出した。記事の平均文字数は約 500 文字、抽出されたキーワード数は総計 6,644 個、平均 69 個 / 記事であった。分類実験には 2 件以上の記事で出現した 606 個のキーワードを使用した。

### 4.2 実験結果

本実験では第 1 階層に 199 個の分類カテゴリが生成された。記事数が 3 件以下と少ない分類カテゴリは無効として残りの 103 個を評価の対象とした。

各分類カテゴリについて予め記事に付与された 9 分野それぞれを正解とみなしたときの適合率 (precision) を求め、「フォルダの統合可否判断、ノイズ文書の除去」処理の有無による影響を評価した。なお、処理有りのときには統合不可と判断される場合があるため、処理無しのときよりも最終的に生成されるフォルダ数が多くなっている。図 2に示すように処理有りの場合に適合率が向上することから、キーワードの統計処理が有効に作用することが確認できる。

分類カテゴリ 103 個の中から適合率の高かった 9 個の例を図 3に示す。平均すると一つの記事が 11 個の分類カテゴリに多重に分類され、予め付与された 9 分野の小分類に相当するような詳細な分類カテゴリが生成された。

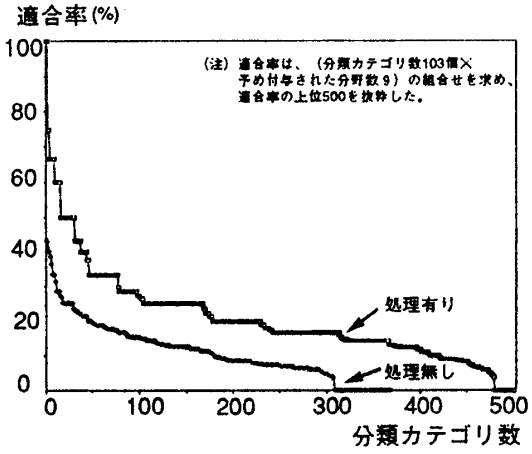


図 2: 多重分類によって生成された分類カテゴリの適合率

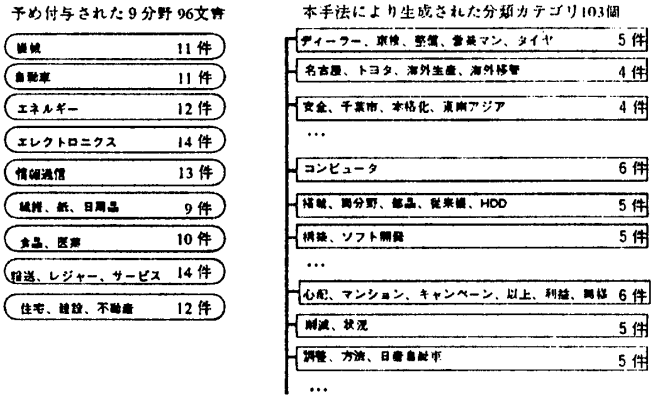


図 3: 「産業」関連の 9 分野と多重分類手法で生成された分類カテゴリの例

## 5. おわりに

キーワードの統計分布情報を用いた文書の多重分類手法による新聞記事の分類実験を行った結果、適合率の向上と、記事の話題を反映した詳細な分類カテゴリの生成が確認できた。今後は、階層的な分類カテゴリを生成する場合について評価を行い、多重分類システムの構築を目指す。

## 参考文献

[1] 豊浦：自己組織型ニューラルネットワークによるドキュメントの自動分類、情処 NL 研資料 88-6、pp.41-48、1992

[2] 内田：ART を利用した多義語の分類と評価、情処 NL 研資料 101-15、pp.113-120、1994

[3] 湯浅：情報のブロードキャッチシステム、情処研資料 IM13-6、GW-4、pp.37-44、1993

[4] 海野: 出現頻度情報に基づく単語重みづけの原理、Library and Information Science No.26、1988