NDC の階層構造を利用した図書の自動分類の試み

宮田洋輔（慶應義塾大学大学院）　miyayo@slis.keio.ac.jp
石田栄美（駿河台大学）　emi@surugadai.ac.jp
神門典子（国立情報学研究所）　kando@nii.ac.jp
上田修一（慶應義塾大学）　ueda@slis.keio.ac.jp

抄録　これまでの図書の自動分類実験から得られた機械学習や統計分析などの分類手法，書名，目次，帯，著者名，出版社名などの特徴素を考慮しつつ，分類できるカテゴリ数が多く，付与される分類記号に偏りのある NDC を対象として，自動分類実験を行った。二段階で分類を行う方法を提案し，第一段階では，いくつかの分類手法で分類した結果を組み合わせて分類を行い，さらに第二段階（類の中での分類）では，相対出現率で分類した結果，一度で全体を自動分類した結果に近い分類性能を得た。NDCの類への分類である第一段階の再現率を比較すると，提案手法による分類のほうが大きく上回っていた。類の中での分類にも組み合わせ手法を用いるなら，さらに分類性能の向上が見込まれる。

1. はじめに

自動分類の応用を考慮し，分類作業の実務で用いられているデータを使用して，自動分類の性能を向上させるための方法の提案と実験，検討を行った。

対象とするのは，日本語の図書であり，用いる分類法は日本十進分類法（NDC）である。

2. 本研究の課題

図書に付与された NDC の分類記号には，分類先となる分類記号（カテゴリ）の種類が多く，さらに分類記号間の付与数の偏りが大きいという特色がある。今回使用したデータでは，約 6000 もカテゴリがあり，NDC では，3 類に分類される例が全体の 3 割を占めている。（表1，表 2 参照）

一般に自動分類で使用されるテストコレクションは，英語であって，カテゴリ数は，NDC に比べて格段に少なく，偏りも小さいので，ここで用いる目録のデータベースは，これとはかなり異なったものである。

一方，現在では，これまでは入手し辛かった目次や帯のデータを収録したデータベースも登場し，図書を自動的に分類するための手がかりとして，書名，目次，帯，著者名，出版社名などの要素を用いることができるようになった。

これまでに，サポートベクターマシン（SVM），相互情報量，相対出現率の三種の手法を用いて自動分類実験を行った結果から，書名，目次，帯の持つ特色が明らかになり，さらに NDC の類ごとにも特徴があることが判明した [1)]。

そこで，一度に全体を分類するのではなく，段階に分け，適切な特徴素と手法を組み合わせていく方法が有望と考えられる。1)。

全体を一度に分類した場合，自動分類の性能の向上は，学習集合を極めて大きくするか，全く新しい手法を導入する以外には困難である。しかし，段階に分ける法では，細かな工夫の余地が生まれる。

そこで，図1に示すように全体を一度に分類する方法に対し，最初に類（10 区分）に分け，類ごとに自動分類を行う方法を試みた。

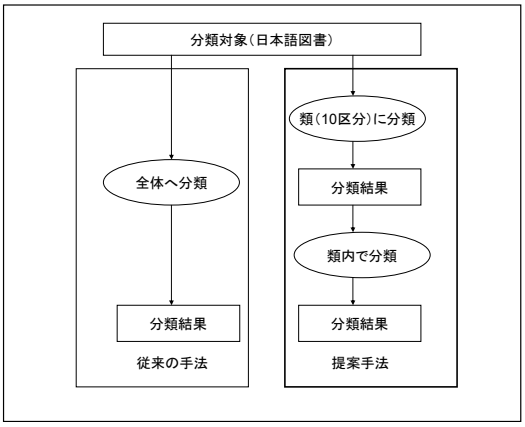

図 1　本研究のアプローチ

表＊ 分類に用いる手法･要素

二段				flat	
第一次区分		全桁			
相互情報量	書名	相互情報量	書名	相互情報量	書名
SVM	目次	相対出現率	目次	相対出現率	目次
	帯	SVM	帯	SVM	帯
	著者名				著者名
	出版社名				出版社名

3. 実験用データ

分類実験には，「BOOK」データベースとNACSIS 書誌データベースを統合した実験用のデータを用いた。

3.1 「BOOK」データベース

「BOOK」データベースは，日外アソシエーツなどにより構築されている書誌データベースである。実験には，1999 年度版の 51,171 件と 2000 年度版の 53,707 件を用いた。「BOOK」データベースの特徴は，目録規則上は採られない目次や帯の内容情報を収録している点にある。

3.2 NACSIS データベース

「BOOK」データベースには付与されていない NDC のデータを得るため，NACSIS-CAT 書誌データベースを用いた。NACSIS-CAT 書誌データベースには，国立情報学研究所の NACSIS-CAT に 1990 年から 2000 年の間に入力されたレコードが収録されている。

3.3 実験用データ

上記の 2 つのデータベースを統合し，実験に用いたデータ集合を作成した。

①「BOOK」データベースから書名，目次，帯，著者名，出版者名と ISBN を抜き出す。

②NACSIS-CAT から NDC と ISBN を抜き出す。

③ISBN をキーに①と②を統合する。

以上のように，書名，目次，帯，著者名，出版者名と分類記号のデータを持った 39,580 件のレコードからなる実験用データを作成した。実験用データの基礎データを表＊に示す。表＊は実験用データの第一次区分ごとの文献数を示している。

これらの表からも，先述した NDC の，カテゴリの種類が多く，さらに分類記号間の付与数の偏りが大きいという特色が明らかになっている。

このデータを 5 分割して，31,664 件を語とカテゴリとの関係を学習するための学習用集合，残りの 7916 件をシステムが分類を付与する評価用集合として用いた。

表＊ 実験用データの基礎データ

異なりカテゴリ数	5,884	最高出現頻度	512
延べカテゴリ数	41,348	出現頻度1	2,344
平均出現頻度	7	出現頻度1の割合	39.8%

表＊ 実験用データの第一次区分内訳

第一次区分	件数	割合
0 類 総記	2,483	6.0%
1 類 哲学	2,771	6.7%
2 類 歴史	3,618	8.8%
3 類 社会	13,395	32.4%
4 類 自然	5,206	12.6%
5 類 技術	4,962	12.0%
6 類 産業	2,186	5.3%
7 類 芸術	3,284	7.9%
8 類 言語	1,559	3.8%
9 類 文学	1,884	4.6%
合計	41,348	100.0%

4. 分類実験概要

4.1 特徴素

分類を行うための手がかりには，書名，目次，帯，著者名，出版者名をそれぞれ用いた。

4.2 語の切り出し

各特徴素から語を切り出すには，奈良先端科学技術大学院大学の形態素解析システム「茶筌」[茶筌]を用いた。茶筌に切り出された語の中から名詞あるいは未知語と判定された語を用いた。

4.3 統計的手法

語の重み付けには，相対出現率と相互情報量(MI)の 2 手法を用いる。

相対出現率は，全出現回数に対する相対的な出現回数を用いて決定される。カテゴリ C_j での語 t_i の重み w_{ij} は以下の式で算出される。

$$w_{ij} = \frac{f_{ij}}{\sum_i f_{ij}}$$

相互情報量は語とカテゴリが共起するときの情報量を重みに用いる手法である。語 t_i の重みは以下の式で算出される。

$$w_{ij} = \log \frac{N n_{ij}}{n_i n_j}$$

ここで N は集合中の全文献数，n_i は語 t_i の出現する文献数，n_j はカテゴリ C_j の文献数， n_{ij} はカテゴリ C_j の語 t_i の出現する文献数である。

相対出現率と相互情報量の 2 つの重み付けによって表現される各カテゴリのベクトルは，評価用

データのベクトルと類似度を計算し，その類似度が最も高いカテゴリに評価用データは分類される。

類似度は以下のように求めた。カテゴリ C_j における語 t_i の重みを w_{ij} とすると，カテゴリのベクトル x_j は以下のように表される。

$$x_j = (w_{1j}, w_{2j}, \cdots, w_{ij})$$

また語 i の出現確率 S_i を用いて，分類対象のベクトル u は以下のように表される。

$$u = (S_1, S_2, \cdots, S_i)$$

対象文献とカテゴリとの類似度 sim(d,q)は以下の公式から算出される。

$$sim(d,q) = \sum_{j=1}^{M} S_i \cdot w_{ij}$$

4.2 機械学習手法

Joachims が提供している SVM ソフトウェア SVMlight [SVM)]を用いて実験を行った。カーネル関数には，線形カーネル関数を用いた。入力ベクトルは各分類カテゴリに対し，単語の出現回数を重みとして与えた。

5. 分類実験結果

5.1 評価尺度

システムの評価には，NACSIS-CAT 書誌データベースで付与された分類を正解として，f 値を用いて評価を行う。再現率 R は以下の式で算出される。

$$\text{再現率}\ (R) = \frac{\text{システムが分類した正解の総数}}{\text{正しい分類の総数}}$$

また，同様に以下の式で，精度Pも算出される。

$$\text{精度}\ (P) = \frac{\text{システムが分類した正解の総数}}{\text{システムの分類の総数}}$$

計算された精度，再現率を用いて，fcj は以下の式で算出される。

$$F_{1,C_j} = \frac{(1+1)P_{C_j}R_{C_j}}{P_{C_j} + R_{C_j}}$$

5.2 従来手法による分類結果

従来の手法での分類結果を表＊に示す。特徴素に書名，重み付け手法に相対出現率を用いた場合にもっとも良い結果が得られた。この結果をベースラインとして，提案手法との比較を行う。

表＊ 従来手法での分類結果(f 値)

	書名	目次	帯	著者名	出版者名
相対出現率	16.7%	13.4%	12.1%	6.4%	2.6%
相互情報量	19.9%	17.3%	16.1%	8.3%	1.8%
SVM					

5.3 二段式分類による分類結果

本研究で提案する手法では，第一次区分の 10 区分に分類したあとに各特徴素による分類結果を組み合わせ，その結果をもとに残りの全桁への部縫いを行う。

分類結果の組み合わせには，以下の 4 手法を重み k(l,m)として用いて組み合わせた。ここで用いた精度は第一次区分の分類を 5 交差検定行った平均値である(表＊)。

(1) $k_{C_j}(l,m) = 1$

(2) $k_{C_j}(l,m) = \frac{1}{M}\sum_{j=1}^{M} P_{c_j}(l,m)$

(3) $k_{C_j}(l,m) = P_{c_j}(l,m)$

(4) $k_{C_j}(l,m) = \frac{1}{M}\sum_{j=1}^{M} P_{c_j}(l,m) \times P_{c_j}(l,m)$

これらの重みを用いて分類結果を組み合わせ，以下の式から算出される $score_q(c_j)$ のもっとも高いカテゴリに分類される。

表＊ 各手法での精度

		0類	1類	2類	3類	4類	5類	6類	7類	8類	9類	全体
MI	書名	54.4%	52.4%	52.4%	89.5%	77.8%	76.8%	39.2%	66.6%	52.8%	47.5%	60.9%
	目次	60.2%	49.3%	50.7%	88.6%	81.1%	85.8%	36.5%	70.0%	67.4%	45.5%	63.5%
	帯	55.8%	54.9%	52.5%	87.8%	74.9%	73.5%	44.0%	70.1%	58.0%	47.5%	61.9%
	著者名	22.3%	25.7%	30.3%	64.9%	30.6%	33.1%	13.8%	28.6%	15.4%	13.8%	27.9%
	出版者名	43.6%	29.3%	31.7%	76.8%	58.7%	59.9%	23.1%	35.4%	36.1%	17.2%	41.2%
SVM	書名	87.0%	81.3%	80.0%	87.1%	86.3%	83.8%	82.0%	85.7%	84.6%	86.1%	84.4%
	目次	84.3%	83.7%	78.6%	88.8%	88.3%	86.7%	86.1%	88.4%	89.1%	87.9%	86.2%
	帯	84.8%	82.0%	77.7%	86.1%	86.4%	84.7%	84.3%	85.6%	88.2%	87.6%	84.7%
	著者名	75.5%	77.6%	79.5%	82.3%	85.9%	77.4%	69.9%	85.8%	89.2%	77.5%	80.1%
	出版者名	60.0%	66.0%	59.5%	71.6%	79.0%	68.4%	67.1%	70.2%	73.3%	73.3%	68.8%

$$score_q(c_j)=\sum_{l=1}^{P}\sum_{m=1}^{Q}k_{C_j}(l,m)R_{C_j}(l,m)$$

$$R_{C_j}(l,m)=\begin{cases}1 & if\ \ q\ \ is\ \ classified\ \ category\ \ C_j\\ 0 & otherwise\end{cases}$$

5.3.1 一桁目の分類結果

相互情報量と SVM での各類への分類結果を表＊に示す。各類への分類では目次を特徴素として用いて相互情報量で重み付けを行ったときにもっとも良い結果を示している。

第一次区分での分類結果を各組み合わせ手法によって組み合わせ結果の f 値を表＊に示す。この結果がもっとも良かった手法(3)を第一次区分での分類結果として採用した。

表＊ それぞれの特徴素を用いた f 値(類)

	書名	目次	帯	著者名	出版者名
MI	62.4%	63.6%	63.8%	26.7%	40.3%
SVM	59.4%	54.0%	56.6%	28.1%	36.2%

表＊ 各組み合わせ手法の f 値

(1)	(2)	(3)	(4)
71.0%	72.1%	72.4%	72.2%

5.3 全桁への分類結果

5.2 で得られた組み合わせ結果をもとに全桁への分類実験を行った。書名，目次，帯を特徴素として用いた時に得られた f 値を表＊に示す。書名を特徴素として用いた場合の相対出現率，flat がもっとも良い結果を示した。

表＊ 各手法での f 値

	相対出現率		相互情報量	
	2 段	flat	2 段	flat
書名	16.5%	16.7%	17.9%	19.9%
目次	12.3%	13.4%	16.8%	17.3%
帯	11.1%	12.1%	14.2%	16.1%

表＊ 各手法での正解数

	相対出現率		相互情報量	
	2 段	flat	2 段	flat
書名	2534	2665	2043	2277
目次	2196	2393	1955	1943
帯	2014	2172	1649	1835

6. 今後の考察

本研究では，図書分類の，偏りが大きくまたカテゴリ数が極端に多いという特徴を鑑み，各種特徴素を用いて十分な学習用文献数が確保できる第一次区分での分類を行ったのちに，各類内で残りの全桁への分類を行うという 2 段階での分類手法を提案し，従来の手法との比較を行った。

実験の結果から，第一次区分での結果は，各特徴素と各手法を，分類結果の精度を用いて組み合わせることで 7 割以上の f 値が得られ，高い結果を示した。また全桁への分類でも，従来の手法と大差のない結果を示すことができた。

しかし，一桁目での分類では，従来手法が 6 割程度の再現率であったのに対して，提案手法では 7 割以上の再現率を示すことができた。このことは，提案手法による分類はより近い主題分野に分類されていたと考えることができる。今後はこの時点での差を生かせるような手法を考案する必要がある。

引用文献

1)石田栄美，宮田洋輔，神門典子，上田修一目次と帯を用いた図書の自動分類
情報処理学会情報学基礎研究会発表要綱．2006.(2006-03-22)
茶筌）奈良先端科学技術大学院大学松本研究室．茶筌．[2006-04-24],
〈http://chasen.naist.jp/hiki/ChaSen/〉
SVM ）SVM-Light Support Vector Machine．[2006-04-24],< http://svmlight.joachims.org/>