社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－FI－82（12）
2006－DD－54（12）
2006／3／22

目次と帯を用いた図書の自動分類

石田栄美*　　宮田洋輔**　　神門典子***　　上田修一**

*駿河台大学文化情報学部
E-mail:emi@surugadai.ac.jp
**慶應義塾大学文学部
***国立情報学研究所

書名だけでなく目次や帯情報を用いて，図書を日本十進分類法の分類カテゴリに自動分類する実験を行った。分類手法には，相対出現率と相互情報量にもとづく重み付けと Support Vector Machine(SVM)による手法を用いた。「BOOK」データベースと NII-CAT データを統合した 24,000 件を用いて学習させたところ，重み付けに相対出現率と相互情報量を用いた場合は，書名に加えて目次と帯情報を用いた場合の効果が認められた。また，機械学習手法よりも，統計的手法のほうが有効であった。分類カテゴリの分野ごとの再現率を調べたところ，分野によって再現率に大きな差があった。さらに，帯情報などが有効な分野もあり，書名，目次，帯を用いた効果が分野ごとに異なることが明らかになった。

Text Categorization using Title and a Table of Contents

Emi ISHIDA*　Yosuke MIYATA**　　Noriko KANDO***　　Shuichi UEDA**

*Surugadai University
E-mail:emi@surugadai.ac.jp
**Keio University
*** National Institute of Informatics

In this paper, we describe methods of classifying Japan MARC records to class number of Nippon Decimal Classification. We compare the performance of three categorization method, based on mutual information(MI), relative frequency and SVM. In each method, training data are title and table of contents and blurb on the flap in Japan MARC records. The experimental results show that the best performance is MI using title and table of contents and blurb on the flap, but other methods are not. In failure analysis, we found the performance depends on subject of class number.

１．はじめに

図書館ばかりでなく書店，オンライン書店，古書店において本の目録データは，基盤となる情報を提供している。目録データの作成と提供は，国立国会図書館や取次，そして NACSIS-CAT などにより行われており，その整備は進んでいる。しかし，主題からのアクセス手段として広く分類がもちいられているものの，膨大な量の分類記号の付与されていない目録レコードがある。例えば，1990 年から 2000 年までに NACSIS-CAT に入力された目録データ(622,295 件)に対し，日本十進分類法（NDC）による分類記号が付与されている割合は 76.6%(476,812 件）でしかない。このような状況となっている大きな原因は，分類記号の付与には，多大な労力を必要とするためである。

目録データベース中に分類記号が付与されていないレコードが大量に存在すれば，分類による主題アクセスが保証されない。そのため，

目録データに対する分類記号の自動付与，もしくは分類記号付与支援システムを開発する実用的な意義が認められる。

本研究では，(1)図書の自動分類に有効な手法の検討，(2)分類の手がかりとして，書名だけでなく目次および帯情報を用いた場合の有効性について検討する。具体的には，「BOOK」データベースの目次，帯情報と NACSIS-CAT の書誌データを用いる。そして分類手法として統計的手法と機械学習手法の比較を行うことにする。

なお，分類カテゴリとしては，日本十進分類法(以下，NDC)で用いられている分類記号を用いた。

2. 本研究のねらい

テキストの自動分類研究は，1990 年後半からさかんに行われるようになってきており，機械学習手法である Support Vector Machine(以下，SVM とする)を用いた手法が有効であるといわれている。Frank と Paynter は SVM を用いて米国議会図書館件名標目表（LCSH）からアメリカ議会図書館分類法（LCC）へのマッピングの実験を行い，その実用性を示しているが [1]，SVM の適用は新聞記事 [2][3]や Web ページ [4]などを対象としたものが多く，図書に対する分類例はあまりみられない。図書の自動分類に関する研究例はいくつかあるが，それらは主に統計的手法がもちいられている [5][6]。本実験では，統計的手法と機械学習手法の分類手法を用いて実験を行い，どちらの手法が図書の分類に有効であるかを検討する。

また，これまでの自動分類実験では，分類の手がかりとして，書名を用いた例が主流であった。これは入手可能な情報が書名であったためもあるが，図書の書名は，“資料の内容を端的に表現しており，副書名はそれを補っている”[7]とされているように分類を行う際の主題の把握の手がかりとして最も適切なものと考えられているからである。しかしながら，書名のつけられ方として，“購買意欲をそそるために誇張したり，奇抜な書名がつけられている通俗書や，形式書名としての文学作品などの書名に注意すべきである”[7]ともいわれているように，必ずしも書名が主題そのものを表しているものばかりではない。一方で，目次は“具体的に資料の内容を表現している”[8]と言われる。本研究では，書名だけでなく，目次や帯情報が図書の分類にとって，どの程度有効であるかも検討する。目次や帯情報を用いた研究例もあまりみられないが，章見出しを用いた例としては，Cheng と Wu の研究がある [9]。Cheng らは数学分野に限定した図書を対象に書名と目次を用いて自動分類実験を行い，自動分類における目次の実用性を示している。

目次や帯情報の利用は，従来まで分類することが困難だと考えられてきた分野にとって，重要な分類の手がかりを提供する可能性もある。NDC は，主題による分類が原則であるが，“文学(9 類)については言語区分の上，文学共通区分という叙述形式の区分によって分類”[10]している。そのため，文学分野の自動分類は書名からだけでは困難と考えられている。例えば『東京育ちの京都観光』という書名の図書は，書名をみただけでは，2 類の「地理･地誌･紀行」に分類されることも適当と思えるが，実際には，9 類「日記，書簡，紀行(915.4)」に分類されているなど主題による分類が行われない場合もある。本研究では，目次や帯情報も用いることができることから，文学分野に対する自動分類も試み，目次や帯情報の効果について検討する。

3 分類実験に用いたデータ

3.1 「BOOK」データベース

日外アソシエーツなどにより構築されている「BOOK」データベースの 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，目録規則上採られない目次や帯の情報を含んでいる点にある。

図 1 は「BOOK」データベースに収録されているデータの例である。「BOOK」データベースはそれぞれのフィールドがタグ付けされている。「＼01＼」のタグの付いた箇所が書名，「＼08＼」のタグが帯の情報，「＼09＼」が目次の情報をそれぞれ示している。

「BOOK」データベースのデータは NDC を付与されていないため，そのまま実験に用いるのではなく，次節の NACSIS-CAT 書誌データベースとの統合を行い，実験用のデータベースを作成した。

3.2 NACSIS-CAT 書誌データベース

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を対象とした。図 2 は NACSIS-CAT 書誌データベースに収録されているレコードの例である。「BOOK」データベース同様，それぞれのフィ

ールドがタグ付けされている。

```
＼00＼Ｂ9900688
＼000Ｙ＼Ｂ9900688
＼01＼小脳―神経科学の基礎と臨床〈7〉
＼011＼小脳
＼011Ｙ＼ショウノウ
＼012＼神経科学の基礎と臨床
＼012Ｙ＼シンケイカガクノキソトリンショウ
                ＜＜中略＞＞
＼10＼ＩＳＢＮ4−89242−161−8
＼100Ｙ＼4892421618
＼A0＼ブレーン出版
                ＜＜中略＞＞
＼Ｉ00Ｙ＼990110
＼08＼小脳に関する解剖，発生，機能，病態，手術が主な内容で，
この分野の最高の執筆陣が総力を結集しここに刊行。
＼09＼プルキンエ細胞シナプスとグルタミン酸シグナル伝達機
構；小脳登上線維シナプス成熟に関与するシグナル伝達系；小脳
の形態形成と機能発現；歩行運動の適応制御と小脳のシナプス可
能性；小脳の働きと脊髄小脳変性症；脊髄小脳変性症の分子遺伝
学；小脳病変の外科的治療
＼700Ｊ＼小脳
                ＜＜後略＞＞
```

図 1「BOOK」データベース収録データ例

```
ID = BA39084771
CTGL = jpn
ISBN = 4892421618
TR = 小脳 / 板倉徹, 前田敏博編著
PBLC = ブレーン出版
SOUCE = ORG
YEARA = 1999
CLSKND = NDC8
CLSKND = NDC9
CLSKND = NDC9
CLSN = 493.73
CLSN = 493.73
CLSN = 491.371
SHTBLKND = BSH
SHTBLKND = BSH
SHTBLKND = BSH
SHTBLKND = NDLSH
SH = 脳
SH = 脳 -- 疾患
SH = 脳神経 -- 疾患
SH = 脳髄
```

図 2　NACSIS-CAT 書誌データベース収録データ例

3.3　実験用データベース

上記の「BOOK」データベースと NACSIS-CAT データベースを以下の手順で統合し，実験に用いたデータベースを作成した。

①「BOOK」データベースから書名，目次，帯のデータと 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。

以上の手順で作成されたデータは，27,000 件であった。その中から 24,000 件を語と分類カテゴリ（分類記号）との関係を学習するための学習用集合, 残りの 3,000 件を評価用集合として用いた。

実験に用いたデータベースの基礎データを表 1 に示す。1 件のレコードに対して，複数の分類カテゴリが付与されている例もある。NDC 第一次区分ごとの件数では，他の類に比べて 3 類「社会」が 3 割以上を占めるかなり偏りのあるデータであった。

表 1　実験用データベースの基礎データ

レコード数	27, 000 件
延べカテゴリ数	28, 257 件
異なりカテゴリ数	5, 088 件

表 2　各分類キーの基礎データ

	書名	目次	帯
平均文字数	19. 4 文字	148. 1 文字	126. 4 文字
延べ語数	143, 996 語	763, 246 語	686, 724 語
異なり語数	18, 862 語	59, 997 語	38, 480 語
平均語数	5. 6 語	45. 6 語	30. 5 語

4　自動分類実験の概要

4.1　自動分類の手順

自動分類は, 学習フェーズと評価フェーズに分けることができる。学習フェーズでは学習用集合中から分類キー（分類の手がかりとなる部分，実際には書名，目次，帯情報)から語を切り出し, 切り出された語を用いて分類システムを学習する。

評価フェーズでは, 評価用データを分類システムに投入し, 実際に評価用データを分類する。

4.2　分類カテゴリ

分類先カテゴリとして，NDC 第 9 版に基づく分類記号を用いた。NDC の分類記号は階層構造になっており，第一階層では 0 類から 9 類まで 10 区分されている。0 類は総記，1 類は哲学，2 類は歴史，3 類は社会科学，4 類は自然科学，5 類は技術，6 類は産業，7 類は芸術，8 類は言語，9 類は文学という主題にわけられている。

自動分類実験では, 分類が困難な 9 類文学を対象外とすることや上位 3 桁で分類を行うなど制限が加えられることが多いが, 本実験では全桁を対象に分類を行った。NDC 全桁への分類を対象にすると，表 1 に示したように 5,000 以上の分類カテゴリとなり, 分類の難度はカテゴリ数に制限を加えるときに比べて大幅に増加することが予想される。しかしながら，実際にそれらの種類の分類記号が用いられていることも考慮し, 本実験では全桁を分類記号として用いることにした。

4.3　分類キー

分類の手がかりとなる分類キーには，書名，目次, 帯をそれぞれ単独で用いた 3 通りとそれ

ぞれを組み合わせた 4 通りの計 7 通りを用いた。

4.4　語の切り出し

各分類キーのテキストからの語の切り出しには奈良先端科学技術大学院大学形態素解析システム「茶筌」[11]を用いた。

また，切り出した単語のうち，名詞と未知語を用いた。これは，切り出された語全てを用いた場合と切り出された語から名詞と未知語を用いた場合を比較した予備実験で，より良い結果が得られたためである。

4.5　分類手法

分類手法は，統計的手法と機械学習手法を比較した。統計的手法では重み付け手法として，相対出現率による重み付けと相互情報量による重み付けの 2 種類を用い，機械学習手法では SVM を用いた。

a　統計的手法

相対出現率による重み付けは，カテゴリと語の相対的な出現回数を語の重みに用いる手法で，カテゴリ C_j における語 t_i の重み w_{ij} は以下の式で算出される。

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

ここで f_{ij} は各カテゴリにおける語 t_i の出現回数である。

相互情報量による重み付けは，語とカテゴリの共起するときの情報量を重みとして用いる手法である。相互情報量では語 t_i の重み w_{ij} 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。

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

b.　SVM

Joachims が提供している SVM ソフトウェア SVM^{light} [12]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　実験結果

5.1　評価尺度

NACSIS-CAT 書誌データベースで付与されている分類記号を正しい分類カテゴリ，正分類とみなし，システムが行った分類が正分類と一致すれば正解として，その評価を行った。

評価は再現率をもとに行う。再現率は，システムの分類がどれだけもとの分類を再現できたかを示している。再現率は以下の公式で算出される。

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

正しい分類の総数とは，NACSIS-CAT 書誌データベースで実際に付与されている分類記号の延べ数のことで，本実験用データには複数の分類記号が付与された文献もあるために，文献数よりも多い件数となる。また，この件数は評価用に用いるデータセットごとに異なる。

5.2　実験結果

表 3 に，書名，目次，帯それぞれを単独で用いた場合と，それぞれを組み合わせた場合の再現率を示す。表中の「全て」は，書名，目次，帯の全てを分類キーとして用いた場合である。評価用データセットを 3,000 件にしたまま，9 交差検定を行ったので，結果はそれらを平均したものである。

最も分類性能が高いのは，重み付けに相互情報量を用い，分類キーとして，全てを用いた場合であり，ついで，分類キーは同様で重み付けに相対出現率を用いたものであった。

しかし，相対出現率を用いた手法では，全ての組み合わせと，書名のみを用いた場合でほぼ同程度の再現率であり，SVM では書名のみのほうが高い再現率を得られている。これらのことから，目次や帯情報を分類キーとして用いることの効果はある程度みられるが，圧倒的な性能の向上に繋がるというわけではなかった。

表 4 に各手法，各分類キーによる NDC 第一次区分ごとの再現率を示した。0 類「総記」は，他分野に比べて高い再現率を示している。特に，相互情報量を用いた手法の場合，目次を組み合わせた場合の再現率が高く，目次が有効であることがわかる。9 類「文学」は，書名だけを用いた場合でもある程度の再現率であることはわかった。また，9 類は，書名と帯情報の組み合わせをみてみると，書名と目次を用いるよりも高い再現率であり，文学は目次よりも帯情報が分類に効果的であるこがわかる。7 類「芸術」，8 類「言語」でも同様のことがいえる。

以上のように，分野ごとに有効な分類キーが異なっていると考えられる。

6 結果の分析

実験の結果から，書誌データに有効な手法，および分類キーとして有効な組み合わせが明らかになったが，それらの理由を探るために，以下では分類結果を分析した。

6.1 各分類キーが結果に与える影響

SVM を除いて，書名のみを分類キーとして用いた場合よりも，書名，目次，帯情報の全てを分類キーとした場合の再現率が上回ったが，目次や帯情報が，具体的にどのような効果をもたらしているか，分析した。

本実験で書名による分類で最も良い結果が得られた相互情報量による重み付けによって分類したもののうち，書名のみを分類キーとして分類した場合に，誤分類された書名を，その書名の特徴によって分類した。ただし，重み付け手法ではカテゴリとの類似度による分類カテゴリのランク付けがおこなえるため，ここでは 10 位までに入った分類カテゴリは正解とした。分析対象は 1,490 件である。先に書誌データの分類で行った誤分類の分析 [6] にしたがっ

表3　各分類手法，分類キーによる再現率

		分類キー						
		書名	目次	帯	書名+目次	書名＋帯	目次＋帯	全て
手法	相対出現率	31.00%	26.50%	25.40%	30.30%	30.80%	30.60%	32.70%
	相互情報量	26.40%	25.80%	22.70%	30.20%	29.30%	31.80%	33.90%
	SVM	10.50%	5.80%	5.60%	7.60%	8.40%	7.90%	9.30%

表4　各分類手法，分類キーによるNDC第一次区分ごとの再現率

		0類	1類	2類	3類	4類	5類	6類	7類	8類	9類
相対出現率	書名	55.1%	26.1%	22.2%	30.4%	30.0%	35.3%	27.3%	33.4%	36.2%	24.6%
	目次	50.6%	18.8%	18.8%	25.7%	27.8%	34.3%	23.9%	27.9%	25.5%	14.9%
	帯	50.2%	17.1%	18.2%	22.7%	24.2%	31.4%	22.5%	32.7%	27.9%	21.8%
	書名+目次	54.4%	23.6%	21.9%	29.1%	30.9%	36.6%	26.8%	34.6%	32.5%	19.3%
	書名+帯	55.2%	23.9%	21.8%	28.7%	29.3%	35.9%	27.8%	39.7%	32.9%	26.0%
	帯+目次	53.5%	22.2%	22.8%	28.4%	30.6%	37.1%	27.0%	38.3%	32.0%	23.3%
	全て	54.9%	25.0%	25.0%	30.7%	32.6%	38.4%	28.7%	41.1%	34.0%	25.2%
相互情報量	書名	40.3%	20.8%	17.5%	27.2%	27.8%	30.1%	23.5%	24.3%	36.3%	19.2%
	目次	47.8%	16.6%	16.1%	27.2%	28.0%	30.6%	23.1%	26.9%	24.3%	13.4%
	帯	40.6%	15.3%	15.9%	21.6%	22.4%	25.8%	18.3%	30.6%	27.3%	16.3%
	書名+目次	52.7%	21.9%	21.0%	31.1%	31.0%	34.6%	27.6%	32.6%	32.8%	16.9%
	書名+帯	48.7%	21.8%	19.8%	28.2%	29.6%	33.3%	25.7%	36.7%	37.2%	21.4%
	帯+目次	53.2%	21.9%	22.3%	31.5%	33.1%	35.8%	28.4%	38.7%	35.4%	22.2%
	全て	55.5%	25.2%	24.5%	33.2%	34.2%	38.3%	30.0%	41.4%	39.1%	23.3%
SVM	書名	33.5%	4.9%	2.2%	11.4%	6.4%	12.9%	10.5%	10.0%	21.6%	3.5%
	目次	24.2%	0.9%	0.9%	6.9%	5.3%	8.3%	5.6%	3.2%	2.9%	0.3%
	帯	25.3%	0.5%	1.7%	5.2%	3.7%	8.9%	3.9%	5.3%	8.2%	1.5%
	書名+目次	29.5%	1.7%	1.1%	8.7%	6.3%	10.8%	7.4%	5.0%	6.8%	0.8%
	書名+帯	33.1%	1.2%	2.1%	8.4%	5.6%	11.6%	7.1%	8.4%	15.1%	2.4%
	帯+目次	32.2%	1.2%	1.5%	8.5%	6.0%	12.0%	7.3%	6.2%	7.0%	1.5%
	全て	35.0%	1.8%	1.6%	10.1%	7.0%	13.3%	8.3%	9.0%	11.4%	1.4%

て分類し，その中で目次や帯を用いることによって分類が成功した件数を表5に示す。表中の「誤分類件数」は，誤って分類された件数で，重複を含めている。誤分類件数の割合は誤分類された文献全体に対する割合である。「正解となった件数(目次と帯)」は，書名以外のいずれかの分類キーを用いたことで正解となった件数であり，割合はそれぞれの誤分類件数に対する割合である。

これらの結果から，目次や帯を用いたことで正解カテゴリに分類されている例が多く，目次や帯の情報は書名だけでは不足していた情報を補う可能性を持っていると考えられる。しかしながら，全ての例がうまくいっているわけではなく，書名では正解していたが，帯や目次情報では誤って分類された例もあった。

表6は，書名のみを分類キーとした場合に分類不能であり，帯情報もしくは書名と帯情報を用いた分類をした場合に正解となった例である。この例では，カテゴリ 930.278(英米文学・作家の個人伝記)に分類されることが正解であるが，「アンダーソン」や「フォークナー」など書名中の語が学習用集合の語にマッチせずに，分類することができなかった。帯情報に含まれる「オハイオ」が，他のシャーウッド･アンダーソンに関する図書の帯にも出現しており正解カテゴリにおいて大きな重みを持っていたために正解となった。この例のように，書名中の語が図書の主題をあらわしていない場合などには，帯や目次の情報が有効であることが考えられる。

カタカナ語やアルファベット，固有名詞を含む書名では，その固有名詞やアルファベットが頻繁にでてこない可能性もあり，分類を行うための情報が不足することが考えられる。このような場合には，目次や帯の情報を用いることで不足した情報を補え，目次や帯の情報も用いる意義はあると考えられる。

しかし，一方で，目次や帯を用いることで正解カテゴリに分類されない例もある。表7は，書名では正解のカテゴリ 493.7(神経科学，精神医学)に分類されているが，他の分類キーと組み合わせると不正解となる。これは，目次に含まれる「ギリシア」や「近代」「西欧」，帯に含まれる「渉猟」や「史」「精神」「ヨーロッパ」などの語が他のカテゴリでの重みが大きかったためである。

すべての例を分析したわけではないが，このような分析を進めていくことで，質的にどのような原因があるかを突き止めることができるのではないかと考えている。さらに分析を進めることが必要である。また，質的な分析だけでなく，データ量の分析など量的な分析も必要である。

7　まとめ

本研究では，目次や帯の情報を手がかりとして，図書をNDCカテゴリに自動分類する実験を行った。その結果，目次や帯を書名と組み合わせて用いることで，性能の向上につながることがわかった。自動分類においても，人手による分類作業と同様に，書名が図書の主題を表すものとして最も適切なものであるが，目次や帯の情報は書名では不足する情報を補完する可能性を持っていることがわかった。

しかしながら，最大でも33.9%の再現率であ

表5　誤分類された書名の特徴と正解となった件数

		誤分類件数		正解となった件数（目次や帯）	
問題	例	件数	割合	件数	割合
カタカナ文字列を含むもの	『ベン＆ジェリー　アイスクリーム戦略―「価値主導のビジネス」が生んだ成功物語』	658	45.90%	311	47.3%
アルファベットを含むもの	『ＯＬＤ　ＪＡＰＡＮ　ＲＥＤＩＳＣＯＶＥＲＥＤ―「日本」‐美と文化の再発見』	98	6.50%	55	56.1%
固有名詞を含むもの	『知られざる大隈重信』	270	15.30%	122	45.2%
人手でも分類困難なもの	『江藤さんの決断』	90	10.50%	44	48.9%
分類不可	『Ａｔｍｏｓｐｈｅｒｉｃｓ』	22	1.50%	4	18.2%

表6　目次や帯を用いて正しく分類された例

分類キー	内容	システムの分類	正誤
書名	アンダーソンとフォークナー	分類不能	×
目次	第1章　出会い；第2章　それまでのアンダーソン；第3章　それからのアンダーソン；第4章　それからのフォークナー；第5章　フォークナーが見たアンダーソン	312.21	×
帯	詩作に悩みつつ密造ウィスキーを呷る若者。「君はどこかに出発点を持たねばならない」と若者を諭す作家。若者は作家をオハイオの肥沃なとうもろこし畑に準えた。作家の名はシャーウッド・アンダーソン。そしてこの若者こそ後年ノーベル文学賞を授与されるウィリアム・フォークナーだった。	930.278	○
書名＋目次	‹‹省略››	312.21	×
書名＋帯	‹‹省略››	930.278	○
目次＋帯	‹‹省略››	930.29	×
全て	‹‹省略››	930.29	×

表7　目次や帯を用いたことで誤分類された例

分類キー	内容	システムの分類	正誤
書名	西欧精神医学背景史	493.7	○
目次	古代ギリシア；ギリシア治療文化の外圧による変貌；ヘレニズムに向かって；ローマ世界とその滅亡；中世ヨーロッパの成立と展開；魔女狩りという現象；魔女狩りの終息と近代医学の成立―オランダという現象；ピネルという現象―一つの十字路；ヨーロッパ意識の分利的下熱；ピューリタニズムと近代臨床；フランス革命第一帝政時代と公式市民医学の成立；啓蒙君主制下の近代臨床建設；新大陸の“近代”；大学中心の西欧公式精神医学；力動精神医学とその反響；一九世紀の再展望二〇世紀における変化；西欧“大国”の精神医学；西欧“小国”の精神医学；	371.2	×
帯	「人類史の中でヨーロッパとは何か」。若き日の著者をとらえつづけた問いへの答えを探りつつ、広範な分野の研究書を渉猟し、構築された異色の精神医学史。	702.05	×
書名＋目次	‹‹省略››	371.2	×
書名＋帯	‹‹省略››	702.05	×
目次＋帯	‹‹省略››	162	×
全て	‹‹省略››	162	×

ることを考えると，未だ実用化にはさらなる性能の向上が必要である。

誤り分析の結果からは，目次や帯がよい効果をもたらす場合と，そうでない場合があることがわかった。これらの傾向をつかみ，性能の向上に役立てたい。

また，書誌データを対象とした分類の場合，非常に多い分類カテゴリに対して，それを学習するためのデータ数が少ないことや分類されているデータ数に偏りが大きいことがいえる。現時点では，統計的手法が有効であったが，データ数を増やした場合に，どのような変化があらわれるかを検討する必要がある。また，現実的に十分なデータ数が揃わない場合を考慮すると，まず，十分なデータ数が揃うと想定できる 10 区分で先に分類を行い，各分野でさらに細かな分類を行うなど段階的な分類手法の可能性もある。

謝辞

本研究は，国立情報学研究所共同研究「異なるオントロジ間のマッピングの試み」，及び文部科学省科学研究費若手研究(B)16700241 の補助を受けた。

引用文献

1. Frank, Eibe; Paynter, Gordon W Predicting Library of Congress Classifications from Library of Congress Subject. Journal of the American Society for Information Science and Technology. Vol.55,No.3, p214-227 (2004)
2. Joachims,T.Text categorization with support vector machine：learning with many relevant features.Proceedings of the 10th European Conference on Machine Learning(ECML'98).p.137-142.

（1998）
3. 高村大也;松本裕治. SVM を用いた文書分類と構成的帰納学習法. 情報処理学会論文誌. Vol.44 No.SIG3, p1-10.
4. Aixin Sun;Ee-Peng Lim;Wee-Keong Ng. Web classification using support vector machineWorkshop On Web Information And Data Management archive. Proceedings of the 4th international workshop on Web information and data management. p96 – 99（2002）
5. Larson, R. Ray. Experiments in Automatic Library Congress Classification. Journal of American Society for Information Science. Vol.43, No.2, p.130-148（1992）
6. 石田栄美. 図書を NDC カテゴリに自動分類する試み. Library and Information Science. No.39, p.32-45（1998）
7. もりきよし. 資料分類法概論改訂版. 東京, 理想社, 1983, 163p.
8. 鮎沢誠, 芦谷清. 資料分類法. 東京, 東京書籍, 1984, 323p.（現代図書館学講座, 4）
9. Cheng, PTK ; Wu, AKW. ACS - An Automatic Classification-System . Journal of Information Science. Vol.43, No.2, p.130-148（1995）
10. もりきよし. “3.4 分類規程”. 日本十進分類法新訂 9 版本表編. 日本図書館協会, 1995, 418p.
11. 奈良先端科学技術大学院大学松本研究室. 茶筌. [2006-02-20], <http://chasen.naist.jp/hiki/ChaSen/>
12. SVM-Light Support Vector Machine. [2006-02-20], < http://svmlight.joachims.org/>