論文

日本十進分類法と基本件名標目の相互マッピングの試み

石田栄美

［要旨］本研究では，国立情報学研究所が提供するNACSIS-CATに入力された目録データを用いて，日本十進分類法第9版による分類記号と基本件名標目表の件名標目を対象に相互マッピングを試みた。100,000件の目録データを用いて，Support Vector Machineによる手法と相対出現率を用いた手法によるマッピングを行い，10,000件の目録データを用いて評価した。11交差検定でマッピングを行った結果，分類記号から件名標目，件名標目から分類記号へのいずれのマッピングも，相対出現率を用いた手法の正解率が高いという結果であった。しかしながら，その正解率は半分程度にとどまっており，さらなる改善が必要である。

［キーワード］日本十進分類法，基本件名標目表，マッピング，自動分類

1．はじめに

コンピュータ性能の向上，インターネットの普及などに伴って，図書館業務の効率的運営に対する期待が高まり，日本でも書誌ユーテリティの基盤が整ってきた。その結果，1985年に学術情報センターがNACSIS-CATの提供を開始した。NACSIS-CATとは目録・所在情報サービスであり，国内の大学図書館を中心とする各機関が所蔵する図書や学術雑誌についての総合目録データベースを形成するシステムである。現在，多くの参加機関の手によって総合目録の整備が進んでいる。目録は，一つの図書館または複数の図書館が所属する図書館資料について記述したものであり，目録には書誌的記述，標目，所在記号が含まれる。標目は資料を検索する手がかりとなるもので，標目になるものにはタイトル，著者名の他に，分類記号や件名がある。分類記号は，分類表にある主題に対応する記号であり，資料の主題を表現するものである。件名は，件名標目表から選び出された統制語であり，資料の主題を表すものである。記号と言葉という違いはあるが，分類記号，件名標目とも，資料の主題を表現するものであり，主題から目録を検索する際のアクセスポイントとして，重要な役割と果たすといわれている。

以上のように，総合目録データベースシステムが提供され，目録データの整備が進み，利用者の情報アクセスは容易になりつつある。しかしながら，登録された目録データ全てに分類記号や件名標目が付与されているわけではない。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，分類記号が付与されているのは481,894件（77.4%）であり，件名標目が付与されているのは401,441件（64.5%）に過ぎない。また，分類記号が付与されているレコード中（481,894件）で，件名標目が付与されている割合は，393,028件（81.6%）である。このように，目録データベースには，分類記号や件名標目が付与されていないレコード，分類記号か件名標目のいずれかしか付与されていないレコードが多数存在する。現在の目録における検索では，分類記号や件名をアクセスポイントとした網羅的な検索ができず，本来であれば検索できるはずの目録が検索できないことになる。つ

まり，現状では，主題からの検索が有効になってないといえる。

分類記号や件名標目が付与されていない目録データに対して，補完を行うためには，分類記号と件名標目の相互マッピングが必要となる。相互マッピングが可能になれば，分類記号を手がかりとした件名標目の付与，反対に件名標目を手がかりとした分類記号の付与が可能になる。これにより，分類記号または件名標目しか付与されていないレコードに対して補完を行うことができ，その結果として，主題からの網羅的な検索を行うことができるようになる。また，これから入力される目録データに対して，分類目録作業，件名目録作業の効率化も図ることができる。

本研究では，すでに入力された目録データを用いて，１レコード中に付与されている分類記号と件名標目のペア情報を用いて，分類記号と件名標目の相互マッピングを試みる。テキストに対する分類記号の付与や件名標目を付与した例はいくつかあるが，分類記号と件名標目の直接的なマッピングを自動的に行う試みは少ない。本研究では，目録データを用いた相互マッピングが可能かどうかを検証する。

以下では，分類記号と件名標目の相互マッピングの概要，用いたデータ，マッピング結果の評価，結果に対する考察などについて述べる。

２．関 連 研 究

件名標目から分類記号の推測を行った例としては，Frankの研究[1)]がある。米国議会図書館件名標目（以下，LCSH）から米国議会図書館分類表（以下，LCC）の分類記号の推測を行っている。これは，LCSHとLCCのペアである学習用データを用いて，機械学習手法であるSupport　Vector Machine（以下，SVM）手法によりLCSHのセットからLCCの分類記号への推定を行っている。10,000件から800,000件と量を変化させた評価用データを用いて，SVMによる手法と共出現情報から高出現のペアであったLCCの分類記号を推定先とする手法とを比較した。その結果，800,000件の学習用データを用いて，SVMによる手法の精度が55.32%と最も高かった。Frankの研究は，本研究とほぼ同じ目的であるが，本研究では件名標目から分類記号の推定だけでなく，分類記号から件名標目の推定も試みている。

タイトルやテキスト自体の情報を用いて，分類記号や件名標目を付与した例には，以下のようなものがある。

Japan　MARCのデータを用いて，図書に日本十進分類法の分類記号を付与した研究[2)]がある。９類を除いた範囲を対象に，書名からの文字列の切り出し手法３種類と重み付け手法４種類が比較されている。最も分類精度が高かったのは，切り出し手法に形態素解析システムを用いて単語を切り出した場合と，重み付け手法に，相対出現率を用いる方法であった。岸田[3)]は『図書館情報学文献目録』のレコードを用いて，文献目録独自の分類記号と件名標目の付与を行っている。分類記号の付与実験では，現在までに図書館情報学や情報検索分野の自動分類や自動索引研究において提案されてきた統計的手法を用いて分類記号と件名標目の付与を行っている。実験の結果は40%～60%の再現率であった。様々な統計的手法の比較の結果，「単純な」手法の性能がより優れていると指摘している。その他には，Larson[4)]がLCCの分類記号の付与を行っており，その際にタイトル情報だけでなく，LCSHの情報を用いている。

Frankの研究では，SVMによる手法を用いている。最近のテキスト自動分類研究においては，このSVMによる手法が最も精度が高いとされており，様々な応用例がある。しかしながら，日本語のテキストを対象にした分類記号や件名標目の付与については，従来の情報検索分野で提案されている統計的手法を用いた研究例のみであり，SVMを用いた手法は未だにない。そこで，本研究では，分類記号や件名標目の相互マッピングにおいて，SVMを用いた手法の有効性の検討も行う。

3．実験に用いたデータ

3. 1　NACSIS-CATの総合目録データ

対象とするデータは，NACSIS-CATの総合目録データである。NACSIS-CATへの参加機関は，2005年2月現在で1,000機関を超えており[5]，「総合目録データベース図書統計」[6]によれば，毎年登録される図書書誌レコードは，450,000件を超えている。日本で最大の総合目録である。

本研究では，NACSIS-CATに1990年から2000年までに入力された目録データ622,295件を対象にした。

目録データの実例を図1に示す。これは，『生活の中に学ぶ心理学：大学生の視座から』というタイトルの図書レコードの書誌的記述部分を示した例である。それぞれの書名，責任表示，出版者，分類記号や件名標目などの項目がある。「CLSKND」はそのレコードに付与されている分類法の種類であり，「CLSN」は分類法に対応する分類記号である。同様に，「SHTBLKND」は付与されている件名の種類であり，「SH」はそれらに対応する件名標目である。複数の分類法や件名標目に対して，対応する分類記号や件名標目が付与されている場合もある。この例では，

```
ID＝BA29910113
CTGL＝jpn
ISBN＝4563056065
TR＝生活の中に学ぶ心理学：大学生の視座から／竹内健児，小林哲郎共編
PBLC＝培風館
SOUCE＝TRC
YEARA＝1997
CLSKND＝NDC9
CLSKND＝NDC8
CLSN＝371.47
CLSN＝140
SHTBLKND＝BSH
SHTBLKND＝NDLSH
SHTBLKND＝NDLSH
SH＝心理学
SH＝青年心理学
SH＝臨床心理学
```

図1　NACSIS-CATのレコード例

NDC9（日本十進分類法第9版）の分類記号「371.47」とNDC8（日本十進分類法第8版）の分類記号「140」が付与されており，BSH（基本件名標目表）の件名標目「心理学」とNDLSH（国立国会図書館件名標目表）の件名標目「青年心理学」，「臨床心理学」が付与されている。目録には様々な項目があるが，本研究では，この中で分類記号と件名標目の組み合わせをペア情報「371.4：心理学」とし，この情報をもとに分類記号と件名標目のマッピングを行う。

3. 2　基礎統計

目録データ622,295件に対して分類記号と件名が付与されているレコード数を表1，表2に示す。一つのレコードに対して同じ種類の分類記号や件名が付与されている場合には1件としたが，異なる種類の分類記号や件名が付与されている場合にはそれぞれ1件とした。表1の「その他」には，

表1　分類記号の付与状況

分類表名	件数	割合
日本十進分類法第8版（NDC8）	452,120	72.7%
国立国会図書館分類表（NDLC）	187,603	30.1%
日本十進分類法第9版（NDC9）	184,467	29.6%
日本十進分類法第7版（NDC7）	39,621	6.4%
日本十進分類法第6版（NDC6）	9,776	1.6%
国立医学図書館分類表（NLM）	5,247	0.8%
米国議会図書館分類表（LCC）	2,315	0.4%
その他	1,404	0.2%

N＝622,295

表2　件名標目の付与状況

件名標目名	件数	割合
基本件名標目表（BSH）	354,466	57.0%
国立国会図書館件名標目表（NDLSH）	285,624	45.9%
国立医学図書館件名標目表（MESH）	16,529	2.7%
米国議会図書館件名標目表（LCSH）	8,517	1.4%
その他	8,013	1.3%

N＝622,295

デューイ十進分類法の18，19，20，21版などが含まれている。表２の「その他」には，米国議会図書館児童図書用件名標目表（JUSH,JVSH）やその他の件名標目表等（FREE）などが含まれている。

表１をみると，日本十進分類法第８版が付与されているレコードが最も多く，ついで国立国会図書館分類表の分類記号が付与されているレコードが多い。表２からは，基本件名標目表（BSH）による件名標目が付与されているレコードが最も多く，ついで国立国会図書館件名標目表による件名標目が付与されているレコードが多い。表１，２から，目録データベース中では，日本十進分類法による分類記号と基本件名標目表による件名標目が付与されているレコードが最も多いことがわかった。

3. 3　用いたデータ

本研究で実際にマッピングに用いたデータは，日本十進分類法の分類記号と基本件名標目表（以下，BSH）の件名標目が両方付与されているレコードである。日本十進分類法に関しては第８版が付与されているレコード数が最も多いが，現在は第９版が提供されているため，日本十進分類法第９版（以下，NDC9）を用いることにした。NDC9とBSHが両方付与されているレコードは114,974件であった。レコードに付与されていたNDC9とBSHについて，それぞれの異なり種類数，総件数，１レコードあたりの平均付与件数を表３に示す。異なり種類数は，BSHがNDC9のほぼ３倍と多くの種類が存在することがわかる。BSHは，単一の標目形だけでなく，より主題を的確に表現するために，細目を用いることができる。例えば，「伝記」という標目に，細目「--日本」を結合し，「伝記--日本」とすることができる。このため，BSHはその種類数が多くなると考えられる。

また，１レコードに付与される平均レコード数をみると，BSHの方がNDC9よりも多く付与されている傾向がある。

114,974件のうち，第一次区分をもとにした分類記号の分布を表４に示す。表４からは，３類「社会科学」の件数が最も多く，８類「言語」の件数が最も少ないことがわかる。件名標目のうち出現回数が多かった上位20位を表５に示す。表５

表３　基礎的な統計データ

	異なり種類数	総　件　数	平均
NDC9	11,738	184,466	1.6
BSH	31,373	354,466	3.1

表４　分類記号の分布

第一次区分（類）	総　件　数	
	件　数	割　合
0	9,230	8.0%
1	9,816	8.5%
2	20,280	17.6%
3	51,470	44.8%
4	18,907	16.4%
5	18,763	16.3%
6	9,032	7.9%
7	17,943	15.6%
8	5,827	5.1%
9	23,198	20.2%

N＝114,974

表５　出現頻度が高い件名（上位20）

出現回数	件　名	出現回数	件　名
4,139	電子計算機--データ処理	1,216	経営管理
		1,074	料理
3,227	電子計算機--プログラミング	1,064	人生訓
		1,063	英語
2,606	電子計算機	1,032	会計
2,517	データ通信	956	コンピュータ・グラフィックス
2,349	通信網		
1,372	太平洋戦争	890	英語--会話
1,310	環境問題	868	健康法
1,276	看護学	844	学習指導
1,241	教育	814	日本--歴史--古代
1,231	老人福祉		

をみると，情報科学に関する件名標目が多く付与されていることがわかる。これは分類記号の分布とは一致していない。また，細目が結合された形の件名標目も多いことがわかる。

マッピングには，この114,974件からランダムに抽出した110,000件を用いた。データの偏りによるマッピングへの影響を防ぐために，データを11分割し，交差検定を行った。11分割したもののうち，学習用データに100,000件（10分割分），評価用データに10,000件（1分割分）を用いて，それぞれ評価用データセットを入れ替えることによって11回のマッピング実験を行った。

4．相互マッピングの概要

4.1　実験の手順

分類記号と件名標目の情報を用いて，分類記号から件名標目へのマッピング，件名標目から分類記号へのマッピングを行った。本研究の概要を図2に示し，具体的な手順を以下に示す。

⑴ NACSIS-CATの目録データからNDC9とBSHの両方が付与されているレコードを抽出する。

⑵ 各レコードに対して，NDC9とBSHのペア情報を取り出す。

⑶ レコード集合を，マッピングに用いるための学習用データ100,000件と，マッピング結果を評価するための評価用データ10,000件に分割する。

⑷ ペアの情報を用いて，各マッピング手法を用いて，マッピングを行う。

⑸ 評価用データを用いて，マッピングの結果を評価する。

手順の⑴から⑷までが図2の罫線で囲まれたマッピング部分に相当し，⑸は点線で囲まれた評価の部分に相当する。用いたデータで述べたとおり，この操作をデータセットによって11回行った。また，分類記号から件名標目へのマッピングと，件名標目から分類記号へのマッピングの2種類を，それぞれ2つの手法を用いて作成した。

以下では，マッピング手法とマッピング結果の評価方法について述べる。

4.2　マッピング手法

マッピングは，分類記号と件名標目のペア情報をもとに行うが，分類記号と件名標目は実際のデータからは様々な組み合わせが存在する。マッピングとは，ペア情報からより適切な対応関係を抽出することが問題となる。本研究では，分類記号や件名標目の付与方法において従来から情報検索分野で適用されている手法と，テキストの自動分類において精度が高いとされている機械学習手法を用いて，マッピングが可能かどうかを検証すること，また，どちらの手法がマッピングに適しているかを検証することが目的であるため，マッ

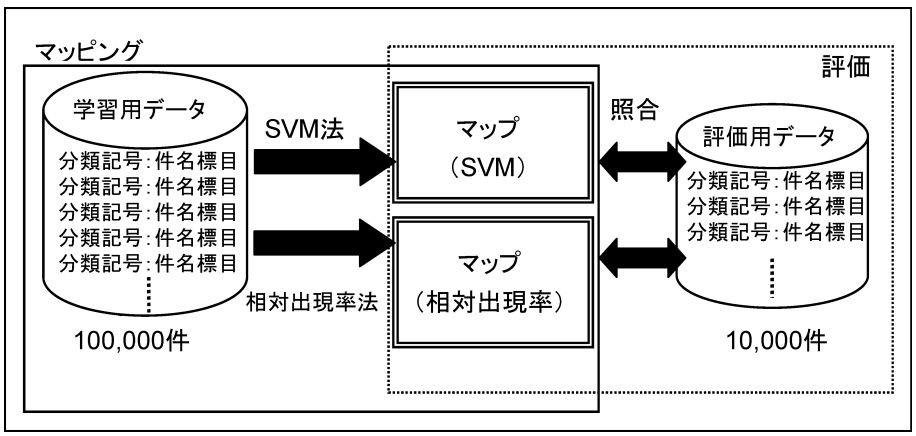

図2　マッピングの概要

ピングには，機械学習法のSVMによる手法と情報検索分野の統計的手法である相対出現率による重み付けを用いた２種類のマッピング手法を用いた。

以下では，２種類のマッピング手法について説明する。なお，分類記号から件名標目へのマッピングを例とするが，件名標目から分類記号へのマッピングも，分類記号と件名標目のデータをそのまま入れ替えることにより，可能となる。

4.2.1　SVMによる手法

SVMは，パターン識別手法の一つであり，“訓練データを正例と負例に分け，かつ，正負例間のマージンが最大になるような超平面を求める学習器”[7]である。テキストの自動分類研究では，テキストがあるカテゴリに分類されている場合を正の例，されていない場合を負の例にあてはめることができ，学習用データを，正例と負例に分けることができるため，様々な応用[8]が試みられている。現在では，精度が最も高い手法として注目されている。

分類記号から件名標目へのマッピングが，テキストの情報をもとに一つの分類記号を付与するという問題と対応することができるため，本研究でもSVMを用いることにした。テキストを分類する際には，テキストに出現する大量の単語をもとにしているが，分類記号と件名標目の場合は，そのもととなる量が絶対的に少ない。このようなデータがどのようにSVMの精度に影響を与えるかを検討する。

実験には，Joachimsが提供しているSVMソフトウェアSVM^{light}[9]を用いた。カーネル関数には，線形カーネル関数を用いた。分類記号から件名標目のマッピングの場合，入力ベクトルは分類記号に対し，ペアとなる件名標目が出現する場合は１，出現しない場合は０を重みとして与えた。件名標目から分類記号のマッピングの場合も同様に，件名標目に対し，分類記号が出現するか（重み１），出現しないか（重み０）とした。入力ベクトルは，各分類記号，件名標目に対して作成する。

4.2.2　相対出現率による手法

相対出現率による重み付け手法は，書名から図書に分類記号を付与する実験[2]において，分類精度が最も高かった手法を用いた。書名から分類記号を付与する場合には，書名を単語に分割し，その単語の出現と，書名に付与されている分類記号との関係から，重み付けを行い，最終的に分類先を決定する。単語を件名標目と想定すると，SVMによる手法と同様に，この手法もマッピングに応用することができる。

相対出現率による重み付けのマッピング手法は，分類記号と件名標目のペアから，分類記号に対して，ペアとなる件名標目の出現回数を利用し，件名標目に重み付けを行う。以下のような重み付け手法を用いた。

NDC9とBSHのペアから，NDC9の分類記号をCi（ｉ＝1，2，3，……，N），NDC9の分類記号とペアとなる件名標目Sj（ｊ＝1，2，3，……，M）の出現回数Fijとしたとき，出現率による重みwijは，以下の式で求める。

$$W_{ij} = \frac{F_{ij}}{\sum_{i=1}^{N} F_{ij}}$$

この重みは，該当する分類記号が付与されているレコードにおいて，件名標目が出現している回数をそのまま重みに適用した非常に単純な方法である。分類記号に対する件名標目のペアが多ければ多いほど，その件名標目の重みが高くなる。最も高い重みを持つ件名標目をその分類記号に対応する件名標目とした。件名標目が同じ重みを持つときには，そのすべてを対応する件名標目とした。

4.3　評価方法

評価用データを正解とし，マッピング結果の一致率を正解率として求めた。正解率は分類記号，件名標目ごとに求め，それらを平均したものをマッピング結果全体の正解率とした。各分類記号や件名標目の正解率は以下の方法で求めた。図３に示すように，各分類記号に対する正解率を求める方法を例として説明する。まず，評価用データ

からある分類記号に対応しているペア情報を抽出する。そのペア情報とマッピングの結果を比較する。図の例では，評価用データでは，分類記号「371.47」に対応する件名標目は，それぞれ「青年期」「青年」「高校生」「青年心理学」である。一方，マッピングの結果は，分類記号「371.47」には件名標目「青年心理学」にマッピングされている。この場合，4件の評価用データに対して，マッピングの結果は1件のみが一致していたことになり，正解率は25%となる。これらを各分類記号に対して行い，平均したものが最終的な正解率である。

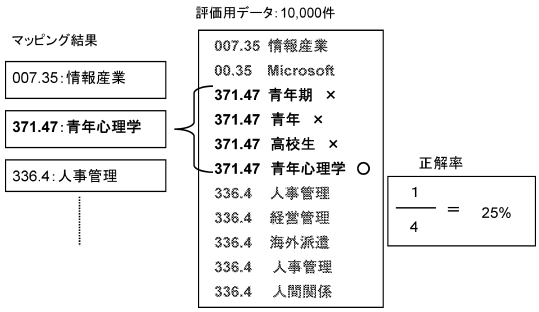

図3　各分類記号における正解率の求め方

5．実 験 結 果

5.1　マッピング結果

図4に相対出現率による手法で作成したマッピング結果の一部を示す。分類記号「318.2」は件名標目「地方自治--日本」にマッピングしていることを示している。また，件名標目「地方自治--日本」と「地方自治」は分類記号「318」にマッピングしていることを示している。多くの場合は分類記号と件名標目は一対一の対応であったが，「318.2」のように，件名標目からのマッピングがない分類記号や件名標目の例もあった。

5.2　実 験 結 果

10,000件の評価用データを用いてマッピング結果を評価した。11交差検定を行ったので，11通りのマッピングとなる。交差検定のために作成した

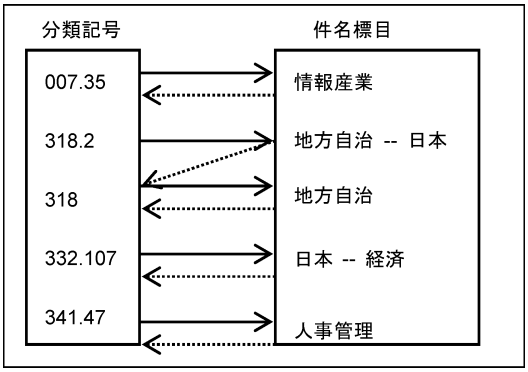

図4　相対出現率による手法のマッピング結果の一部

11の実験データセット名を0から10とし，各データセットにおける正解率とその平均を表6と7に示す。「カテゴリ数（件）」の「評価した」とは，正解率の平均を求めた分類記号や件名標目の種類数である。以後，ここではこれをカテゴリ数と呼ぶ。「評価できない」とは，学習用データに出現した分類記号や件名標目の中で，評価用データに出現しないカテゴリ数であり，これらは評価の対象外とした。

表6から，分類記号から件名標目へのマッピングの平均の正解率は，相対出現率による手法は47.80%，SVMによる手法は33.59%であり，相対出現率による手法が正解率が高いことがわかる。また，表7から，件名標目から分類記号へのマッピングの平均の正解率は，相対出現率による手法は53.92%，SVMによる手法は16.23%であり，相対出現率による手法の正解率が高いことがわかる。いずれのマッピング結果もSVMを用いた手法よりも相対出現率を用いた手法の方が正解率が高かった。また，各データセットの正解率をみると，全ての場合において相対出現率による手法の正解率が高いことを示していることがわかる。

それぞれの手法ごとに正解率を見てみると，相対出現率を用いた手法では，分類記号から件名標目へのマッピングよりも，件名標目から分類記号へのマッピングの正解率が高かった。しかしながら，その差は5%前後にとどまっており，マッピングの対象による違いは小さいといえる。反対に，

SVMを用いた手法では，件名標目から分類記号へのマッピングは件名標目から分類記号へのマッピングと比べて，その半分程度と非常に低い値となっている。SVMによる手法はマッピングの対象によって違いがあることがいえる。

表６　分類記号から件名標目へのマッピングの正解率

データセット名	正解率（%）		カテゴリ数(件)	
	相対出現率	SVM	評価した	評価できない
0	47.93	34.24	3,106	508
1	47.26	33.10	3,147	499
2	47.21	33.48	3,125	523
3	47.42	34.18	3,077	458
4	48.43	34.14	3,153	512
5	47.06	32.80	3,121	508
6	48.87	33.83	3,140	471
7	47.59	32.99	3,038	517
8	48.37	33.82	3,067	501
9	47.82	34.11	3,117	511
10	47.80	32.75	3,113	511
平　均	47.80	33.59	3,109.45	501.10

表７　件名標目から分類記号へのマッピングの正解率

データセット名	正解率（%）		カテゴリ数(件)	
	相対出現率	SVM	評価した	評価できない
0	54.27	16.42	4,583	1,072
1	53.30	16.08	4,567	1,129
2	54.15	16.05	4,518	1,147
3	54.72	16.60	4,510	1,099
4	54.07	16.39	4,516	1,071
5	53.56	16.13	4,540	1,133
6	54.82	16.44	4,583	1,076
7	53.15	15.65	4,525	1,067
8	53.46	15.83	4,520	1,104
9	54.17	16.74	4,568	1,092
10	53.40	16.16	4,568	1,182
平　均	53.92	16.23	4,545.27	1,110.00

６．結果の分析

実験結果から相対出現率による手法の正解率が高いことが明らかになったが，その原因を考察するため，結果をいくつかの観点から分析した。ここでは，相対出現率による手法とSVMによる手法が分類記号や件名標目が持つ特性とどのようにかかわっているのかを中心として分析した。

その一例として，分類記号から件名標目へのマッピングについて分析した結果を述べる。

6.１　分類記号別の正解率

分類記号や件名標目が持つ特性の一つは，記号や言葉を用いて主題を表しているということである。主題によって正解率が異なるかを検証するために，分類記号ごとの正解率を調べた。

各正解率の平均を分類記号の第一次区分ごとに求めたものを表８と図５に示す。これらから，相対出現率による手法の正解率が高いため，SVMによる手法よりも，正解率の平均がすべて高くなっているが，分類記号ごとに違いをみると，両手法とも「8（言語）」において正解率が比較的高

表８　各分類記号における正解率の平均

第一次区分（類）	正解率の平均（%）		総件数
	相対出現率	SVM	
0	47.18	27.60	89
1	50.54	28.45	226
2	39.12	32.02	296
3	45.70	30.08	873
4	49.44	34.24	453
5	51.48	40.44	381
6	49.57	26.90	206
7	48.93	37.40	303
8	66.21	54.38	120
9	34.16	25.40	159

いということがいえる。また，「9（文学）」は両手法とも正解率が低いことがわかる。「0（総記）」や「1（哲学）」「6（産業）」に関しては，相対出現率による手法とSVMによる手法では差がついていることがわかる。以上のように，ある程度，分類記号によって，正解率に違いや差があることがわかった。

分類記号ごとに正解率をみていくと，一つの特徴として，正解率が0％と100％であるカテゴリの割合が多かった。そのため，これらの正解率を示した分類記号に何らかの特徴があるのかを検証した。各分類記号において，正解率0％と100％だったカテゴリ数を調べ，分類記号の第一次区分ごとにまとめた結果を表9と10に示す。正解率は分類記号ごとに算出しているが，表9は正解率が0％だったカテゴリ数，表10は100％だったカテゴリ数である。割合は，第一次区分に属する各分類記号に対して，該当する正解率であるカテゴリ数の割合である。

相対出現率による手法では正解率0％である割合が「8（言語）」では15.97％と最も低く，次いで，「4（自然科学）」の16.44％，「1（総記）」「3（社会科学）」の18.89％となっている。正解率100％の結果をみると，「8（言語）」の44.54％で最も割合が高く，次いで，「5（技術）」の29.84％，「1（哲学）」の27.74％となっている。SVMによる手法は，全体的に正解率が低いのは明らかだが，ほとんどの分類記号において，全カテゴリ数の半分以上が正解率0％になっており，非常に大きな割合を占めていることがわかる。正解率100％のカテゴリ数については，分類記号ごとに大きな違いがあることがわかる。

表8の各分類記号における正解率の結果からは，「8（言語）」のように正解率0％の割合が低い分類記号において，正解率100％の割合が高くなってはいたが，他の分類記号をみてみると，必ずしも正解率の高い分類記号はすべて正解率100％の

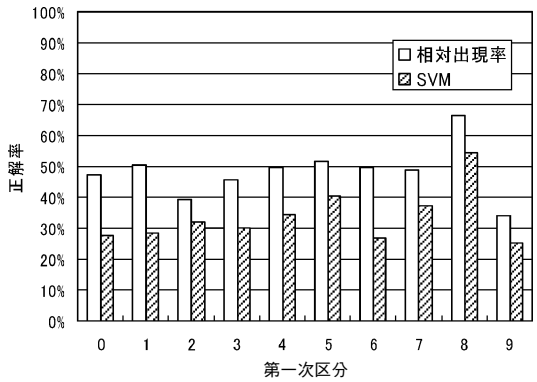

図5　各分類記号の正解率の平均

表9　各分類記号における正解率0％の件数

第一次区分（類）	相対出現率		SVM		総件数
	件数	割合	件数	割合	
0	17	18.89%	53	59.55%	89
1	50	21.46%	133	58.85%	226
2	72	24.32%	129	43.58%	296
3	164	18.89%	454	52.00%	873
4	74	16.44%	223	49.23%	453
5	73	19.11%	161	42.26%	381
6	57	27.67%	122	59.22%	206
7	72	23.76%	134	44.22%	303
8	19	15.97%	46	38.33%	120
9	65	40.88%	94	59.12%	159
平均	66.3	22.74%	154.9	50.64%	

表10　各分類記号における正解率100％の件数

第一次区分（類）	相対出現率		SVM		総件数
	件数	割合	件数	割合	
0	19	21.11%	11	12.36%	89
1	64	27.47%	36	15.93%	226
2	37	12.50%	34	11.49%	296
3	158	18.20%	115	13.17%	873
4	102	22.67%	85	18.76%	453
5	114	29.84%	92	24.15%	381
6	52	25.24%	31	15.05%	206
7	81	26.73%	56	18.48%	303
8	53	44.54%	46	38.33%	120
9	28	17.61%	22	13.84%	159
平均	70.8	24.59%	52.8	18.16%	

割合が高く，正解率０％の割合が低いとは言えない。それぞれの分類記号において，正解率が高いカテゴリ，低いカテゴリの例のような極端な例が見られるといえる。

6. 2　各分類記号における手法の比較

各分類記号において，相対出現率による手法とSVMによる手法の正解率を単純に比較した結果を表11に示す。合計は，その第一次区分に属する分類記号のカテゴリ数であり，割合は評価したカテゴリ数に対する件数である。マッピングの結果からもわかるように相対出現率の方が正解率が高い例が多いが，カテゴリによってはSVMによる手法が相対出現率による手法を上回る例があることがわかる。各分類記号においてSVMによる手法が正解率が高い場合があるということは，これらの手法は，データの特徴に影響を受けるといえるだろう。

6. 4　結果の分析のまとめ

分類記号，すなわち，主題によって，マッピングの結果に影響ができるかを検討したが，主題ごとに正解率が異なるという現象がみられた。しかしながら，正解率が高い例も低い例も混在しており，分類記号ごとに一定の傾向が見られるということではなかった。個々のカテゴリに関して正解率の比較をすると，事例によってはSVMによる手法が高い場合があった。今後はこれらの各事例について，具体的にどのような場合に，相対出現率による手法の正解率が高くなるのか，SVMによる手法の正解率が高くなるのかを検証していく必要がある。

表11　手法の正解率を比較した結果

第一次区分（類）	相対出現率による手法の方が高い		SVMによる手法の方が高い	
	件　数	割　合	件　数	割　合
0	39	43.8%	16	18.0%
1	100	44.2%	24	10.6%
2	81	27.4%	62	20.9%
3	352	40.3%	153	17.5%
4	182	40.2%	85	18.8%
5	111	29.1%	68	17.8%
6	84	40.8%	27	13.1%
7	88	29.0%	61	20.1%
8	32	26.7%	15	12.5%
9	46	28.9%	31	19.5%
合　計	1,115	35.9%	542	17.5%

本研究でマッピングに用いた学習用データは，分類記号と件名標目のペア情報のみであり，これは一般的なテキストの自動分類などに比べて，データが疎である。つまりマッピングに用いる情報量が少ないといえる。これらのデータを用いてもある程度のマッピングが可能であることは示せたが，データ量を増やすことにより，さらに正解率の高いマッピングが行える可能性がある。前述したFrankの研究[3]においても，データ量を10,000件から800,000件まで用いて分類記号の推定を行っており，800,000件用いた場合の結果が最も高かったことが示されている。このことからも，データ量を増やすことは有効ではないかと考えられる。

また，正解率の向上にはマッピングに用いる情報量をさらに増やすために，タイトル情報を介した分類記号と件名標目の相互マッピングが考えられる。しかしながら，タイトル情報は，現在に比べてその情報量が多くなるのでノイズが含まれる可能性もあり，今後の検討が必要である。

本実験では，テキストの自動分類と同様の問題と考えられるとし，テキストの自動分類において用いられている手法を適用したが，SVMによる手法よりも相対出現率による手法が正解率が高かった。これは，従来のテキストの自動分類研究においてSVMによる手法が優れているとされている結果とは異なっていた。日本語テキストの自動分類において，SVMによる手法と今回用いた相対出現率による手法と比較した実験例はないので，日本語テキストの自動分類も同様の結果となるのか，どのような手法が有効であるのかという

検証も必要である。

6．おわりに

本研究では，分類記号と件名標目のペア情報から分類記号から件名標目，件名標目から分類記号への直接的なマッピングを行った。その結果，これらのデータを用いてもマッピングがある程度可能なこと，またマッピング手法として相対出現率を用いた手法が有効であることはわかった。

しかしながら，その正解率は半分程度であり，今後はさらに，目録データの特徴の分析やマッピング結果の詳細な分析をさらに行い，手法の改良を目指す必要がある。

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, E. and Paynter, W.G. “Predicting Library of Congress Classifications from Library of Congress Subject Headings.” *Journal of the American Society for Information Science and Technology*. vol. 55, no. 3 (2004), p. 214–227.

2）石田栄美「図書をNDCカテゴリに分類する試み」『Library and Information Science』no. 39 (1998), p. 31–45

3）岸田和明「論文標題に基づく分類記号とディスクリプタの自動付与のための統計的手法」『日本図書館情報学会誌』vol. 47, no. 2 (2001), p. 49–66.

4）Larson, R.R. “Experiments in automatic Library of Congress Classification.” *Journal of the American Society for Information Science*. vol. 43, no. 2 (1992), p. 130–148.

5）NACSIS–CAT統計情報，http://www.nii.ac.jp/CAT–ILL/contents/ncat_stat_org.html (accessed 2005–05–15)

6）NACSIS–CAT統計情報・総合目録データベース図書統計，http://www.nii.ac.jp/CAT–ILL/contents/ncat_stat_bookdb.html (accessed 2005–05–15)

7）永田昌明，平博順「テキスト分類―学習理論の「見本市」―」『情報処理』Vol. 42, No. 1, (2001), p. 32–37.

8）Joachims, T. “Text categorization with support vector machines: learning with many relevant features.” *Proceedings of the 10th European Conference on Machine Learning* (*ECML '98*). 1998, p. 137–142.

9）SVM–Light Support Vector Machine. http://svmlight. joachims. org/［2004．10．11］

Mutual mapping of Nippon Decimal Classification and Basic Subject Headings

［**Abstract**］In this research, I attempt to make a mutual mapping of Nippon Decimal Classification (NDC) and Basic Subject Headings (BSH). The training data which I used were the catalogs of books and serials inputted into NACSIS–CAT. NACSIS–CAT is the online cataloging system which NII provides. We tested two methods, one is Support Vector Machine (SVM), and the other relative frequency for mapping. The relative frequency method proved better than performance of SVM method.

［**Keyword**］Nippon Decimal Classification, Basic Subject Headings, mapping, text categorization