分類記号と件名標目の相互マッピング

石田栄美(駿河台大学文化情報学部)　[emi@surugadai.ac.jp]
神門典子(国立情報学研究所)　上田修一(慶應義塾大学)

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 9 版による分類記号と基本件名標目の件名を対象に相互マッピングを試みた。相互マッピングは, 分類記号から件名の推定, および件名から分類記号の推定で評価した。SVM による推定と相対出現率を用いた重み付け手法による推定を行ったところ, 分類記号から件名標目への推定は, SVM による推定手法の正解率が高かった。

1．はじめに

情報を検索する場合には, それらの情報に関するメタデータなどを利用することが有効である。しかしながら, 異なるコミュニティの情報を横断的に検索する際に問題となるのは, それぞれのコミュニティが異なるメタデータを用いている場合である。異なるメタデータの相互変換システムも必要となるが, メタデータに用いられているオントロジ間のマッピングが可能になれば, 情報を見つけることがより容易になるといえる。ここでいうオントロジとは, 概念の階層関係や概念間の関係を表したものである。

本研究では, その例として, 図書館目録で用いられている分類記号と件名標目に着目し, 異なるオントロジ間のマッピングの手法について研究する。分類記号と件名標目のマッピングとして, 分類記号から件名標目の推定, 件名標目から分類記号の推定を行う。ここでは, まず, 同じ目録データに対して付与されている分類記号と件名標目の関係を利用することで, 分類記号と件名標目の相互マッピングが可能かどうかを検証する。

現在, 目録データの整備が進んでいるが, その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, 日本十進分類法(以下, NDC)による分類記号が付与されているのは 476,812 件（76.6%）である。件名標目は, 国立国会図書館件名標目表と基本件名標目(以下, BSH)などを合わせても 401,441 件(64.5%)に付与されているに過ぎない。目録データベース中に分類記号や件名標目のないレコードが大量に存在すれば, 分類や件名による検索の意義が損なわれるのは明かであろう。しかし, 既存の目録レコードに新たに件名標目を付与するには膨大な人手を必要とする。

一方, 目録データベースには, 分類記号のみで, 件名標目のないレコードが大量に存在している。既に付与されている分類記号と件名標目の間の自動的なマッピングを行い, 分類記号から件名標目を推定できるようなシステムを構築することにより, 件名標目未付与レコードへの件名標目付与の効率化を図ることができる。

本発表では, 分類記号と件名標目が両方付与されている目録データを用いて, 分類記号から件名標目の推定, 及び件名標目から分類記号の推定を行うことができるか, 実験を行った。推定手法には, Support Vector Machine（以下, SVM）による推定手法と相対出現率による重み付けを用いた推定手法を採用した。

2．関連研究

分類記号や件名標目の付与に関する研究には以下のようなものがある。

まず, 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)の分類記号の予測を行った Frank の研究 [1]がある。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。5 万件の評価用データを用いて推定手法の評価をしたところ, 80 万件の学習用データを用いて SVM を用いた手法の精度が 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, 本研究では件名から分類記号の推定だけでなく, 分類記号から件名の推定も試みる。

また，岸田[2)]は『図書館情報学文献目録』のデータを用いて, 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。この研究では，分類記号から件名への直接的な付与ではなく, 書名を介して，同じ書誌データに対して，双方の付与を行っている。その他には，Laron[3)]がLCC の分類記号の付与を行っており，その際に書名情報だけでなく，LCSH の情報を用いている。

３．NACSIS－CAT の総合目録データ

3.1　基礎統計

対象とするデータは，国立情報学研究所のNACSIS－CAT の総合目録データである。

1990 年から 2000 年までに NACSIS-CAT に入力された目録データ 622,295 件に対する基礎データを示す。分類記号，件名の付与状況を表 1 に示す。「いずれかの分類記号」とは，分類法の種類を問わず, いずれかの分類記号が付与されている, もしくは付与されていないレコード数である。「いずれかの件名」も同様である。「分類記号と件名」は，分類記号が付与されているレコード中で件名が付与されている, もしくは付与されていないレコード数である。表 1 からは, 前述したように件名が付与されていないデータが多いことがわかる。

表 2,3 に分類記号と件名が付与されているレコード数を示す。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 2 の「その他」には，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 3 の「その他」には，米国議会図書館件名標目表(JUSH,JVSH)やその他の件名標目表等(FREE)などが含まれている。

表 2 をみると，NDC8 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が付与されているレコードが多い。表 3 からは，BSH 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が付与されているレコードが多い。表 2,3 から，目録データベースの中では，NDC による分類記号とBSH 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。

表1　分類記号, 件名の付与状況

	付与	未付与
いずれかの分類記号	481,894	140,401
いずれのかの件名	401,441	220,854
分類記号と件名	393,028	88,866

表2　分類表の付与状況

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

表3　件名の付与状況

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

3.2 用いたデータ

本研究では，NDC の分類記号と BSH の件名標目が両方付与されているレコードを対象とした。ただし，NDC に関しては第 8 版のレコード数が最も多いが, 現在は 9 版が提供されているので，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 件である。114,974 件のうち，第一区分をもとにした分類記号の総件数, 分類記号の種類数を表 4 に示す。表 4 からは，３の「社会科学」の件数が最も多く，8 の「言語」の件数が少ない。分類記号は，114,974 件に対して，184,466 件付与されているので，1 レコードに対し，約 1.6 の分類記号が付与されていることになる。一方, 件名は 31,373 種類あり，出現回数が多い件名を表 5 に示す。表 5 をみると，情報科学に関する件名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，114,974 件に対して，総件数は 354,466 件であり，1 レコードあたり 3.1 の件名が付与されていることになる。

表4 分類記号の分布

分類記号	総件数		分類記号の種類数	
	件数	割合	件数	割合
0	9,230	5.0%	445	3.8%
1	9,816	5.3%	843	7.2%
2	20,280	11.0%	957	8.2%
3	51,470	27.9%	3,038	25.9%
4	18,907	10.2%	1,536	13.1%
5	18,763	10.2%	1,487	12.7%
6	9,032	4.9%	1,344	11.4%
7	17,943	9.7%	1,028	8.8%
8	5,827	3.2%	469	4.0%
9	23,198	12.6%	591	5.0%
合計	184,466	100%	11,738	100.0%

表5 出現頻度が高い件名(上位14)

出現回数	件名	出現回数	件名
4,139	電子計算機 -- データ処理	1,310	環境問題
		1,276	看護学
3,227	電子計算機 -- プログラミング	1,241	教育
		1,231	老人福祉
2,606	電子計算機	1,216	経営管理
2,517	データ通信	1,074	料理
2,349	通信網	1,064	人生訓
1,372	太平洋戦争	1,063	英語

推定実験には，この 114,974 件のうち，ランダムに抽出した 100,000 件を学習用データに，10,000 件を評価用データに用いた。また，分類記号は表 4 からもわかるように種類数が多いため，NDC 上位 3 桁のみを用いた。

4. 分類記号と件名の推定実験

4.1 実験の手順

分類記号から件名, 件名から分類記号の推定には, 学習用データを用いて学習した推定手法を用いる。分類記号から件名の推定実験は以下の手順で行った。

(1) NDC9 と BSH の両方が付与されているレコードを抽出する。約 110 万件である。
(2) 110 万件の中からランダムに 10 万件を学習用集合とした。
(3) 抽出したデータから，NDC9 の BSH のペアを取り出す。
(4) ペアのデータを用いて，各推定手法を学習する。
(5) 評価用データの分類記号を，件名から推定する。推定には，(4)で学習した推定手法を用いる。
(6) 評価用データのレコードにすでに付与されている件名を正解とし，推定された件名と比較し，正解率を求める。

なお, 件名から分類記号の推定も上の手順と同じ方法で行う。

4.2 推定手法

手順(3)では 2 つの推定手法を用いた。SVM による推定手法と相対出現率による重み付けを用いた推定手法である。以下ではそれぞれについて説明する。

4.2.1 SVM

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”[4]である。テキストの分類問題では，テキストをあるカテゴリに分類するか，しないかという問題にあてはめることができ，様々な応用[5]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。

本研究でも, 分類記号から件名が推定できるかできないかということに応用できるため，SVM を用いた。Joachims が提供している SVM ソフトウェア SVM^{light} [6]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。分類記号から件名の推定の場合，入力ベクトルは分類記号に対し, ペアとなる件名が出現する場合は 1, 出現しない場合は 0 を重みとして与えた。件名から分類記号の推定の場合も同様に，件名に対し，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。

4.2.2 相対出現率

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[7]において，分類精度が高かった手法である。SVM 同様に，この手法も推定に応用することができる。

相対出現率による重み付けの推定手法は, 分類記号と件名のペアから，分類記号に対して，ペアとなる件名の出現回数を利用し, 件名に重み付けを行う。具体的には，分類記号から件名への推定の場合, 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 を Ci（ i = 1 , 2 , 3 , ‥‥ , N ）としたとき，NDC9 とペアとなる 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, 件名が出現している回数をそのまま重みに適用したものである。分類記号に対して件名のペアが多ければ多いほど, その件名の重みが高くなる。推定方法は, 最も高い重みを持つ件名をその分類記号から推定する件名とした。また, 件名が同じ重みを持つときには, そのすべてを推定結果で得られた件名とした。

件名から分類記号への推定は, 分類記号と件名のデータを入れ替えて行った。

5. 実験結果

10,000 件の評価用データに対して, 分類記号から件名の推定(NDC9->BSH), 件名から分類記号の推定(BSH->NDC9)を行った正解率を表 6 に示す。評価用データにおいて, すでに付与されている分類記号や件名を正解とし, 推定手法が推定した分類記号や件名との一致率が正解率である。ただし, BSH->NDC9 の SVM に関しては全体の 8 割のデータを用いた結果である。

表 6 から, NDC9->BSH に関しては SVM を用いた手法が相対出現率を用いた手法よりも正解率が高いことがわかる。しかしながら, 正解率は半分にも満たなかった。

また, 相対出現率では, 分類記号から件名への推定の正解率よりも, 件名から分類記号への推定の正解率が高かった。これは, NDC は上位 3 桁のみを用いているのに対し, BSH は付与されているものをそのまま用いているので, BSH の種類の多さが影響してためであると思われる。BSH は細目を組み合わせた件名が多く種類数が増えるので, 今後はそれらを考慮する必要がある。

表6 推定手法による正解率

	SVM	相対出現率
NDC9->BSH	43.50%	28.60%
BSH->NDC9	1.61%	66.66%

6. おわりに

本研究では, 分類記号から件名, 件名から分類記号の推定を行った。その結果, SVM を用いた分類手法が有効であることはわかったが, 推定を行うのに十分な精度までは達していなかった。目録データの特徴を理解し, 間違って推定された例の分析を行い, 更なる手法の改良が必要である。

また, 基礎統計からもわかるように, 分類記号や件名は必ずしも一つしか付与されないものではない。本実験で用いた手法はどちらも一つの分類記号や件名を付与しない方法をとったが, 複数付与する手法の検討が必要である。

7. 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) 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.
3) 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.
4) 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.
5) 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.
6) SVM-Light Support Vector Machine. http://svmlight.joachims.org/ [2004.10.11]
7) 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