分類標数の相互参照に基づく多言語書誌データ検索システム

河手 太士, 藤井 敦, 石川 徹也
図書館情報大学
{kawate, fujii, ishikawa}@ulis.ac.jp

1 はじめに

日中国交の正常化および中国の改革解放によって、日中両国間の交流が活発になっている。これにともない、両国間の情報流通が活発になっている。中国においては各大学・研究所などで日本研究が盛んになっており、それらが所蔵する日本語資料の数も増加している。一方、日本ではNACSIS-CATや国立国会図書館のアジア文献センターによって中国語資料のデータベース化が進められている。そのため、日本では日本語キーワードを用いて中国語資料を、中国では中国語キーワードを用いて日本語資料を検索することが望まれている。そこで本研究は、この問題を解決するために日本語で中国語書誌データベース (China-MARC) を、中国語で日本語書誌データベース (Japan-MARC) を検索するシステムの開発を行っている [1–4]。

図書館における蔵書管理は、配架分類および書誌分類に基づいて行われる。日本では「日本十進分類法」(NDC)[5] が、また中国では「中国図書館 図書分類法」(CLC)[6] が一般的に使われている。そこで我々が提案するシステムは、日本語でNDC分類標数を検索し、参照表を用いてCLCに変換しChina-MARCを、中国語でCLC分類標数を検索し、参照表を用いてNDCに変換しJapan-MARCを検索する。

本システムはCLIR (Cross-Language Information Retrieval) システムである。CLIRシステムの検索方式としては現在、

(1) 対訳辞書・コーパスを利用する方式 [7–11]

(2) 機械翻訳システムを利用する方式 [12–14]

(3) 中間言語を利用する方式 [15–18]

がある。本研究は分類標数を一種の中間言語として用いるので、方式 (3) に属する。

2 分類法とMARC

2.1 分類法

図書館情報学における分類法とは、体系的に配列された分類概念をあらわす記号 (分類標数) とこれに対応する概念範疇を示す語 (分類名辞) をもつ体系である。

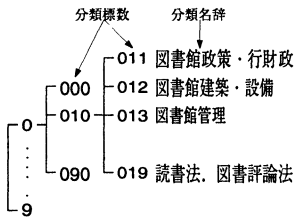

図 1: NDCの例

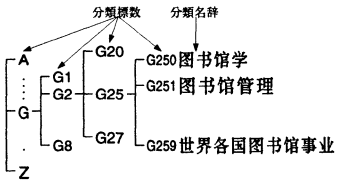

図 2: CLCの例

2.2 MARC

MARC (MAchine Readable Cataloging) は、図書館資料の題名や著者、出版関係などの書誌情報を機械可読化した書誌データベースである。書誌データのフォーマットの規格は、ISO-2709 (書誌的情報交換用レコード・フォーマット) として定められている。日本では国立国会図書館が Japan-MARC (図 3) を作成し、中国では中国国家図書館が China-MARC (図 4) を作成している。

ISBN :4-621-05073-7
タイトル:情報を考える
著者 :仲本秀四郎
出版地 :東京
出版者 :丸善
出版年 :1993.1
ページ数:178p
大きさ :18cm
NDC :007

図 3: Japan-MARC のデータ例

ISBN :7-5013-1044-0
书名 :情报学
著者 :塔拉卡诺夫
出版地 :北京
出版者 :书目文献出版社
出版年 :1993.5
ページ数 :318页
大きさ :19cm
CLC :G350

図 4: China-MARC のデータ例

3 研究システム

3.1 概要

本研究システムは、分類表データ、参照表データと以下の 5 つのシステム機能を必要とする (図 5)。

1. 分類名辞検索機能:キーワードによって NDC または CLC の分類名辞を検索する機能

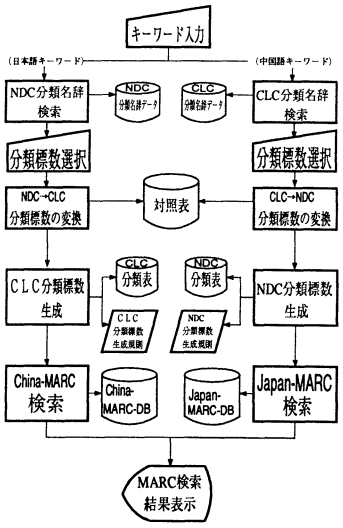

図 5: システムの構成

2. 分類標数変換機能:検索した分類標数を参照表を用いて変換する機能
3. 分類表数自動生成機能:分類標数生成規則に基づいて分類標数を自動生成する機能
4. MARC 検索機能:変換・生成された分類標数をもちいて MARC データを検索する機能
5. 文字コード変換機能:表示のために MARC データの文字コードを変換する機能

本システムを実現するための課題として以下の 3 点がある。

1. 参照表の作成
2. 分類名辞の拡張
3. 分類標数の自動生成

3.2 参照表の作成

NDC と CLC の間で分類体系が異なるため、双方の分類標数に対する参照表データの作成が必要となる。そこで、CLC 分類標数に対し NDC 分類標数を人手で対応させて参照表を作成した。

作成は以下の基準に従って行った。

- CLC 分類標数に対して 1 つ以上の NDC 分類標数を対応させる。
- CLC 分類標数に対し NDC 分類標数が複数になる場合は、そのすべてを対応させる。

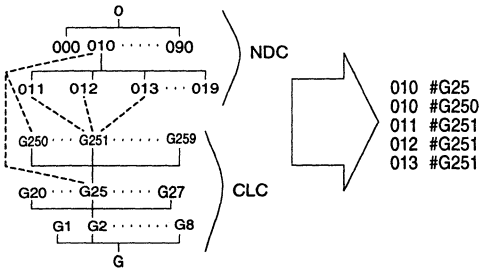

図 6: 参照表の例

このようにして作成した参照表 (図 6) の総参照数は 29,330、NDC 分類標数 1 つあたりの CLC 分類標数の参照数は平均 6.28 である。

3.3 分類名辞の拡張

NDC における収録語数は 33,365 語であり、NDC 分類標数 1 つあたり平均 3.72 語である。そこで、分類表の分類標数に対する分類名辞は十分ではないと考え、分類名辞を拡張した。1983 ～1996 年収録の Japan–MARC データ 929,758 件に対して、分類標数と書名中の名詞の関連度を TF・IDF 法 [19] を用いて計算し、分類標数との関連度が高い名詞を用いて分類名辞の拡張を行った (式 (1))。

$$R(cn,n) = f(cn,n) \times \log_2 \frac{C}{C_n} \qquad (1)$$

ここで、$R(cn,n)$ は 分類標数 cn と名詞 n の関連度、$f(cn,n)$ は 名詞 n と分類標数 cn の共起頻度、C は 分類標数の総数、C_n は 名詞 n と共起する分類標数の数である。

書名から形態素解析によって抽出した名詞 384,564 について関連度を計算した。さらに、利用者の質問に対し図書館員の回答事例を集めたレファレンス事例集 [20, 21] から、主題検索質問 15 件を用いて検索実験を行った。この結果、関連度 30 以上の名詞を分類名辞として用いたとき適合率、再現率ともに向上した。そこで、このような名詞を拡張語として採用し、NDC 分類標数 1 つあたりの分類名辞の数を平均 7.20 語に拡張した。

3.4 分類標数の自動生成

分類標数は、資料の「図書館」や「情報科学」などの主題をあらわす主題標数と資料の形態や言語などをあらわす補助標数との組合わせによって構成されている。そこで、分類標数を自動的に生成する機能を作成した。

事前分析として、NDC 分類標数が付与されている Japan–MARC のデータ 51,666 件について主題標数と補助標数との組合せの分析を行ったところ、「主題標数のみ」と「主題標数+形式区分」で 86.0%であったので、これらの組合せを対象に NDC 分類標数自動生成機能を作成した。Japan–MARC に付与されている「主題標数+形式区分」の分類標数 57 件を用いて評価実験を行ったところ、正解率は 92.6%であった。また、CLC 分類標数が付与されている China–MARC のデータ 22,010 件について主題標数と補助標数との組み合わせの分析を行ったところ、「主題標数のみ」と「主題標数+総記再区分」で 79.7%であったので、これらの規則を対象に CLC 分類標数自動生成機能を作成した。

4 おわりに

本論文は、参照表の作成、分類名辞の拡張、分類標数自動生成のシステム化の研究を行った。本システムでは、日本語で中国語書誌データベースを、中国語で日本語書誌データベースを検索することが可能である。

今後の課題として以下の点があげられる。

- 分類表と分類表の参照表 (たとえば NDC と UDC の参照表) をすべて人手によって作成することは時間がかかり、困難である。そこで、(半) 自動的に参照表を作成する必要がある。

- 分類表は改訂されるたびに体系が変化する。分類表の改訂ごとに参照表もを効率よく修正しなめればならない。
- 分類標数と書名中の名詞の関連度による分類名辞の拡張では、書名中に専門語があらわれる可能性が低い。そこで、専門用語シソーラスなどを利用して分類名辞の拡張を図る必要がある。

参考文献

[1] F. Kawate and T. Ishikawa. A Mutual Reference Retrieval System for Japan/China-MARC using NDC and CLC. *Proc. of the 2nd International Conference on Terminology, Standard ization and Technology Transfer*, pp. 516–523, 1997.

[2] 河手太士, 石川徹也. 日本十進分類表－中国図書館・図書分類表相互参照システムに基づく Japan–MARC および China–MARC 検索システム. ディジタル図書館, No. 10, pp. 13–23, 1997.

[3] 河手太士, 石川徹也. 日本十進分類表－中国図書館・図書分類表相互参照システムにおけるシソーラス展開. 第 5 回国立国語研究所国際シンポジウム第 1 専門部会「言語研究とシソーラス」, pp. 119–125, 1997.

[4] 石川徹也, 河手太士, 石間衛. Common Indexing/Retrieval Language としての「分類表」資源の活用. 電子情報通信学会・自然言語処理シンポジウム「実用的な自然言語処理に向けて」, 1997.

[5] もり・きよし（編）. 日本十進分類法 新訂 8 版. 日本図書館協会, 1981.

[6] 中国図書館図書分類法編輯委員会（編）. 中国図書館図書分類法 (第三版). 書目文献出版社, 1990.

[7] Lisa Ballesteros and W. Bruce Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 64–71, 1998.

[8] Mark W. Davis and William C. Ogden. QUILT: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 92–98, 1997.

[9] David A. Hull and Gregory Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–57, 1996.

[10] Akitoshi Okumura, Kai Ishikawa, and Kenji Satoh. Translingual information retrieval by a bilingual dictionary and comparable corpus. In *LREC workshop on translingual information management: current levels and future abilities*, 1998.

[11] Atsushi Fujii and Tetsuya Ishikawa. Cross-language information retrieval using compound word translation. In *Proceedings of the 18th International Conference on Computer Processing of Oriental Languages*, 1999. (To appear).

[12] Denis A. Gachot, Elke Lange, and Jin Yang. The SYSTRAN NLP browser: An application of machine translation technology in multilingual information retrieval. In *ACM SIGIR Workshop on Cross-Linguistic Information Retrieval*, 1996.

[13] Douglas W. Oard and Paul Hackett. Document translation for cross-language text retrieval at the University of Maryland. In *The 6th Text Retrieval Evaluation Conference (TREC-6)*, 1997.

[14] 酒井哲也, 梶浦正浩, 住田一男. Cross-language 情報検索のための BMIR-J2 を用いた一考察. 情報処理学会 自然言語処理研究会, Vol. 99, No. 2, pp. 41–48, 1999.

[15] Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. Translingual information retrieval: A comparative evaluation. In *Proceedings of the 15th International Joint Conference on Artficial Intelligence*, pp. 708–714, 1997.

[16] Susan T. Dumais, Thomas K. Landauer, and Michael L. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In *ACM SIGIR Workshop on Cross-Linguistic Information Retrieval*, 1996.

[17] Gerard Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, Vol. 21, No. 3, pp. 187–194, 1970.

[18] Páraic Sheridan and Jean Paul Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 58–65, 1996.

[19] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[20] 東京都立多摩図書館参考奉仕課（編）. 調べて探して見つけ出す：都立多摩図書館レファレンス回答事例集. 東京都立多摩図書館, 1990.

[21] 図書館学演習研究会. 参考業務：原論から演習まで. 学芸図書, 1985.