DESIDOC Bulletin of Information Technology, Vol. 17, No. 4, July 1997, pp. 17-19

Application of OCR in Building Bibliographic Databases

ARD Prasad *

Abstract

Bibliographic Databases tend to be very verbose and pose a problem to libraries due to the huge amount of data entry involved. In this situation, the two technologies that offer solutions are retro-conversion and optical character recognition (OCR). The application of building an intelligent system for automatic identification of bibliographic elements like title, author, publisher, etc is discussed here. This paper also discusses the heuristics in identifying the elements and resolving conflicts that arise in situations where more than one bibliographic element satisfy the criteria specified for identifying the various elements. This work is being carried out at the DRTC with the financial assistance of NISSAT.

1. INTRODUCTION

The huge amount of data entry involved in building bibliographic databases is a major hurdle for building library databases. Larger the library the more time it takes to build a database of its collection.

Although incredible speed has been achieved in the area of computerisation especially in data processing, it is well known that input and output of information to the system is a bottleneck to be reckoned with. Usually, data input into the computer is done using a keyboard. Keying in the data greatly restricts the speed of input and varies with the speed of typist or data entry personnel. On an average, a typist can key in 40-70 words per minute. However, the recent advances in computer peripheral technology, like scanners and laser printers have greatly reduced the problems of input and output to a large extent.

OCR technology is used to convert printed information into electronic text that can be processed by a computer. Similarly, image scanners translate graphics (e.g., line, art, photographs) into digitized images for computer processing. Both OCR and image scanners use a similar technique to convert their data. Software within the computer determines how the data is to be manipulated and eventually stored.

A few years ago, most OCR systems on the market could not read dot matrix print, and few could read typeset documents. Today, many OCR systems are able to read dot matrix print, and several vendors offer scanners that can read almost any font or format (omnifont) including typeset material. The trend toward omnifont readability vastly broadens the range of document sources produced from typewriters and daisy wheel printers to laser printer, offset presses, typesetters, photocopiers, and letter quality dot matrix printers. Nowadays, with powerful software like Omnipage professional, Word Scan Plus, the success of recognition is almost 100%.

* *Documentation Research and Training Centre Bangalore - 560 059*

This paper presents results of the application of OCR technology in building bibliographic databases incorporating heuristics into the system. The work is being carried out at the Documentation Training & Research Centre (DRTC) of the Indian Statistical Institute, Bangalore with financial assistance from the National Information System for Science & Technology (NISSAT).

2. METHODOLOGY

The objective of the project assignment is to automatically generate bibliographic records from the title pages of monographs. The various steps involved are :

1. Scanning the title page
2. Using OCR software to convert the scanned images into text.
3. Identification of bibliographic data elements using artificial intelligence.
4. Assigning CCF tags to the data elements.
5. Generating of ISO 2709
6. Importing into CDS/ISIS.

An HP Scan jet III C scanner for scanning and OmniPage Pro software have been used for the purpose. Steps 1-2 are done mechanically using the scanner and OCR. Steps 4-6 are fairly straight forward once the proposed system identifies the various bibliographic data elements. The crux of the present work is the Step 3 with regard to identification of bibliographic data elements in the scanned text.

Out of the major two alternative approaches, viz., (i) based on the clues of the title page layout and (ii) based on the linguistic analysis of the scanned text, the present paper mainly focuses on the latter.

3. LINGUISTIC ANALYSIS OF THE SCANNED TEXT

3.1 Heuristics with regard to title

In order to automatically identify the title of a document, one of the best clues is that normally the line or lines that speak of a particular subject. In other words, as title usually consists of subject terminology and if any keyword appears in a string, there is every likelihood that it might be the title. One problem with this solution is that, it requires a large lexicon to find whether a particular term in the title is in the lexicon. Secondly, it needs regular updation as new terms keep appearing.

A few rules of thumb have been developed. The basic approach aims at keeping the dictionary at a minimum level. The three hypothesis formulated are:

1. Title in most cases contain
   - articles
   - prepositions

   E.g. An Introduction to Philosophy, Principles of Cataloguing, etc.

2. In case of titles with single words or adjectival phrases the system should look for common suffixes, e.g. '-tion' in Information; '-logy' in technology; '-sophy' in philosophy.

3. In addition, the lexicon can contain some most commonly occurring words like Introduction, Basic, Modern, Fundamentals, Principles, etc.

Prefixes have not been taken into consideration for the simple reason that most prefixes contain only two letters like 'in-' 'un-' 'im-' and the likelihood of conflict with words occurring in author or other fields is very high.

3.2 Identification of Publisher

For the identification of publisher, either the system should contain the entire list of publishers or alternatively it should look for most commonly used words as clues like 'Inc.', 'Press', 'Company', 'Publishing', 'Publishers', 'Pvt', 'Ltd', etc. In addition, there may be a few words like 'book' as is the case of 'book house'.

However, there is a likelihood that some words like 'and', 'the', etc., may occur in title as well as publisher fields. Such conflicting words have been identified and the problem has been resolved through other means like, whether the line is on the top of the page or the bottom of the page.

3.3 Identification of place

Compared to the lexicons of title terms and publisher terms, place names present a fairly limited set.

In addition, the basic clue that it occurs in the last line or in the last but one line, obviates the need to build a lexicon. However, the system does contain a lexicon and looks into it only when it fails to identify the bibliographic element or when it gets confused with place names having more than one word such as New York, New Delhi, etc. When the place name is included in publisher's name, e.g., Calcutta Printers, Oxford University Press, the identification becomes non-trivial.

3.4 Identification of Author

One of the most difficult problems of the present work is the identification of author element. It is fairly simple when the author name is preceded by 'by', 'edited by', etc. It is difficult to build a lexicon of either first-name or last-name of persons. As a last resort, one can use the method of residue, that is, if a string does not represent either title, or publisher, or place, it could be that of the author.

4. CD-ROM DATABASES

To build lexicons for different bibliographic data elements, about 88,000 records (1990+) from BOOK FIND CD-ROM database have been downloaded. A study is underway to identify the most frequently occurring words in each bibliographic data element. We hope that a statistical analysis of the downloaded data would help us identifying heuristics to build an expert system for identification of bibliographic data elements from the title pages of documents.

5. CONCLUSION

Much of the present attention is on the title pages. Since most of the title pages do not contain the year of publication, it is planned to scan the verso of the title page also. The page may also contain a prenatal cataloguing information, which can be used to countercheck some of the decisions made by the system while analysing the title page proper.

For human intelligence, it is very easy to identify the bibliographic elements on a title page. It is however, difficult to develop a computational model for the same purpose.

A simpler approach is to manually identify the elements and leave some mark preceding each bibliographic element. This does not require any artificial intelligence component in the system. However, the goal of the present project is to evolve a fully automatic system.

6. REFERENCES

1. Cisler, Sterve. Smart scanning: Big improvements in page recognition hardware and software. *Online*, 1989, **13**(1), 105-07.
2. Diehl, Stanford and Stein, Haward Eglow. Tame the paper tiger. *Byte*, April 1991, 220-41.
3. Howard, Alan. Optical character recognition and images. *Information Media and Technology*. 1989, **22**(3), 117-19.
4. Jean DE, D. OCR teaches your PC to read. *PC Computing*. August 1990, 95-100.
5. Jennings, PG, *et al*. Data capture by optical scanning of published material for database enhancement. *Program*, 1982, **16**(1), 1 7-26.
6. Jones, Mitt. OCR gets you from there to here. *PC Magazine*. July 1992, 267-319.
7. Mundgod, Madhwacharya. Application of optical character recognition and expert systems to cataloguing: A state-of-the-art report. Guided project II, DRTC, 1993.
8. Prasad, ARD. Optical character recognition (OCR) and its relevance to library and information work. *In* Current Research in Library and Information Science. RBSA Publishers, Jaipur, 1993, 157-63.
9. Rice, James. OCR for libraries: Only a few years away. *Library Journal*, 1981, **106**(15), 1603-05.
10. Smith, John, *et al*. Optical character recognition: The technology and its application in information units and libraries. British Library Research Report. No. 33, 1985.