Heuristics for identification of bibliographic elements from title pages

Durga Sankar Rath and A.R.D. Prasad

The authors

Durga Sankar Rath is a Lecturer in the Department of Library and Information Science, Ravindra Bharati University, Kolkata, India.

A.R.D. Prasad is an Associate Professor, Documentation Research and Training Centre, Indian Statistical Institute, Bangalore, Karnataka, India.

Keywords

Bibliographic systems, Data handling, Cataloguing, Classification schemes, Information operations

Abstract

This paper presents a methodology for automatic identification of bibliographic data elements from the title pages of books. Also enumerates the various steps like scanning the title pages, running optical character recognition (OCR) software, generating HTML files out of title pages and applying heuristics to identify the bibliographic data elements. Much of the paper deals with the surveys undertaken to analyze the characteristics of various bibliographic descriptive elements like title, author, publisher and other elements. The first survey deals with the sequence of the bibliographic data in the title pages. The second survey deals with the font size, font type and the proximity of each bibliographic element on the title pages. The survey results are then used to develop heuristics, in order to develop a rule-based expert system which can identify the bibliographic elements on the title pages. The results of the system are presented, along with problems encountered.

Introduction

One of the most time-consuming technical operations in libraries is cataloguing. The cataloguing process describes each item in a collection, organizes the description into a coherent structure of relationships, and provides a tool in the form of a catalogue to access any document in a library. Although the work involved in cataloguing is very time consuming and not easily automated, libraries have long tried to reduce the amount of time and effort involved (Akiyama, 1990). The process of determining bibliographic data from title pages of the documents is complex, yet systematic. An investigation of the intellectual process involved may yield a few heuristics to design an expert system paradigm that can automatically identify the bibliographic data elements from the title pages.

The process of descriptive cataloguing begins with the identification of bibliographic data about an item. They include the following (Hagler and Simmons, 1982):

- Title and Statement of Responsibility area (i.e. the name of the item and names designating its intellectual responsibility).
- Edition area.
- Publication and distribution area.
- Series area.
- Note area.
- Standard Number area.

Some bibliographic data can be easily found in the item itself, others may come from other sources. In order to ensure that all items are described in the same way using at least the same starting point for gathering data, the notion of "chief source of information", has been introduced by cataloguers. The "chief source of information" is defined as "the source of bibliographic data to be given first preference as the source from which a bibliographic description (or portion thereof) is prepared" (Gorman and Winklet, 1988). For monographs it is prescribed that "the page that occurs very near the beginning of a book that contains the most complete bibliographic information about the book" (Gorman and Winklet, 1978), called the title page, is to be the first preference as a source of information for descriptive cataloguing. "The title page serves a purpose of information ... and as a means of distinction and identification" (Wyner, 1980).

The purpose of this study is to investigate ways in which artificial intelligence techniques can be applied to cataloguing process. The basic problem is to analyze, in terms of the conceptual level and logical flows, the way a computer can be taught to recognize bibliographic elements from the title

Library Hi Tech
Volume 22 · Number 4 · 2004 · 389-396

page of a document (Jeng, 1986). The assignment of descriptors falls in the realm of subject indexing, where the decision making is based on subject analysis, which is an intellectual process. A few automated systems have been developed that take clues from human-mediated expressive titles of documents (Aptagiri *et al.*, 1995). However, it must be noted that the present study concerns only with recognition of descriptive cataloguing elements from scanned title pages and does not deal with automatic assignment of subject descriptors to the document.

A considerable part of the cataloguer's expertise lies not so much in the ability to execute the rules as in the ability to recognize the bibliographic conditions which determine the choice of rules. Even greater expertise is required to understand fully the purpose of the rules and to examine existing codes in a critical manner and suggest improvements. One important result of the development of expert systems and artificial intelligence in general is that it has made us aware of the importance of knowledge in a new way. For example, medical knowledge, as opposed to medicine itself, had not really been studied in a systematic manner (Hjerppe and Olander, 1985). The plethora of medical consultation systems from MYCIN onwards has helped to change that situation. To the extent that expert systems in cataloguing has value, it too is worthy of study.

A logical place to start is the title page. Even when confronted by a book in a language we cannot read, we can reasonably distinguish the title and the name(s) of the author(s) and publisher(s). If we could discover the heuristics we use to do this, then we might eventually be able to develop intelligent systems capable of cataloguing with the least human intervention. The interpretation of title pages is a rather complicated business, as librarianship students quickly realize when they are first taught cataloguing. Later the skills involved come to be taken for granted. Cognitive analysis using both students and experienced cataloguers might be one way of uncovering heuristics and identifying pitfalls. Another approach (Bertin, 1983) is suggested by semiotics, the theory of signs. We can also study the title page from syntactic, semantic, and pragmatic points of view.

At the syntactic (structural) level, we consider just the layout, using clues provided by the positions of the various features, sequence of occurrence, the size of the spaces, print size, and changes in the type font. At the semantic level, words and phrases like "by", "edited by", "ISBN", "Library of Congress Cataloguing-in-Publication Data", could be used to identify the various bibliographic data elements from the various sections of the title page or the verso of the title page of the document. Finally, at the pragmatic level, character strings which might represent names of authors, places, or publishers could be checked against authority files for verification, and the layout of the title pages could be compared with the patterns typical of books from various publishers to aid in the identification of uncertain features (Burger, 1984).

The present work is an attempt to study the physical features of title page – more specifically sequence of occurrence, font size, format, special characters, and so on.

Sequence of bibliographic data elements

For identifying the sequence of bibliographic data elements in the title pages of monographs, a survey was conducted by physically checking the documents. Mainly four major bibliographic elements have been identified, which usually appear in the title pages of monographs. These are title, author, publisher and place of publication. However, it does not mean that only these four data elements appear in the title pages. There may be some documents where some of them may not appear – we may have three elements or even two elements. Besides data elements like series or conference information may also appear. For this purpose a sample of 485 document title pages were taken to study the sequence of the four bibliographic data elements. The results are shown in Table I.

The study of the order of data elements on Title page shows that in more than 90 per cent of documents in the sample taken, "title" is the first in the sequence. The most common pattern found is "Title", "Author", "Publisher" and "Place of publication".

Study of the bibliographic data elements

Although we can say that the most probable sequence is TAPuPl, it should be noted that these

Table I Probability of each sequence

Order	Frequency	Probability
TAPuPl:	368	0.75876
TAPu :	60	0.12448
TAPlPu:	45	0.09278
ATPuPl:	3	0.00619
ATPlPu:	3	0.00619
TPuPl:	3	0.00619
TAPl :	2	0.00412
PuTA :	1	0.00206
Total	485	1.00000

Notes: "T" stands for Title; "A" stands for Author, "Pu" stands for publishers and "Pl" stands for place. Thus TAPuPl indicates the order of these elements appearing in a document

Library Hi Tech
Volume 22 · Number 4 · 2004 · 389-396

elements do not appear as simple four lines in a title page. Definitely there is much more information in many lines about these four descriptive elements, and perhaps a few more descriptive elements. To make an analysis of the information that appears in a title, another survey has been undertaken. This survey is a minor modification of the survey conducted earlier at DRTC by Mr Madhwacharya Mundgod (Mundgod, 1993).

Data collection

Data have been collected in the form of a questionnaire intended to collect information on each bibliographic data element like – line number, font size, preceding field, presence of terms (e.g. "Inc", or "Press" in case publisher field). The data collected includes reference books, text books, conference/seminar proceedings. Only English language books have been taken into consideration. The method of data collection adopted was systematic sampling (Cox and Miller, 1982).

Data analysis

Presence of bibliographic information

Out of the 500 samples, 499 have title, 458 have author, 483 have publisher, 441 have place fields. Frequency of occurrence of other fields, are as follows: sub-title (164), volume (4), edition (48), conference proceedings (34), year (161), series (39). Following sections present a summary of the analysis of various bibliographic descriptive elements.

Title information

- Titles are found in upper or upper middle portion of the title page.
- The title appears as first in the title page (75.15 per cent) (In few cases author or series occupies first position.)
- Fonts used in title field are the largest fonts (94.99 per cent) compared with the size of fonts in other fields.
- If the title and sub-title occurred in the same line, they are separated by ":" (colon) or "-" (hyphen).
- It is not necessary that title should have only alphabetic characters. Title string may have numerals, punctuation marks like comma, hyphen and others.
- Usually titles have the terms like "The", "An", "Introduction", "Theory", "in", "to".

Sub-title information

- If a sub-title occurs, it immediately follows the title.
- The sub-title occurs in between second and fifth line (94.74 per cent).
- If the title and sub-title are of same font size, they are separated by ":" or " a line" (horizontal line), or a "blank space" (a vertical space between the title and sub-title and the vertical height of blank space will be more than the vertical height of characters used in the title string).
- The sub-title may also have terms like "a", "an", "the", "to".
- Like the title, the sub-title may have numerals, punctuation marks.

Edition information

- Mostly edition is found in reference books (e.g. Encyclopaedias, Dictionaries, Manuals), because reference books require periodical updating.
- Always the term "Edition" appears in the "edition" string.
- The edition string generally consists of an edition number written either in numerals or by using alphabetic characters. (e.g. Edition 2, Second Edition).
- The edition field is either preceded by the author (43.18 per cent), or title (34.10 per cent) or sub-title (18.18 per cent). That is, the edition field occurs after author, or title or sub-title field.

Volume information

- A volume field is often found in reference books.
- The term "Volume" or "V" or "Vol." is present in the volume string.
- A volume string generally consists of a volume number. The volume number may follow or precede the term "Volume".
- Title or Author are most probable preceding fields for volume field.

Author/contributor information

- The author field usually occurs in the third or fourth line (51.96 per cent). Less frequently, it occurs in fifth line with 13.97 per cent, or the first line with 9.83 per cent, or the sixth line with 9.3 per cent.
- Terms like "edited by", "by", "editor", occur in the author field (32.75 per cent) of which "Edited" by occurs more frequently (54.67 per cent) and "By" with 30.00 per cent.
- Usually the author name does not exceed three or four words.

Library Hi Tech
Volume 22 · Number 4 · 2004 · 389-396

- It is most common that the author field has a single author (70.74 per cent). Less frequently it will have two authors (23.80 per cent) and still less frequently three authors (4.37 per cent).
- In case of multiple authorship, authors' names are separated by "Different line" (78.36 per cent), or "and" or "&" (17.91 per cent).
- The usual preceding fields for the author field are, title (57.14 per cent), sub-title (29.54 per cent), conference proceedings (5.33 per cent) and edition (3.63 per cent).
- The horizontal position of author field is centered (57.21 per cent), or left-aligned (34.28 per cent), or right-aligned (8.51 per cent).
- More than half (55.90 per cent) of the author fields have author affiliation. Usually it will be in italic fonts.

Conference proceedings information

- This field occurs less frequently (6.8 per cent).
- The probable line numbers for conference proceedings are 4th line (35.29 per cent), 3rd line (17.65 per cent), 2nd line (11.76 per cent), 6th line (8.82 per cent), 8th line (8.82 per cent).
- The field normally has any one or more of the following terms: Conference; Proceedings; Seminar; Symposium; Workshop.
- The title with 87.5 per cent, or the sub-title with 12.5 per cent are only the preceding fields for conference proceedings fields, if it did not occur in the first line.

Publisher information

- The publisher field appears in the lower portion of the title.
- It is common that the publisher field occurs in between fifth to ninth line (64.8 per cent).
- The publisher field contains a symbol, or publisher's logo, or publisher's trademark. (45.96 per cent).
- The symbol, or logo, or trademark appears in preceding line (60.36 per cent). Sometimes these precede (34.68 per cent) or follow (4.96 per cent) the publisher's name.
- The terms like "Inc", "Press", "Published by", occur frequently in publisher field. "Press" occurs with 29.21 per cent, "Publishing" with 22.86 per cent, "Company" with 16.51 per cent, "Inc" with 16.19 per cent and "Publishers(s)" with 7.30 per cent.
- The publisher field will be of third-largest font size (51.67 per cent), or second-largest font (31.88 per cent) and less frequently of fourth-largest with 10.35 per cent.
- The most probable preceding field for the publisher field is the author (61.88 per cent). Less frequently it precedes the year with 10.00 per cent, or title with 8.95 per cent, or place with 6.25 per cent, or sub-title with 4.37 per cent.

Place of publication information

- The place name also appears in the lower portion of the title page.
- Frequently it occurs in between 6th to 10th line (63.23 per cent) and less frequently in 12th line (7.26 per cent) or fifth line (6.12 per cent) or 11th line (5.67 per cent).
- If both the publisher's name and the place name occur in the same line, they will be separated by a "Blank space" (47.75 per cent) or "." (a dot) (18.92 per cent), or by "," (11.71 per cent) or by the publisher's logo (9.91 per cent). And less frequently "/" or "-"are also used to separate publisher's name and place name.
- A blank space (40.97 per cent) or a dot (".") (25.69 per cent), or "," (12.15 per cent), or "and" or "&" (11.11 per cent) are used to separate place names, if there are more than one place name. A different line, or "/" or "-" are also used less frequently to separate place names.
- It is common that in most cases, the "place" field follows the publisher field (90.93 per cent). In some cases the author follows (5.90 per cent), or year (1.36 per cent).

Year of publication information

- The occurrence of year field is not so frequent (32.2 per cent).
- The year field also appears in the lower portion of the title page.
- The year field always has four digit numerals (in a few cases a year will be printed, like May 1988).
- The year field occurs in the last line (31.68 per cent), or follows the place name (31.06 per cent) (i.e. both place name and year appear in the same line. Even that could be last line) or before the publisher's name with 29.19 per cent.
- The place name frequently precedes the year field (63.75 per cent) or the author field (24.38 per cent). And less frequently the publisher field precedes the year field (6.87 per cent).

Series information

- The most common position of the series field is the first line (92.5 per cent).

Library Hi Tech
Volume 22 · Number 4 · 2004 · 389-396

- In a series string, the term "series" is found (45.00 per cent).
- The series string usually ends with numerals, which indicates the series number (82.5 per cent).
- If a series occurs in something other than the first line, either it precedes the author field (2.5 per cent) or the publisher (2.5 per cent) or the conference proceedings (2.5 per cent).

Rule base derived through heuristics

Heuristics for title

One
IF a line appears in the first block of the Title Page,
AND IF it is in the largest font,
AND IF it contains some English words,
THEN that element may be the Title of the document.
Two
IF the same size font continues in the succeeding lines,
THEN other lines should be included in the Title.

Heuristics for other title information

One
IF a ':' or '-' appears in the Title,
THEN the words following ':' or '-', constitute the Subtitle.
Two
IF the Title continues to the next line,
AND IF the following line is in a lesser font,
THEN the following part may be the Subtitle.

Heuristics for edition

One
IF a line appears in between Title and Author Block (e.g. First vertical space or break),
AND IF it appears in a separate line,
AND IF it contains some specific words (i.e. edition, ed.),
THEN it may be the edition statement of the document.

Heuristics for volume

One
IF an element appears in between Title and Author Block (e.g. First vertical space or break),
AND IF it appears in a separate line,
AND IF it contains some specific words (e.g. Part, or, Volume, or, Vol., V.),
THEN it may be the volume statement of the document.

Heuristics for author/contributor

One
IF anything (string/s) appears in between Title and Publisher blocks,
AND IF it happens to be in largest font between title and publisher block,
AND IF those are not the English words, (English words may belong to authors' affiliation)
THEN that element may be the Author of the document.
Two
IF anything (line(s)) appears before Title,
AND IF it does not contain any English words, (English words may constitute series statement)
THEN it may be the Author of the document.
Three
IF the Author element is either followed, or, preceded by some specific terms ('editor', 'edited'),
THEN it may be the Editor of the concerned document.

Heuristics for conference proceedings

One
IF a line or continuous lines contain(s) some specific words (e.g. Proceedings, Seminar, Conference, Symposium, and some numeric figures),
THEN it may be the information about Conference Proceedings.

Heuristics for publisher

One
IF some elements are found after the largest gap (vertical space), (it is presumed that the gap between author and publisher is more than the gap between title and author)
AND IF any line of that Block (i.e. the last block) is in larger font,
THEN that may be the Publisher of the document.
Two
IF anything matches with the limited set of Publishers' lexicon,
THEN that may be the Publisher of the document.

Heuristics for place of publication

One
IF an element happens to be the last line of the last Block (after the largest gap),
THEN that may be the Place(s) of Publication.
Two
IF anything matches with the list of the Place in the lexicon,
THEN those may be the Places of Publication.

Heuristics for year of publication

One
IF anything appears with '19..' or, '20..', in the last block
AND IF that happens to be consist of four numeric figures,
THEN it may be the year of Publication.

Library Hi Tech
Volume 22 · Number 4 · 2004 · 389-396

Heuristics for series

One

```
IF anything appears before Title,
AND IF contains some of the terms (e.g.
series, endowment, some numbers,)
THEN it may be the Series of Publication
of the document.
```

Two

```
IF anything appears after Title,
AND IF contains some of the terms (e.g.
series, endowment, some numbers)
THEN it may be the Series of Publication
of the document.
```

Program for identification of bibliographic data elements from title page

After the Title Page is scanned, using OmniPage Pro 10 (OmniPage, 2001), it is saved in HTML file format. OmniPage software can generate a plain text (ASCII) file, however, since such file does not contain information about the font size, font face and other physical characteristics required for the present study, the HTML file format is used. The intention is to use this HTML page, which gives clues regarding physical features like font size, font face, to identify the bibliographic data elements from the Title Page using a program. Even for the creation of an online data base for tables of contents of books, this kind of model could be developed (Jett *et al.*, 1998). The rules derived from the heuristics are implemented in a Java program. Most of the artificial intelligence systems or expert systems are normally implemented in either PROLOG (Programming in Logic) or LISP, though C or Java are not uncommon. However, the problem with Prolog is it lacks good I/O operations. Hence, Java programming language was chosen for the present operation. Since the input is in HTML format, a tokenizer is required to identify the HTML tags. Firstly, the HTML tokenizer is described, this tokenizer is an adoption of the tokenizer given by Vanhelsuwe *et al.* (1996).

HTML Tokenizer

Tokenizing some input means reducing it to a simpler stream of tokens. These tokens represent recurring chunks of data in the stream. Any Java compiler, for example, would check for grammatical correctness of the programs by checking the sequence of tokens representing reserved word strings like class, import, public, void, and so on. By not having to actually deal with the exact character sequences themselves, tokenizing as a technique has the following two main advantages:

(1) It reduces code complexity.
(2) It allows for flexible, quick changes in input syntax.

Class "StreamTokenizer" can be used to turn any input stream into a stream of tokens. The programming model for the class is that a stream can contain three types of entities:

(1) Words (that is, multicharacter tokens).
(2) Single character tokens.
(3) Whitespace (including C/C++/Java-style comments).

Before we start processing a stream into tokens, we have to define which ASCII characters should be treated as one of the three possible input types, called "defining the syntax table for the stream".

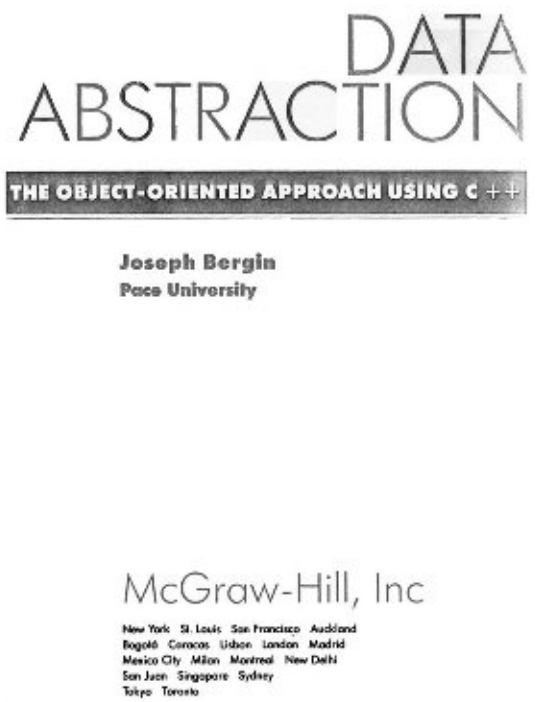
Sample pages

In the previous sections, the discussion is on the physical study of the title pages, the heuristics developed from it, then the corresponding rule base and the program part of it. All of these are aimed at the automatic identification of the bibliographic data elements from the Title pages of the document.

Sample of the actual page

For automatic identification purpose, the title pages of the documents are scanned. For this purpose, a HP ScanJet 6100C flatbed scanner, and the OmnipagePro 10 software are used. See Figure 1 and Figure 2 for an example of an original title page and the HTML result.

Figure 1 Sample of a scanned title page



Library Hi Tech
Volume 22 · Number 4 · 2004 · 389-396

Sample of the program output

The HTML Tokenizer (written in Java) is used for parsing the input files. If we use the HTML file in Figure 2 as the input to the system, we get the following output:

```
7 H1 ArialData: DATA
7 H1 ArialData: ABSTRACTION
0 null nullData: BR
4 null ArialData: THE OBJECT-ORIENTED
4 null ArialData: APPROACH USING C++
0 null nullData: BR
0 null nullData: BR
4 null ArialData: Joseph Bergin
0 null nullData: BR
4 null Times RomanData: Pace University
0 null nullData: BR
0 null nullData: BR
0 null nullData: BR
0 null nullData: BR
7 H1 ArialData: McGraw-Hill, Inc.
0 null nullData: BR
2 null ArialData: New York St. Louis San
Francisco Auckland Bogota Caracas
Lisbon London Madrid Mexico City Milan
Montreal New Delhi San Juan Singapore
Sydney Tokyo Toronto
```

Output for the SampleTitle Page

```
TI: Data Abstraction
PU: Mcgraw-Hill, Inc. (Used largest
Font)
PL: Tokyo
PL: New York
PL: London
PL: Sydney
PL: Toronto
PL: Singapore
PL: Delhi
AU: Joseph Bergin (Author after title)
OT: The Object-Oriented Approach Using
C++
```

Note: the first part of the output is meant to check whether the HTML tokenizer correctly identified the data, font type, font face. The second part is the actual output of the program, i.e. after identification of the bibliographic data elements. The information in parenthesis is to provide an idea about which heuristic rule is used to identify the element.

Conclusion

After development of the program, we have collected 50 documents randomly for final crosschecking. Then we used the scanned title pages (HTML files) as input to the program. The outcome is fairly promising. In 46 cases we have got the outcome clearly. Some of the problems observed in the study are enumerated as follows.

Figure 2 Scanned title page in TML format

Problems relating to OCR

- OmnipagePro 10 claims more than 99 per cent accuracy, and it is found the accuracy is impressive for original documents.
- We have taken photocopies of the Title pages because of the logistic problems in borrowing books from the library. From the photocopies we have generated the HTML pages using OmniPage. In some cases, the recognition is not completely satisfactory, as photocopies lose quality.
- In the case of light colored printing, the OCR has some difficulties. Mostly these problems arise when the letters are not printed in black. Recognition is best when the page is in black and white.
- In this study the library collection was used. The library stamp and barcode sticker in the Title page can confuse the OCR, and sometimes causes distortion to the final output in HTML.
- The HTML file of the scanned page misses some of the vertical spaces (breaks), which create problems in applying heuristics. This is a serious problem, especially when it removes many lines in between the author and publisher blocks. In some cases, though, the gap between the Author block and the Publisher block is bigger than other gaps in original page, but the scanned title page in HTML format shows that gap as smaller than other gaps.
- Sometimes, it happens that the data elements of different lines appear in the same line in the scanned HTML page. In a few cases, the reverse also happens, i.e. a scanned line is split into two lines in the HTML document.

Library Hi Tech
Volume 22 · Number 4 · 2004 · 389-396

- When an emblem or logo is adjacent to the publisher name, a part of the Publisher's name goes above the logo and another part of the publisher's name goes under the logo in the scanned HTML page. For example, in case of "Cambridge University Press", the line having the word "Cambridge" appears above the logo, whereas the line having "University Press" appears under the logo, as if it is a place name.
- In brief, although the character recognition of the OmniPage Pro is impressive, conversion to the HTML document is not always reliable. Since the present study focuses on the physical layout of the title pages, it becomes absolutely necessary that the OCR software should produce an accurate representation of the title page.

Problems relating to the program

- If a person's name appears before the title, but it is neither the author, nor a series title, the system finds it difficult to identify some of the descriptive elements. In the same title page, if a list of authors' names is given after title, this program goes astray, since personal name(s) can appear before the title and after the title. Figure 3 shows an example.
- Another serious problem arises when there is a series title, and then the first line of the title is in a smaller font size, and next line of the Title is in the largest font. So the system fails to recognize the first line as the part of the Title (see Figure 4 and Figure 5).

In the second example (Figure 5), the title is scattered across many lines, and different font sizes have been used to present various segments of the title. In such cases, the system fails in the identification process. The reason is that, the basic heuristics for identification is that the "title appears in the largest font in consecutive lines". The system counts as the title the first line where the largest font starts, and considers the title last line to be where the largest font ends. The rest of the title in the smaller font is considered as other title information, such as a subtitle.

Figure 3 Example of name appearing before title

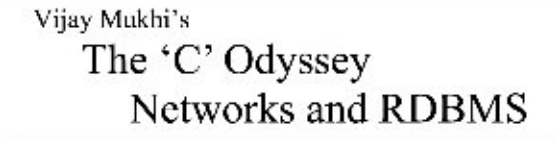


Figure 4 Title problem (example 1)

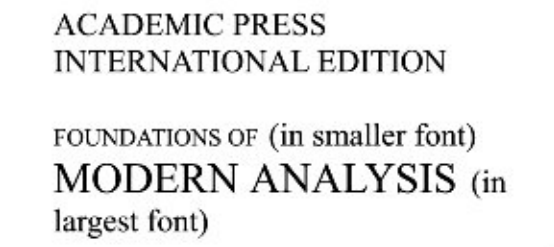


Figure 5 Title problem (example 2)

References

Akiyama, T.N. (1990), "Automated entry system for printed documents", *Pattern recognition*, Vol. 23 No. 11, pp. 1141-54.

Aptagiri, D.V., Gopinath, M.A. and Prasad, A.R.D. (1995), "A frame based knowledge representation paradigm for automating POPSI", *Knowledge Organisation*, Vol. 22 No. 3/4, pp. 162-7.

Bertin, J. (1983), *Semiology of Graphics: Diagrams, Networks, Maps*, University of Wisconsin Press, Madison, WI.

Burger, R.H. (1984), "Artificial intelligence and authority control", *Library Resources and Technical Services*, Vol. 28, pp. 337-45.

Cox, D.R. and Miller, H.D. (1982), *The Theory of Stochastic Process*, John Wiley & Sons, New York, NY, pp. 203-51.

Gorman, M. and Winklet, P.W. (Eds) (1978), *Anglo-American Cataloguing Rules*, 2nd ed., ALA, Chicago, IL.

Gorman, M. and Winklet, P.W. (Eds) (1988), *Anglo-American Cataloguing Rules*, 2nd rev. ed., ALA, Chicago, IL, p. 616.

Hagler, R. and Simmons, P. (1982), *The Bibliographic Record and Information Technology*, ALA, Chicago, IL, p. 118.

Hjerppe, R. and Olander, B. (1985), *Artificial Intelligence and Cataloguing*, Linkoping University, Linkoping.

Jeng, L.-H. (1986), "An expert system for determining title proper in descriptive cataloguing: a conceptual model", *Cataloguing and Classification Quarterly*, Vol. 7 No. 2, pp. 55-69.

Jett, M., Reuse, B. and Kessling, G. (1998), "Implementation of an online database for tables of contents of books", *Electronic Library*, Vol. 16 No. 2, pp. 123-30.

Mundgod, M. (1993), "Application of optical character recognition and expert system cataloguing: a state of the art report", *Guided Project II*, DRTC, Bangalore, pp. 35-62.

Omnipage (2001), *OmniPage Pro 10*, available at: www.caere.com/products/omnipage/pro/, Scansoft Inc., Peabody, MA.

Vanhelsuwe, L. *et al.* (1996), *Mastering Java*, BPB Publications, New Delhi, pp. 360-5.

Wyner, B.S. (1980), *Introduction to Cataloguing and Classification*, 6th ed, Libraries Unlimited, Littleton, CO, p. 640.