Automatic Subject Indexing and Classification Using Text Recognition and Computer-Based Analysis of Tables of Contents

Jan Pokorny

Automatic Subject Indexing and Classification Using Text Recognition and Computer-Based Analysis of Tables of Contents

Jan Pokorný

1 Book titles for the natural and technical sciences do not include information about all topics a book contains. For example, for a book entitled XML Technology: Principles and Applications in Practice, it is unclear whether a chapter on language XQuery is included or not. Even bibliographic records with subject headings cannot solve such problems because catalogers often do not have sufficient knowledge of particular scientific disciplines. Unlike titles, book tables of contents (ToCs) in the natural and technical sciences often have highly precise descriptions of topics covered in chapters and subchapters. This means it is possible to successfully harvest keywords from ToCs with high relevancy. Such keywords can then become reliable input data for subject indexing and subject classification in library catalogs and other search engines, giving users the ability to search all important topics covered anywhere in a book.

2 Library collections of print books are traditionally processed for library catalogs in the form of bibliographic records consisting of fields with descriptive data (author name, number of pages, language of publication, name of publisher) and subject data (keywords, subject headings, classification terms). Selected fields are indexed and can serve as access points or filters/facets in a catalog.

3 Users are usually looking for either a specific book in a catalog (e.g., a user looks for a book entitled Nanotechnology in biology and medicine: methods, devices, and applications by Tauna Vo-Dinh, 2007) or is looking for books on a certain topic (e.g., a user wants some books in the field of nanotechnology in medicine) [Lown & Sierra & Tito, 2013]. These are two completely different search strategies. In this article, we will look at a technology that addresses the second strategy, a search based on book subject.

Subject description in bibliographic records for scientific books

4 Although scientific book titles attempt to express a general topic or focus (what the book is about), they are unable to cover all topics and subtopics contained in a work. For example, the book title Selected Chapters in Nuclear Physics suggests that the main theme of the book is nuclear physics but we do not know what specific topics are included in the book. In this case, a user might refer to subject access data if a bibliographic record for the book is available.

5 Subject access data is traditionally created by a librarian-cataloger with a book in hand. A cataloger discovers what topics the book covers and performs so-called content analysis. Consequently, the cataloger must define clear keywords expressing the subject of the book as expected by a potential user. Finally, the cataloger has to find the appropriate terms from the existing classification systems supported by the library (most commonly, the Library of Congress Classification, the Dewey Decimal Classification, or the Universal Decimal Classification).

6 Content analysis performed by a universal cataloger can be problematic. First of all, there are economic aspects, because human labor is expensive. Another disadvantage is time delay: a cataloger is able to accurately describe about 3–4 books per hour [Read, 2003]. If the library buys many books at one time, it can take an unbearably long time for the books to be processed. In libraries with large numbers of books acquired, the delay may even be several weeks. This issue can be partially solved by sharing records between libraries, meaning previously-created records can be copied by other libraries. Records can also contain subject descriptions in many national languages which require translation by libraries. This, too, consumes time.

7 Another—and perhaps the most serious problem with subject description performed by catalogers—is the limited ability to understand and describe a subject. The quality of description depends on the intellectual assumptions of a cataloger and is always affected by some amount of subjectivity [Cloete *et al* 2003, Romero 1994]. When processing scientific books, a cataloger who is not an expert in a given discipline is not able to correctly and precisely understand and describe the topics contained in the book. This leads not only to inaccuracy of description but also often leads to completely wrong subject classification, with a cataloger putting a book into a different subject classification than an expert would choose [Kgosiemang 2005, Bowman 2006, Romero 1994]. Moreover, catalogers often do not have enough time to study each book in detail. As a result, end users are unable to find books that objectively match their search queries. Let's demonstrate the issue using an example.

8 A user searches for a book about how to work with a file system in the Python programming language. If the user searches for “python AND file system” in a title or subject index, they will not probably find any matching records. The reason for this is simple: a cataloger probably recognized that the book is about the Python programming language but not did not recognize that some chapter talks about the file system at a certain level of significance. As a result, the bibliographic record with subject access created by this fictional cataloger does not describe the book’s topics in enough depth and was not able to recognize nor identify specific topics.

9 As a result, searching by title or subject for scientific books likely leads to high error rates and a library catalog user probably does not find all the books relevant to specific topical interests. There are several ways libraries can solve this problem.

10 One way—unfortunately the most frequent solution in many libraries—is to ignore the problem. Universal catalogers create subject descriptions for library collections and the inaccuracy of description is simply tolerated.

11 Another way to address this problem would be to hire disciplinary specialists who are experts in a particular field. An expert in biology could likely produce fairly accurate descriptions of books from the biological sciences. Unfortunately, this solution is only possible in narrowly specialized professional libraries where the subject range of holdings is so small that the library can employ an appropriate number of catalogers who are also disciplinary experts. However, many general libraries such as national, regional, or large academic libraries build collections across so many scientific disciplines that they cannot afford to employ specially-oriented catalogers for all fields.

12 Another option is to use the folksonomy method, where end users enrich metadata with custom tags. However, this method requires a relatively high minimum number of users who are actively working with a system. This is a kind of crowd processing, with all related advantages and disadvantages [Porter, 2011].

13 Another alternative involves creating relevant subject descriptions without the involvement of expert catalogers. This is possible by implementing automatic recognition of book ToCs and extracting relevant keywords from them.

Keywords extracted from book tables of contents

14 Non-fiction books, in most cases, come equipped with rich information tools that allow users to quickly navigate to specific chapters or pages—namely, the alphabetical index and the ToC.

15 While the alphabetical index contains highly atomized and isolated terms that are too context-dependent to be automated harvesting, the ToC is ideal for outlining what each chapter in a book is about. The ToC consists of headings and subheadings for textual content (typically chapters) in the order in which they appear in the book and also marking the page where each part begins. Chapter headings are created by authors who know the topic and text perfectly and therefore should be able to accurately capture key topics for each chapter. In a scientific text, chapter headings consist of keywords carefully chosen by an author, normalized to nominative case with an emphasis on conciseness and simplicity. In sum, we can say that chapter headings represent chapter content. This applies to non-fiction books only—fiction, poetry, and other artistic books include chapter names created with a different intent rather than strictly summarizing content.

16 ToCs for scientific books thus become ideal data sources for extracting keywords that can be used to enrich subject access descriptions of books. This might seem similar to mining keywords from full-texts but is not actually the case. The basic advantages of using ToCs as data sources when compared to extracting keywords from full-texts can be summarized as follows:

1. Full-text of a book is not required; only a few ToC pages are necessary,

2. ToC items are presented in basic grammar forms,
3. Context for each item is clearly apparent due to the hierarchical relationships of the items (title of the work → section name → chapter name → subchapter name and so on),
4. ToCs contain links to page numbers where chapters begin and these can be used to determine keywords scaled according to how many pages is written about each.

17 The output of this method can be a set of keywords with a parameter that expresses the space an author devoted to each topic in the book. This parameter can be used for sorting search results by relevance.

Method

18 The following illustrative example provides detail about the method of extracting keywords from ToCs. Suppose there is a book we only have in print. The process can be split into following steps:

1. Scanning book pages containing ToCs
2. OCR with text block detection
3. Eliminating irrelevant text blocks
4. Resolving word and numeric blocks, removing stop words
5. Text analysis with focus on the context of ToC items
6. Keyword extraction
7. Subject classification assignment
8. Usage

Scanning book pages contained in tables of contents

19 At our library, we scan all pages where a ToC is present in a book. This depends on how detailed a ToC is and if it is spread over one or more pages. We often see that irrelevant text or some graphic objects are present on ToC pages. Such irrelevant elements will be removed with the following steps. When scanning ToCs, one must be aware that the primary purpose is to retrieve text using OCR. Therefore, it is necessary to check all input errors. Problems can be caused by a missing page, a bent corner, too solid bending, crooked text lines or a book scanned askew. Careful work reduces error rates in the subsequent steps. The result of scan is a bitmap object in the form of a graphic file such as TIFF or JPEG.

OCR with text block detection

20 Software converting images into text objects must be able to detect the blocks that contain text on a page. Since ToCs have a very diverse structures and layouts, different ways of splitting topics into chapters and subchapters, and different typographical and graphical designs, the software may identify some blocks incorrectly. Therefore, the software must have a capacity for human intervention and should be able to learn from new situations to work better and better.

Irrelevant text block elimination

21 Pages with ToCs often contain various irrelevant texts and graphic objects which need to be distinguished and removed from processing. For the next steps, we need to work only with text blocks that contain ToC items.

Resolution of word and numeric blocks, removal of stop words

22 ToCs usually list chapters with the name of each chapter and the page number where each chapter begins. In addition, the page number for the next chapter determines the range of pages that can be used to calculate a relevance scale or to display the importance of a chapter in thematic clouds. Therefore, the software must retrieve the name of each chapter and its location (pages) for each ToC. This can be complicated because of diverse layouts for ToCs on a page.

23 ToC item names often contain words that indicate sections and chapters only formally and do not contain any information about the content itself. Examples are “Chapter 2,” “Book first,” “Index,” or formal numbering such as “2.1.3.” These formal markers can help in the recognition of a ToC hierarchical tree structure but must be distinguished and removed for final processing.

Text analysis with focus on context of TOC items

24 This step is the most demanding in the whole process and puts a high demand on software. Scientific books typically use a multilevel structure for chapters in which subchapters are embedded into parent chapters, inheriting their context. For example, the name of a subchapter entitled “Paths and their description” can be interpreted in many ways. Its meaning is different for tourist travel books than for books on programming. However, when we put this subchapter in context with a parent chapter “Using file system,” we can interpret the meaning of the word “paths” more precisely—paths in a computer file system. If we extend this using another context, resulting from a book title Programming in Python, we can determine the meaning of the word “paths” exactly.

25 Software must be able to handle ToC items not only in an exact order but must also be able to distinguish item positions in ToC hierarchical trees in order to determine parent and child items as well as the extent of their embedding. Automatic detection of these relations is usually based on the layout of items on a page (embedded items are often indented) or by using typographic characteristics (embedded items often have smaller or thinner fonts than their parent items). The meaning of child items needs to be extended by utilizing context from all parent items in order to create definite and significant keywords.

Keyword extraction

26 Keywords harvested from ToC items can be optimized and transformed according to the requirements of particular cataloging policies, excluding duplicate items, and so on. If books written in more than one language are processed, automatic translations can also be performed at this point. The system can also remove keywords from chapters which are considered to be marginal (e.g., if a chapter is under 2 pages, it can be assumed that such a topic is presented marginally or inadequately).

Subject classification assignment

27 Finally, the system suggests terms for selected subject classification systems that best match each book. This functionality leverages a learning mechanism that tracks what keyword combinations have been applied to which classification terms. The system currently supports Universal Decimal Classification, Conspectus, and PSH. Others can be added.

Usage

28 The result of this process is a list of keywords that can serve as output for information systems, typically library systems or discovery systems. The list can be imported directly into selected bibliographic record fields subsequently indexed for searching or harvested for indexing in another system. Another interesting option is to store keyword lists in a shared repository so that other libraries can use them. Keywords can then be downloaded from the shared storage space using a unique identifier such as ISBN.

29 In conjunction with bibliographic records, keywords from ToCs can help in the identification of scientific and technical topics that occur most frequently for a certain time period. This can be used for a number of summary views; for example, to display a timeline where a user can clearly see how often a topic occurs in the literature, to show how a given keyword occurs over time in a variety of disciplines, or to trace the intersection of topics between disciplines. Another potential use is for the display of rankings for popular topics which be an attractive alternative to classical book newsletters.

BIBLIOGRAPHY

References

Bates, J. Marcia (1986). "Subject access in online catalogs: A design model." *Journal of the American Society for Information Science*, 37(6): 357–376.

Bowman, J.H. (2006). "Education and training for cataloguing and classification in the British Isles." *Cataloging and Classification Quarterly*, vol. 41 nos. 3–4: 309–33.

Choi, Youngok, Hsieh-Yee, Ingrid, Kules, Bill. (2007). "Retrieval effectiveness of table of contents and subject headings." In: *Proceedings of the 2007 conference on Digital libraries—JCDL'07* [online]., s. 103– [cit. 2018-01-09]. New York, USA: ACM Press. DOI: 10.1145/1255175.1255195. ISBN 9781595936448. Available from: http://portal.acm.org/citation.cfm?doid=1255175.1255195

Chung, EunKyung, Miksa, Shawne, Hastings, Samantha K. (2010). "A Framework of Automatic Subject Term Assignment for Text Categorization: An Indexing Conception-based Approach." *Journal of the American Society for Information Science and Technology*, 61, no. 4: 688–99.

Chung, Yi-Ming, Pottenger, William M., Schatz, Bruce R. (1998). *Automatic Subject Indexing Using an Associative Neural Network*. ACM DL.

Cloete, Linda M., Snyman, Retha; Cronje, J.C. (2003). "Training cataloguing students using a mix of media and technologies," *Aslib Proceedings*, vol. 55 no. 4: 223–33.

El-Haj, Mahmoud, Balkan, Lorna, Barbalet, Suzanne, Bell, Lucy, Shepherdson, John. (2003). "An Experiment in Automatic Indexing Using the HASSET Thesaurus." *Computer Science and Electronic Engineering Conference (CEEC)*, 13–18.

Golub, Koraljka, Hamon, Thierry, Ardö, Anders. (2007). "Automated Classification of Textual Documents based on a Controlled Vocabulary in Engineering." *Knowledge Organization*, 34, no. 4: 247–263.

Kgosiemang, Rose Tiny. (2005). "Education and training for cataloguing at the University of Botswana Library: an overview." *Cataloguing and Classification Quarterly*, vol. 41 no. 2: 5–25.

Lown, Cory, Sierra, Tito, Boyer, Josh. (2013). *How Users Search the Library from a Single Search Box*. College & Research Libraries, [S.l.], V. 74, n. 3: 227–241, May 2013. ISSN 2150-6701. Available at: < https://crl.acrl.org/index.php/crl/article/view/16303>. Date accessed: 27 Mar. 2018. doi:https://doi.org/10.5860/crl-321.

Maghsoodi, Nooshin, Homayounpour, Mohammad Mehdi. (2011). "Using Thesaurus to Improve Multiclass Text Classification." *Lecture Notes in Computer Science*, 6609: 244–253.

Morris, Ruth. (2001). "Online tables of contents for books: effect on usage." *Bulletin of the Medical Library Association* [online]. 2001, 89(1): 29–36 [cit. 2018-01-09]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC31701/

Peters, Isabella. (2009). *Indexing and retrieval in web 2.0*. Translated by Paul Becker. Berlin: Walter De Gruyter GmbH & Co.

Porter, John. (2011). "Folksonomies in the Library: their impact on user experience, and their implications for the work of librarians." *Australian Library*, vol. 60, no. 3: 248–258.

Read, Jane. (2003). "Cataloguing Without Tears: Managing Knowledge in the Information Society." Chandos Information Professional Series, *Elsevier*, 2003: 47–51. ISBN 1780630638.

Romero, Lisa. (1994). "An Analysis of Entry-Level Cataloging Errors: Implications for Instruction and Training." *Journal of Education for Library and Information Science.* 35 (3): 210–226.

Sebastiani, Fabrizio. (2002). "Machine Learning in Automated Text Categorization." *ACM Computing Surveys*, 34, no. 1: 1–47.

Stock, Wolfgang. (2007). "Folksonomies and science communication: A mash-up of professional science databases and Web 2.0 services." *Journal Information Services and Use* [online]. 2007, 27(3): 97–103 [cit. 2018-01-09]. Available from: https://dl.acm.org/citation.cfm?id=1370665

"Subject Indexing and Classification: 2002–2007." *Association for Library Collections & Technical Services (ALCTS)* [online]. Chicago: ALCTS, 2018 [cit. 2018-01-10]. Available from: http://www.ala.org/alcts/resources/org/cat/research/subjindclass07

ABSTRACTS

This paper will describe a method for machine-based creation of high quality subject indexing and classification for both electronic and print documents using tables of contents (ToCs). The technology described here is primarily focused on electronic and print documents for which, because of technical or licensing reasons, it is not possible to index full text. However, the technology would also be useful for full text documents, because it could significantly enhance the accuracy and relevance of subject description by analyzing the structure of ToCs.

INDEX

Keywords: text mining, computer-generated subject headings, computer-generated keywords, machine learning system, library automatization

AUTHOR

JAN POKORNÝ

National Library of Technology, Prague, Czechia
jan.pokorny@techlib.cz
(corresponding author)