Received 17 February 2015
Revised 1 May 2015
Accepted 15 July 2015

A method for automatic analysis Table of Contents in Chinese books

Jing Chen
School of Information Management, HuaZhong Normal University, Wuhan, China, and
Quan Lu
School of Information Management, Wuhan University, Wuhan, China

Abstract

Purpose – The purpose of this paper is to propose a novel method to analyze Table of Contents (TOC) in Chinese books automatically based on the hierarchy organization rules which gained by investigation.

Design/methodology/approach – This paper analyzed the main literature in this field first, then hierarchy organization rules of Chinese book TOC were generated and the method parsing TOC automatically based on these rules was proposed. A prototype system implementing the method was also developed. The method was evaluated through processing a corpus on the prototype system, and the results were checked with calculation of precision and recall.

Findings – The experiment result illustrated the superiority (extensive application, recall is 95.34 percent and precision is 94.44 percent) of the method.

Practical implications – The result can help Chinese libraries deal with electronic texts from four aspects. First, it can be used to complement or enhance current digitization and optical character recognition methods and cut the financial and labor cost of Chinese libraries. Second, it can help libraries to keep information on indexing words as well as chapters, sections and subsections in Chinese book databases, which ensures easy retrieval and extract any intended portion as demanded by user. Third, it helps to enrich the services and then enhances the user experiences in Chinese libraries. Fourth, it improves the specification and policy of digitalizing Chinese books.

Originality/value – The paper provided insight into the hierarchy organization of TOCs in Chinese books, the method based on the rules has extensive application than other methods. This method for Chinese book TOC automatic analysis is also as reference for English book TOC automatic analysis.

Keywords Digital libraries, Automatic analysis technologies, Chinese books, Document analysis, Hierarchical organization categories, Table of Contents

Paper type Research paper

Introduction

In recent years, the Chinese digital book resources and their applications are booming. Table of Contents (TOC) analysis has drawn attention nowadays because it is a collection of references to the different components of the document and naturally reflects the logical structure of the entire document (Gao *et al.*, 2010). In China's largest comprehensive dictionary, the *Cihai* (Xia and Chen, 2009), the function of TOC is described as "listing the structure condition, title, page number of books and periodicals." Lu (1971) in his dictionary of library science states that, "TOC is an

National Natural Science Foundation of China (project name: Research on Automatic Indexing for Book Hierarchical Topics, Project No. 71303089). Independent research Project of Huazhong Normal University (project name: Research on Automatic Analysis Table of Contents in Chinese Books, Project No. CCNU14A05050).

Library Hi Tech
Vol. 33 No. 3, 2015
pp. 424-438

0737-8831
DOI 10.1108/LHT-05-2015-0043

inherent part of book, the integrated embodiment of its subjects, structure and layout, which can reveal the author's views and theme of the book."

TOC analysis includes:

- Hierarchical logical marker recognition. A logical marker usually locates in initial position of a TOC entry, and also represents the hierarchical structure of the TOC entries, such as "第一章" (as "chapter one" in English) and "第一节" (as "section one" in English), which is often composed of sequence number (may also be combined with punctuations) and root word(s). For example, "第一章" (chapter one) is a logical marker in which "一" (one) is its sequence number and "章" (chapter) is its root word. Recognizing the logical markers and then the hierarchical relationship of these logical markers will make a TOC hierarchy certain.
- Title or description extraction of chapters, sections and subsections. The title or description of a chapter, section and subsection in TOC entry will always reappear in the text body of the book, so it is a guide to a specific part of the text body. It is the main body of a TOC entry and can be extracted by removing the logical marker, linked notation and page numbers in TOC entry.
- TOC may not only include some TOC entries referring to text body, but also some other TOC entries referring to accessories of book. Some of them usually locate in the beginning of TOC, such as "前言" (foreword), "绪论" (introduction), "引言" (similar to foreword), "导言" (similar to introduction) and "序" (preface), and some usually located in the end of the TOC, such as "参考文献" (reference), "后记" (postscript) and "附录" (appendix). However, they often do not have any logical maker to indicate their hierarchy, so it is hard to analyze them. Fortunately, these issues often have relatively fixed usage and they can be look as the first level in TOC hierarchy.

In other words, the aim of TOC analysis is to get the title of all book parts and to get the hierarchical structure of these parts, so user can understand the content distribution. Furthermore, user can analyze every part of the text as his/her wish when TOC analysis is applied to book processing. TOC automatic analysis has lots of benefits, such as it is helpful to better services like TOC enrichment service with easy navigation, fast retrieval and fine-grained analysis into books, it is also necessary for the document databases in libraries to store information hierarchically for easy retrieval of any intended portions, and so on.

However, Chinese books present TOC in various forms, which seriously hindered the application of TOC automatic analysis technologies to Chinese books. The effective management and utilization of Chinese books is still far from implementation. Therefore, the study of Chinese books will raise the following question: how to analyze TOCs automatically in Chinese books?

This research attempts to address this problem by proposing a practical method based on TOC hierarchy organization rules. The authors first investigate TOCs in Chinese books that are stratified and randomly sampled from books stored in the Huazhong Normal University (HNU) library. Based on the data of 920 TOCs, this paper summarizes the markers and their combinations, builds hierarchy organization rules by open coding method, then TOCs can be analyzed top-down based on these rules. Our method is implemented in JAVA programming language. Finally in our experiment, we check this method with calculation of precision and recall.

Review of related literature

A number of methods on TOC automatic analysis have been proposed in recent years. Automatic analysis technologies of TOCs can be cataloged into three kinds: layout based, text information based and joint with text body. Layout of TOC includes content elements' location on the page, font size, word spacing and row spacing, etc. Text information-based method can be cataloged into two kinds further: text semantics based and hierarchical logical marker based. Text information-based method adopts natural language understanding technologies to recognize and extract hierarchical structure of the TOC, and afterwards parse the TOC. Hierarchical logical marker refers to the prefix label in the header information of each logical entry in a Chinese book, such as "第一章" (chapter one) and "第1节" (section 1), which is often composed of sequence number (may also be combined with punctuations) and root word(s), and used to express the hierarchical logical structure of TOC. Joint with text body method match the contents of a reference in TOC pages to the title page in text body, for the content of a reference in TOC pages will repeat on the title page.

In principle, the layout-based technologies adopt optical character recognition (OCR) technology to digitalize the TOC of a book or a paper, and then analyze the digitalized TOC based on layout information. Similar with the research of Tsuruoka *et al.* (2001) and Mandal *et al.* (2003) in English documents, Sun and Su (2004) made use of the rules of indentation in TOCs of Chinese books, and developed an algorithm to digitalize TOCs of Chinese books based on OCR technology and indentation analysis. Gao *et al.* (2010) noticed the style consistence phenomenon of TOCs of Chinese books, and put forward a Chinese book TOC recognition method by detecting decorative elements in the TOC based on clustering techniques. Sarkar and Saund (2008) enumerated various graphical and perceptual cues that provide cues to TOC parsing. Wu *et al.* (2013) classify TOCs into three basic styles, namely "flat," "ordered" according to the space information of blocks and "divided," and propose TOC parsing rules based on the classification of these three TOC styles.

However, Ma (1995) concluded the common errors in Chinese document typography, and pointed out that, layout errors (including font, size and/or spacing errors) are the most common ones. Therefore, the layout-based technologies are limited by low accuracy under the status quo of variable layout quality of Chinese book TOCs.

According to the research of Ma (1995), one important advantage of text information-based technologies is that avoiding the layout errors' influences that layout-based technologies cannot. Abdel (2001) proposed a semantic labeling approach to delimit articles using the contextual rules of part-of-speech tagging. In recent years, several preliminary hierarchical logical marker-based technologies has been developed. Chen and Ding (2002) hypothetically set some rules to describe and reason the hierarchical logical categories in TOCs of Chinese books, and then proposed a book document logical structure extraction method based on these rules. Similar with the idea of Abdel (2001) in English documents processing, Zhang *et al.* (2005) further studied the optimized hierarchy clustering-based extraction algorithm for logical document structures of semi-structured Chinese documents, which requires a predefined hierarchical logical markers set to support the clustering analysis. He *et al.* (2004) combined geometrical rules (indentations) and semantic rules (typical text sequences identifying chapters and sections) to extract the hierarchical structure in Chinese books.

Joint with text body method tend to be more robust, but they are usually time consuming because the computational complexity required by text matching is quadratic to the number of text blocks. For example, Lin and Xiong (2006) and Herve

Dejean and Meunier (2009) independently used text matching between TOC candidate pages and body pages for detecting TOC pages in a document. Dresevic *et al.* (2009) defined a TOC entry to be a single smallest group of words with the same title target somewhere in the book. Marinai *et al.* (2010) also use text similarity measuring techniques to detect TOC entries.

Our method relies on hierarchical logical marker to build the rules for recognize TOC automatically. Although our method is somehow similar with Chen's (Chen and Ding, 2002) , the rules of Chen and Ding (2002) are so limited and implicit that they are not so practical.

Proposed solution

We have observed that TOC entries belonging to the same level share the same logical marker, such as "章" (chapter) and "节" (section), and the sequence number shows the order of the TOC entry in the hierarchy. We call this "TOC intrinsic logical consistency." Although different book may adopt different logical marker and different hierarchical organization rule to indicate the hierarchy of TOC entries, Cao and Wang (2000) states that the TOC hierarchical organization rules in Chinese books are converging. Convergence is a vital characteristic so that hierarchy organization rules can be obtained by quantitative analysis method, and then they can be used to recognize the hierarchy of TOC entry. This method is usually effective and applicable because the "TOC intrinsic logical consistency" reflects the common logical practice and the rules can be quickly captured.

This method is especially suitable for book documents, which usually have multiple pages and contain a large number of TOC entries. We generated the TOC logic hierarchy organization rules by investigation and open coding technique which belonging to Grounded Theory Approach (Wikipedia, 2014). Then we used the rules to analyze other TOCs in Chinese books.

Generation of TOC hierarchy organization rules

Chen and Ding (2002) has set some rules to describe and reason the hierarchical logical categories in TOCs, however, their rules are not so practical. Cao and Wang (2000) states that the TOC hierarchy organization rules in Chinese books are converging. So the rules can be generated through qualitative analysis method. We use open coding technology to get the rules.

Research sample selection

The authors chose more than 500,000 Chinese paper books collected in HNU library as the population for paper books with long history and with more issues than digital books. The authors chose categories in *Chinese Library Classification* (5th edition) compiled by NLC (2010) as stratification variables to organize the population. The population covers all the first-level disciplines of conferring academic degrees in *Chinese Library Classification* (5th edition).

Then the authors stratified randomly sampled Chinese paper books in HNU library from A to Z categories of the *Chinese Library Classification* (5th edition), and got 920 Chinese paper books covering all A-Z categories and publishing dates ranging from 1957 through 2013 as samples. These sampled books were published by 156 different publishers such as Science Press (China), China Higher Education Press and Commercial Press. The data collection time window is between August 12 and September 16, 2013.

Gathering data

The authors makes a comprehensive investigation of the hierarchical organization categories of TOC' in Chinese books, to generate TOC analysis rules and offer more effective automatic analysis technologies of Chinese book TOCs. So the authors manually extracted the following information from each sampled book:

- Hierarchical organization information of the TOC in a sample book, including: the maximal number of layer in the hierarchical organized TOC or in another word, the depth of the TOC tree, the hierarchical logical marker of each layer in the TOC, the expression modes of accessories of book. The information is used to generate the hierarchical organization rules.
- Basic information of a sample book, including: title of the book, publisher's name, International Standard Serial Number and year of publication. The information is basic information of every book and can be used to analysis the development trend of hierarchical organization of TOC.

We also investigate whether the hierarchical organization in the TOC is completely consistent with that in the headers of the text part. The information is used to valid whether researchers can use TOC to direct the segmentation and analysis of the whole book text.

Open coding the logical marker rules in TOC's hierarchical organization

First, from the 920 samples we extracted 920 marker vectors such as "第一部分" (part one), n第1章" (chapter 1), "第1节" (section 1) according to the specific-circumstances of each layer of each TOC, and then only kept those vectors different with each other for open coding in the grounded theory.

The hierarchical logical markers having been used in TOCs vary from layer to layer, so we studied the categories of hierarchical logical marker in all layers, respectively, as shown in Table I.

Although the maximal number of layers in the TOC could be theoretically relatively high, some TOCs only have the first layer, some TOCs only have the first and the second layers, more TOCs have the first to the third layers. The proportion of TOCs having four or more layers is extremely small (only 0.06 percent of 920), and the fourth or higher layer is much less important in book management and utilization than the first three layers, hence we ignored the fourth and above layers in our following analysis. So we analyze the hierarchical organization modes of TOC's based on the combination relationship of every two adjacent layers' markers, such as the

Table I. Types of hierarchical logical marker in TOCs of Chinese books

Layer in the TOC	Types of hierarchical logical marker
1st	"第*章" (chapter*), "第*篇" (discourse*), "第*部分" (part*), "第*编" (series*), "上" (first) or "中" (midst) or "下" (last)+"篇" (discourse) or "编" (series); Chinese numeral, uppercase Chinese numeral, Arabia numeral; "Chapter," "part," "册" (book), "卷" (volume), "单元" (unit), "辑" (division)
2nd	"第*节" (section*), "第*章" (chapter*), "Chapter," "第*篇" (discourse*); Chinese numeral, Arabia numeral, "§"+numeral, numeral+"."+numeral, numeral +"–"+numeral, numeral in brackets
3rd	"第*节" (section*); Chinese numeral, Arabia numeral, "§"+numeral, numeral +"."+numeral, numeral+"–"+numeral, Roman numeral, numeral in brackets

combination of layer 1 marker "第一部分" (part one) and layer 2 marker "第1章" (chapter 1). Furthermore, the combination relationship of every two adjacent layers' markers is an easier object to mine simple and clear rules, just like each link in a chain, to explicate the myriads of changes of hierarchical organization of TOC's in an interlocking way.

Our specific approach was to code the combinations of every two adjacent layers' markers in the marker vectors according to the following principles:

(1) Merging by synonymity. In this research, two or more markers have the relationship of synonymity means they have the same meaning of natural Chinese language but never appear in the same marker vector, so they are called synonyms of each other. Merging by synonymity means to merge those combinations of two adjacent layers' markers having synonymous markers in the same location into one combination. For example, the marker "第一章" (chapter one) with the marker "第1章" (chapter one), "第一章" (chapter one) with "Chapter 1," and "第一编" (series one) with "第1篇" (discourse 1), are all instances of synonymity in this research.

(2) Dividing by co-occurrence, as a patch to the first one. Two or more markers have the relationship of co-occurrence means they can appear and have appeared in the same vector. If two or more markers have the relationship of co-occurrence, then even if they have the same meaning of natural Chinese language, to avoid misunderstanding or confusion, their corresponding combinations and/or vectors should not be merged into only one, but should be kept as different categories. For example, although "第一篇" (discourse one) and "第一章" (chapter one) have very similar meaning in natural Chinese language, they can combine with each other like subcategory 3.1 in Figure 2, so "第一篇" (discourse one) and "第一章" (chapter one) should be kept as different categories. More typically, although Chinese numeral "一" (one) and Arabia numeral "1" have almost the same meaning in natural Chinese language, they can form subcategory like ("一"(one), "1"), so they are not synonyms of each other in this research.

According to the above idea and principles, the authors concluded four synonymous root word groups in Chinese TOC's hierarchical organization, as shown in Table II.

Used interchangeably, the root words in the same group in Table II are completely equivalent to each other in Chinese TOC's hierarchical organization. In addition, any forms of the sequence number are completely equivalent to each other in one group when being combined with root words to form a marker. For example, "第一章" (chapter one) is a synonym of " chapter 1," "第一编" (series one) is a synonym of "上篇" (first discourse) and "第一节" (section one) is a synonym of "第1节" (section 1).

Group number	Synonymous root words
1	部分 (part), Part, 卷 (volume),单元 (unit), 辑 (division)
2	章 (chapter), chapter
3	篇 (discourse), 编 (series)
4	节 (section), section

Table II.
Synonymous root word groups

Numerals can be used as both sequence numbers and logic mark in Chinese book TOC markers. The authors also study the modes to use numerals in markers and partial ordering relation between deformations of numerals for better analysis TOC hierarchy. The arrows in Figure 1 indicate the direction of decreasing priority. For example, Chinese numeral "一" (one) always appears in a higher priority header than Arabia numeral "1" if they both exist in Chinese book TOC. Figure 1 also revealed the underlying function of punctuations to adjust numerals' priority in TOC markers of Chinese books.

TOC may not only includes some TOC entries referring to text body, but also some other TOC entries referring to accessories of book, such as "序" (preface),"前言" (foreword),"绪论" (introduction),"参考文献" (reference) and "附录" (appendix). Some of them usually locate in the beginning of TOC, such as "序" (preface),"前言" (foreword),"绪论" (introduction), and some usually located in the end of the TOC, such as "参考文献" (reference),"后记" (postscript) and "附录" (appendix). However, they are often without logic maker to indicate their hierarchy. But they should be look as the first layer ones for they are independent parts of book. In our method, we treat the TOC entries which without logic maker as first layer ones.

Open coding the categories of TOC's hierarchical organization

For better understanding hierarchy organization of Chinese books, the authors classified TOC's hierarchical organization of Chinese book TOC into six categories and 19 subcategories, through open coding the hierarchical organization from 920 samples, just as shown in Figure 2. The categories and subcategories shown in Figure 2 are a subset of the logical markers' free combination of every two adjacent layers based on the above logical marker rules.

There are five categories and 17 subcategories of TOC's hierarchical organization combining layers 1 and 2 according this research, just as category 1 to category 5 and all their subcategories shown in Figure 2. For example, subcategory 3.3 means the corresponding TOC uses "第一篇"(discourse one) or its synonym as first layer marker and "第一节"(section one) or its synonym as second layer marker.

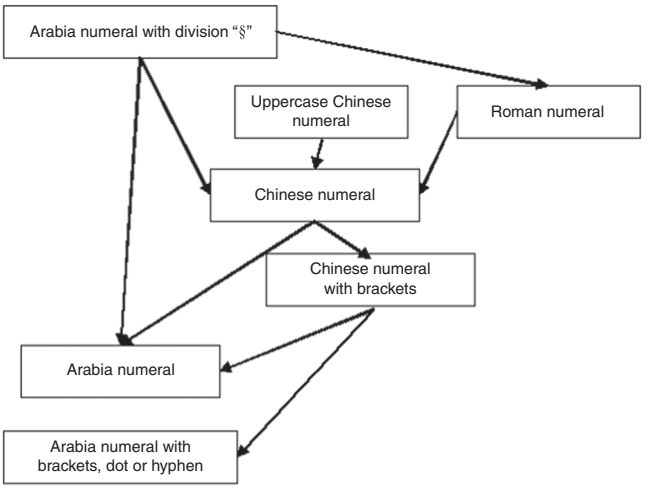


Figure 1.
The partial ordering relation between deformations of numerals

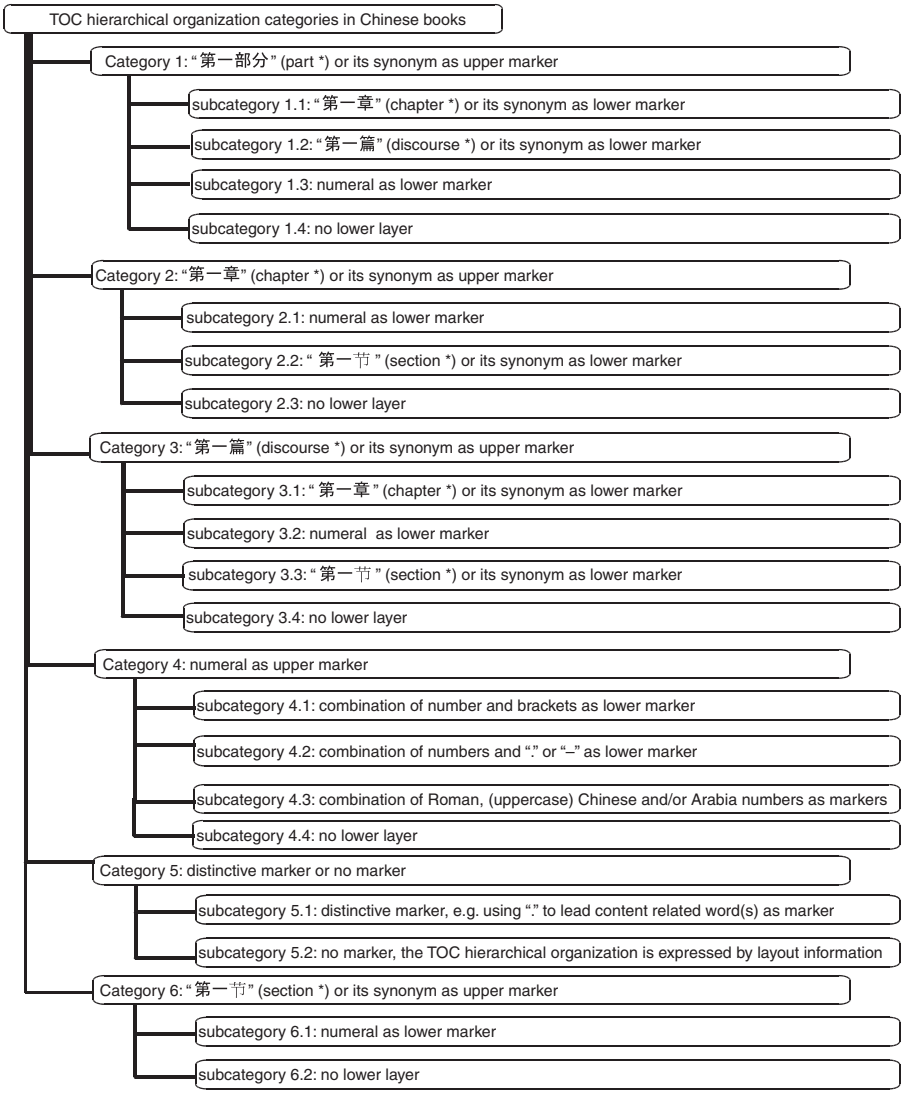


Figure 2.
TOC hierarchical organization categories in Chinese books

Five categories and 15 subcategories of TOC's hierarchical organization combining layer 2 and 3 according this research from 920 samples. The results are listed as from category 2 through category 6 and all their subcategories shown in Figure 2.

Then, the authors observed high similarity between the categories of layers 2 and 3 and the categories of layers 1 and 2. Specifically, they share four categories and 13 subcategories, while each has a private category. Further, the authors found that is because they share most markers and combinations while there are two exceptions. First "**第一部分**"(part one) or its synonym never appears as marker in layers 2 or 3, so category 1 has nothing to do with layers 2 and 3. Second, on the contrary, "**第一节**"(section one) or its synonym never appears as marker in layer 1, so category 6 is not related to combination of layers 1 and 2.

However, category 5 shows that some other Chinese authors prefer unique or personalized TOCs. There do exist some very special and not commonly used markers in subcategory 5.1. For example, a sample weirdly uses "." to lead content-related word (s) as first layer marker, and uses Chinese numeral as second layer marker. Moreover, samples in subcategory 5.2 do not have any marker, their TOC hierarchical structure is expressed by layout information including font, spacing and location. For example, a sample boldfaces the first layer heading and uses Song font in the second layer heading. Some samples unindent the first layer heading and indent the second layer heading.

Since TOC in category 5 can hardly be processed by automatic analysis technologies because of their characteristics of uniqueness and personalization, so the proportion and the trend of category 5 will be analysis to explore the practicability of automatic analysis of TOC. The authors took an analysis on sample distribution by the year of publication range, and find something interesting. The ratio of category 5 appears in layers 1 and 2, gradually increased from 14.3 percent in 1957-1969 to a peak of 25.6 percent before 2000, but then continuously decreases to 7.5 percent till 2010-2013. The peak point is consistent in time with the beginning of TOC enrichment service in China libraries, raising the awareness of importance of normalizing the TOC to book digitization. So broad category 5, which is difficult for automatic analysis, gradually converge in a declining trend. By analyze the sample data, the good news is, category 5 hardly appears in layers 2 and 3 of the TOC (only 0.6 percent).

We also investigate consistency of the hierarchical organization between the TOC and the text. The examination of the hierarchical organization between the TOC and the text of 920 samples turned out that, for the hierarchical logical structure, all samples have full consistency between the TOC and the text; and for the hierarchical logical marker, only one sample is inconsistent with its text in layer 1 (0.19 percent of the population) or layer 2 (0.24 percent of the 416 samples having layer 2), but no inconsistency occurs in layer 3. So there is very high consistency of the hierarchical organization between the TOC and the text in Chinese books, and the authors suggested to apply hierarchical organization information from automatic analysis of TOCs to Chinese book applications such as automatic indexing for book hierarchical topics.

Through above investigation and analysis, the authors could anticipate a high accuracy of automatic analysis of TOCs in Chinese books based on our method. By the way, those necessary manual interventions should be focussed to category 5 of layers 1 and 2. But, the convergence characteristics in the evolution of the categories of layers 1 and 2 predict a better environment for the automatic analysis technologies, so the accuracy rate will continue to improve along with it in the future.

Recognizing the hierarchy of TOC entry based on hierarchy organization rules

The basic technologies for automatically analyzing Chinese book TOCs based on the above categories are natural language understanding and automatic classification, which have been widely studied and are qualified for processing the markers and the categories except category 5, and the technology researches on processing Chinese book TOC. Chen and Ding (2002) and Zhang *et al.* (2005) propose application templates, making it highly feasible for us to apply our method and categories to automatic analysis technologies.

The above rules were collected into two rule bases. One is a marker rule base, which consists of the types, synonyms and partial orders of markers, including expressions of preface and appendix. The other one is a category rule base, which contains hierarchy organization rules mapping a combination of two adjacent layers' markers to a specific category, and all hierarchy organization rules follow the basic form of (upper marker, lower marker, category).

Based on these two bases, the following steps are carried out to recognize the hierarchy of the TOC entries in a Chinese book:

- Step 1: read every TOC entry in the TOC and extract the first up to four words before the first spacing as the phrase to be processed.
- Step 2: mark the first few entries as "foreword" if their first word lies in those words expressing foreword in the marker rule base, such as "前言" (foreword),"绪论" (introduction),"引言" (similar to foreword),"导言" (similar to introduction) and "序" (preface).
- Step 3: for the following phrases, extract their logical markers by comparing each phrase to hierarchical logical markers and expressions of appendix in the marker rule base.
- Step 4: after all logical markers are extracted, explore the hierarchy of TOC entries by their sequence in the TOC and the priority rules of markers in the marker rule base, starting from setting the first marker after "foreword" as first-level marker. The priority rules here consist of layer in the TOC and partial orders. If any corresponding priority rule is violated, then the foreword part is thought to have lower level, so the phrase just after the "foreword" is marked as its lower level, and the exploring is altered to set the next different marker is set as first-level marker. By iterating this process to every two adjacent layers in the TOC, the hierarchy of TOC entries are recognized. The appendix is recognized like that in step 2. In addition, the preface and the appendix are all set as the first level in the hierarchy.
- Step 5: to analyze the explored hierarchy a step further, classify every two adjacent layers in the TOC into one of the hierarchy organization categories, by applying forward reasoning in the category rule base to every two adjacent layer markers in the hierarchy.

After recognition the hierarchy of the TOC entries, extracting titles by removing the logical marker, linked notation and page numbers in the entries.

Prototype system to automatic analyzing TOCs in Chinese books

The authors implemented the approach we mentioned above in a JAVA-based prototype system to automatic analyzing TOCs in Chinese books which is called CBTOC Hierarchier. CBTOC Hierarchier adopts TXT files to implement the marker rule base and the category rule base, which makes them easy to be maintained. For example, it is easy to distinguish category rule {"1," "1.1," "4.2"} between {"1," "1-1," "4.2"} and add new rules. The recognization of markers, hierarchy and categories are implemented as three interdependent modules to gain higher flexibility.

CBTOC Hierarchier can read TXT format TOC files from Chinese books, extract the markers and hierarchy of all the TOCs and analyze their hierarchy organization categories automatically and respectively, then display them in a GUI window as shown in Figure 3. Interactive analysis of customized TOCs and browsing are also supported.

Notes: ① Category classification results; ② tree view of recognized TOC hierarchies; ③ the result of category classification for book named “东方文化史” (The history of oriental culture): category 2.2 in layer 1 and 2 and category 6.1 in layer 2 and 3; ④ the corresponding tree view of recognized TOC hierarchy for the book named “东方文化史 ” (The history of oriental culture)

Figure 3.
Interface of CBTOC Hierarchier

In Figure 3, the left window of CBTOC Hierarchier lists all TOC category classification results with each TOC’s result as a row. Every TOC file name is shown in the head of every row, then, category of layers 1 and 2 and category of layers 2 and 3 in the TOC file are shown after the colon. Click every TOC’s category classification result, the right window will show its corresponding tree view of recognized TOC hierarchy for the book. For example, the first row in left window in Figure 3 means, the book named “东方文化史” (The history of oriental culture) belongs to category 2.2 in layers 1 and 2 and category 6.1 in layers 2 and 3. Then click it, the right window details its logical markers, entry titles and hierarchy of every TOC, in a tree style structure.

Experiment and analysis

The authors checked the recall and precision of our approach by experiment of automatic analyzing TOCs in Chinese books on CBTOC Hierarchier.

The authors stratified randomly sampled Chinese electronic books in Chinese digital library electronic book database from A to Z categories of the Chinese Library Classification (5th edition), and got 236 Chinese electronic books covering all A-Z categories. Chinese digital library electronic book database is one of the most authoritative electronic book databases in Chinese, and it is source database of China Machine-Readable Catalogue. Then we extracted TOCs from these sample books in PDF format and save them as TXT format. So the inputs of our experiment were 236 TOCs in TXT format. Running time of CBTOC Hierarchier to process them was 3.641 seconds. Outputs of CBTOC Hierarchier in this experiment are shown as Figure 3.

We measured the recall and precision of TOC analysis in our experiment. Recall is the fraction of analyzable TOC, which means 1 minus the fraction of TOC in category 5. Precision is the fraction of accurate TOC entries which are analyzable. Precision is measured by analysis precision of TOC entries with logical marker and analysis precision of TOC without logical marker.

After analyzing the experiment result, 11 TOCs cannot be analyzed by our method, including six TOCs in our TOC data set having no logical marker at all, which should be analyzed manually, and five TOCs having logic hierarchy rules which is different from our logic hierarchy rules. So recall of our method is 95.34 percent.

We also calculated the precision of our method. There are two kinds of precision, one is the precision of analysis for TOCs with logical marker, and the other is the precision

of analysis for TOCs without logical marker. We calculated two kinds of logical hierarchy analysis precision here.

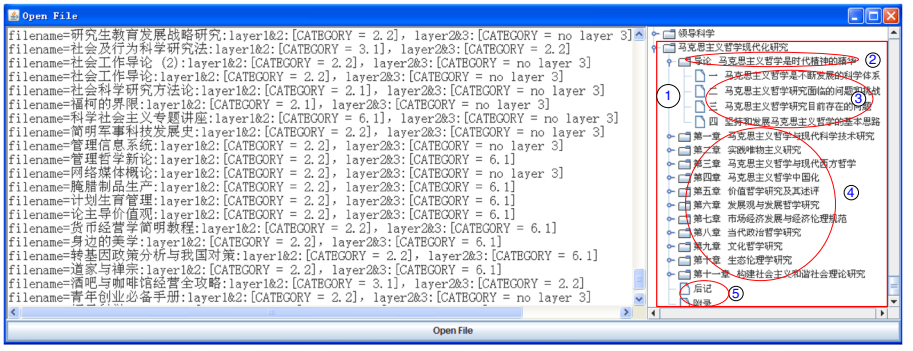
There are 19,947 TOC entries in the 225 analyzable TOCs, which include 18,903 TOC entries with logical marker and 1,044 TOC entries without logical marker. With our method, 18,382 of the 19,947 TOC entries with logical marker can be analysis precisely. So the precision for TOC entries with logical marker is 97.24 percent. In total, 455 of the 1,044 TOC entries without logical marker can be analyzed precisely, so the precision for TOC entries without logical marker is 43.58 percent. The whole precision of our method is 94.44 percent. F1 score is 94.9 percent.

The errors occurred mainly because some TOC entries miss their necessary logical marker. For TOC entries without logical marker, if they are lower-level TOC entries, for their logical marker missing, our method also treats them as first-level ones.

There are 668 TOC entries should have their logical markers but they did not. Especially for TOC entries without logical marker, when we ignore the TOC entries which missing their necessary logical markers, the precision will up to 90.88 percent. Compare to our earlier 43.58 percent, the precision is improved greatly. Then the whole precision of our method is 95.93 percent. F1 score is 95.6 percent.

It is exciting that our method can analyze hierarchy level of some TOC entries which missing logical marker. Such TOC entries have the common characteristic is that they are referring to accessories of book and with relatively fixed usage. For example, in the book which named "马克思主义哲学现代化研究" (*Modern Research on Marx's Philosophy*), there is "导论" (introduction) in TOC, there are some lower-level TOC entries in "导论" (introduction), and there is also "后记" (postscript) and "附录" (appendix) in TOC, our method can parsing them accurately, because rules learned from later part (like ④ in Figure 4) of the TOC can feedback and be used to correct the error occurred in the previous part (like ② and ③ in Figure 4) and our method treat the TOC entry without logical marker as first-level entry. The Tree view of recognized TOC hierarchy of the book is shown in Figure 4.

Comparing the hypothetically set rules of Chen and Ding (2002), the research of Zhang *et al.* (2005) and our results, we can see Chen and Ding (2002) only dealt with



Notes: ① The corresponding tree view of recognized TOC hierarchy for the book named "马克思主义哲学现代化研究" (Modern research on Marx's Philosophy); ② "导论" (introduction) in TOC of the book which without logical marker; ③ lower level TOC entries in "导论" (introduction); ④ TOC hierarchy tree view of text body; ⑤ "后记" (postscript) and "附录" (appendix) in TOC which without logical marker also

Figure 4.
Analysis hierarchy level of TOC entries which without logical marker and have low-level TOC entries

subcategories under category 4, while Zhang *et al.* (2005) assumed books have just one layer in the TOC and only dealt with subcategories 4.4 and 6.2. So their assumptions aim at some specific categories, but ignore the other categories, which is a flaw in their researches.

Discussion

The significant contribution of this study lies in that, it provides insight into the hierarchy organization of TOCs in Chinese books, the rules of hierarchy organization are effectively revealed, and the method based on the rules has extensive application than existing methods. And, the rules and analysis method for Chinese book TOC in this research is also as reference for English book TOC automatic analysis. Based on this study, practical implications are discussed. It can helps Chinese libraries deal with electronic texts from four aspects:

(1) First, it can be used to complement or enhance current digitization and OCR methods and cut the financial and labor cost of Chinese libraries. Digitization and OCR technologies are constantly increasing the number of digitalized books in Chinese libraries, but current digitization and OCR methods are too rough and inefficient for them. For example, because the huge demand for labor and financial resources to recognize TOC in Chinese book with existing digitization and OCR methods, the implement of TOC enrichment service in China libraries is difficult (Yu and Liu, 2012). Our study can automatic recognize and extract TOC in digitalized Chinese books with high accuracy, so it can complement or enhance current digitization and OCR methods applied in Chinese libraries. Applying it to Chinese libraries will observably reduce the workload of manual intervention in digitalizing book resources, so it can cut down the financial and labor cost in Chinese libraries.

(2) Second, it can help libraries to keep information on indexing words as well as chapters, sections and subsections in Chinese book databases, which ensures easy retrieval and extract any intended portion as demanded by user. There is a severe contradiction between the need for TOC description for Chinese books and the cost and efficiency of metadata production in Chinese libraries, and high transforming costs and lack of related standards are two existing limitations in describing the TOCs in Chinese books (Zhu, 2011). The method proposed in our study is flexible and universal to analyze TOC from various publishers or writers with high accuracy. So libraries can use it to describe TOC and hierarchically organize information in Chinese books efficiently and accurately, and support query needs at chapter, section and subsection level.

(3) Third, it helps to enrich the services and then enhances the user experiences in Chinese libraries. Most Chinese libraries treat a Chinese book as just a record unit or a sequence of characters without composite structure. But, because many Chinese books have more than 300,000 words, and Chinese owns far greater amount of information per character than alphabet like English, efficient services inside a digital Chinese book is becoming a vital challenge for Chinese libraries. Few Chinese libraries provide within book services like TOC enrichment service so far and librarians are aware of the gap between within book services and inaccurate TOC processing technologies. Our study provides a method to effectively analyze and organize the markers, hierarchy and hierarchy

organization categories of TOC in Chinese book, this will narrow the gap and bring new thoughts and bases for services like easier navigation, efficient retrieval and fine-grained interactive analysis. An example in English book is that, Lu *et al.* (2014) use a series of concentric circles to visualize chapter, section and paragraph in one book's hierarchy to aid user browse and view the book.

(4) Fourth, it improves the specification and policy of digitalizing Chinese books. Most Chinese libraries provide the way of metadata and downloading for viewing full text of digitalized Chinese books. And most of the Chinese books do not have a structured TOC hierarchy, while only a few books have specially processed TOC hierarchy that can only understand by corresponding specially designed reader. This makes it inefficient and inconvenient to use the electronic books because of the download and install tasks and the large amount of texts to be read. Generally, our study will help to solve the dilemma and develop new standards and policies of hierarchically digitalizing, organizing and displaying Chinese books. Furthermore, hierarchical and fine-grained book services derived from this research will change the traditional way of using Chinese books and in turn the traditional way to sale Chinese books.

Conclusion

Digitized Chinese books are becoming increasingly an important source of knowledge in China. But various forms of TOCs in Chinese books have seriously hindered the application of TOC automatic analysis technologies to digitalize Chinese books. The purpose of this study is to build a practical method based on TOC hierarchy organization rules from investigating TOCs in Chinese books. The findings reveal that there is strong regularity about the markers and categories in TOCs in Chinese books. The markers falls into a relatively fixed word set, and the categories can be merged into six categories and 19 subcategories from every two adjacent layers' combination. With these rules, TOC entries can be analyzed top-down automatically. The experiment results on our prototype system show extensive application, high recall and precision of our method for TOC automatic analysis. This means libraries can use this method to improve their information organization and services on Chinese books.

References

Abdel, B. (2001), "Recognition of table of contents for electronic library consulting", *International Journal on Document Analysis and Recognition*, Vol. 4 No. 1, pp. 35-45.

Cao, N. and Wang, D. (2000), "Research on content information description of Chinese book", *Journal of The National Library of China*, Vol. 9 No. 2, pp. 26-31 (in Chinese).

Chen, G.G. and Ding, X.Q. (2002), "A rule-based book document logical structure extraction method", *Computer Engineering and Applications*, Vol. 38 No. 19, pp. 53-57, 143 (in Chinese).

Dejean, H. and Meunier, J.-L. (2009), "On tables of contents and how to recognize them", *International Journal on Document Analysis and Recognition*, Vol. 12 No. 1, pp. 1-20.

Dresevic, B., Uzelac, A., Radakovic, B. and Todic, N. (2009), "Book layout analysis: TOC structure extraction engine", 7th International Workshop of the Initiative for the Evaluation of XML Retrieval, Dagstuhl Castle, pp. 164-171.

Gao, L.C., Tang, Z., Lin, X.F., Yu, Y.Y. and Fang, J. (2010), "A table of content recognition method of book documents based on clustering techniques", *Acta Scientiarum Naturalium Universitatis Pekinensis*, Vol. 46 No. 4, pp. 531-538 (in Chinese).

He, F., Ding, X. and Peng, L. (2004), "Hierarchical logical structure extraction of book documents by analyzing tables of contents", *Proceedings of the SPIE Conference on Document Recognition and Retrieval XI, San Jose, CA*, pp. 6-13.

Lin, X.F. and Xiong, Y. (2006), "Detection and analysis of table of contents based on content association", *International Journal on Document Analysis and Recognition*, Vol. 18 Nos 2-3, pp. 132-143.

Lu, Z.J. (1971), *Tu shu xue da ci dian*, Taiwan shang wu, Taibei (in Chinese).

Lu, Q., Liu, G. and Chen, J. (2014), "Integrating PDF interface into java application", *Library Hi Tech*, Vol. 32 No. 3, pp. 495-508.

Ma, Y.X. (1995), "Common error analysis of computer typesetting", *Journal of Yangzhou Teachers College (Natural Sciences)*, Vol. 15 No. 3, pp. 66-70 (in Chinese).

Mandal, S., Chowdhury, S.P., Das, A.K. and Chanda, B. (2003), "Automated detection and segmentation of table of contents page from document images", *Seventh International Conference on Document Analysis and Recognition*, pp. 398-402.

Marinai, S., Marino, E. and Soda, G. (2010), "Table of contents recognition for converting PDF documents in e-book formats", *Proceedings of the 10th ACM Symposium on Document Engineering, September*, pp. 73-76.

NLC (2010), "Chinese library classification", available at: http://clc.nlc.gov.cn/ (accessed August 2, 2014).

Sarkar, P. and Saund, E. (2008), "On the Reading of Tables of Contents. Document Analysis Systems", The Eighth IAPR International Workshop on Document Analysis Systems (DAS'08), Vol. 6 No. 4, pp. 386-393.

Sun, P. and Su, D.C. (2004), "An algorithm to automatically generate tables of contents of e-book based on OCR", *Modern Information*, Vol. 24 No. 9, pp. 151-153.

Tsuruoka, S., Hirano, C., Yoshikawa, T. and Shinogi, T. (2001), "Image-based structure analysis for a table of contents and conversion to XML documents", Workshop on Document Layout Interpretation and its Application (DLIA 2001), Seattle, September 9.

Wikipedia (2014), "Grounded theory", available at: https://en.wikipedia.org/wiki/Grounded_theory (accessed August 2, 2014).

Wu, Z., Mitra, P. and Giles, C.L. (2013), "Table of contents recognition and extraction for heterogeneous book documents", *12th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1205-1209.

Xia, Z.N. and Chen, Z.L. (2009), *Cihai*, Shanghai Lexicographical Publishing House, Shanghai.

Yu, X. and Liu, Y. (2012), "Analysis of tables of contents enrichment of western languages bibliographic records in the national library of China", *Journal of The National Library of China*, Vol. 21 No. 2, pp. 33-36 (in Chinese).

Zhang, K., Xu, P., Li, J.Z. and Wang, K.H. (2005), "Optimized hierarchy clustering based extraction for logical document structures", *Journal of Tsinghua University (Science and Technology)*, Vol. 45 No. 4, pp. 471-474.

Zhu, J. (2011), "Contents information description for library paper books", *Library and Information Service*, Vol. 55 No. 5, pp. 60-63, 59 (in Chinese).

Corresponding author
Dr Quan Lu can be contacted at: mrluquan@whu.edu.cn