2011 International Conference on Document Analysis and Recognition

Metadata Extraction System for Chinese Books

Liangcai Gao, Yuan Zhong, Yingmin Tang, Zhi Tang
Institute of Computer Science & Technology
Peking University, Beijing, China
{gaoliangcai, zhongyuan, tangyingmin, tangzhi}
@icst.pku.edu.cn

Xiaofan Lin
Vobile Inc
Santa Clara, CA, USA
xiaofan@vobileinc.com

Xuan Hu
College of Software
Beihang University
Beijing, China
huxuan@sse.buaa.edu.cn

Abstract—**Extracting metadata from academic papers has attracted much attention from researchers in past years. But how to extract metadata automatically from books is still seldom discussed. In this paper, we address this task on Chinese books and present a system to extract metadata from the title page of a book. This system consists of three components: metadata segmentation, metadata labeling, and post-processing. Different strategies are adopted in the system to identify different metadata types, and a variety of information sources, including geometric layout, linguistic, semantic content and header-footer, are used to accommodate the wide range of metadata layouts. Experimental results on real-world data have demonstrated the effectiveness of the proposed system.**

Keywords-metadata extraction; electronic book; page segmentation

I. INTRODUCTION

Nowadays more and more electronic books are available on Internet. In order to retrieve these books efficiently, it is necessary to get their metadata information such as title, author, publisher, and ISBN. Unfortunately, the metadata of a book are usually printed on its cover page and title page in the form of unstructured or semi-structured text, so that the metadata cannot be harvested directly. Up to now, metadata of electronic books is often manually extracted, which is very time-consuming and expensive. Our research is motivated by the need of a system that enables computers to automatically extract metadata from books.

Automatic metadata extraction, especially utilizing the title pages of academic papers, has attracted much attention from researchers. A number of methods on this task have been reported in recent years. Those approaches can be classified into three types: rule-based, learning-based and template-based approaches [2]. Rule-based methods are usually simple and effective, and have been widely used in real-world applications. For example, CiteSeer [1] uses heuristic rules to extract metadata from head parts and bibliographic attributes of research papers, and has become a well-known academic search engine. Wei et al. [14] also proposed a rule-based method for metadata extraction with layer-upon-layer tagging. Among the learning-based methods, Hidden Markov Model (HMM) [12, 13] was first used for the task, and achieved great success. However, it is difficult to model correlated features. Support Vector Machine (SVM) [6] is also applied to the task and has achieved better results in handling correlated features through loosening requirements on the relationship between state transformation and observation sequence. Conditional Random Fields (CRF) enjoys the advantages of both HMM and SVM by allowing the use of arbitrary correlated features and joint inference over entire sequences, and thus has achieved the best overall word accuracy of 95.37% on the public Cora dataset [11]. Template-based methods use template databases with various styles of title pages and reference sections. M. Y. Day et al. [3] designed metadata templates for computer science literature and constructed a hierarchical knowledge representation framework (INFOMAP) to extract reference metadata. Eli Cortez [4] proposed an unsupervised citation metadata extraction method to automatically generate templates from the training data set. ParaCite [7] is also a widely used system that parses metadata based on templates.

Compared to metadata extraction from academic papers, the related work on books is much more scarce. On the other hand, the metadata in a book has its own characteristics as listed below, and we propose a number of techniques dedicated to extracting metadata from books according to the those characteristics.

(1) It usually appears in the non-body pages such as cover pages and title pages. Our research focuses on the title page, since it contains most metadata of a book. Two typical examples of the title pages of Chinese books are shown in Figure 1 and that marked region contains the metadata of a book (called "MRegion" henceforth). The left example has an "indicator cell" before each metadata cell while no indicator cells appear in the right example. The text belonging to the same metadata type is called as a metadata entry. For the metadata entries with indicator cells, we can directly utilize the indicators to label the data cells and the rule-based methods can be sufficient.

(2) Some metadata types (e.g. title, author, publisher) usually appear in more places such as cover pages, title pages, and back pages. Furthermore, the book title and the author name(s) sometimes appear in the headers and footers of the body pages as well. Such redundancy can provide valuable clues for identifying them. We will utilize this characteristic to identify book titles and authors by page association.

(3) The styles of metadata entries vary significantly in different books. For example, a text line can contains one metadata entry or multiple entries depending on the books. Also, the spaces between entries, indicator cells and data

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.156

cells are quite different in different books, or even in the same book. In this paper, we propose an iterative dual-threshold page segmenting algorithm to handle various metadata styles.

In the rest of this paper, we will introduce an Automatic Extraction System of Book Metadata ("AESBM"). Specifically, we aim at extracting metadata of PDF-based books where the low-level information of fonts, characters and bounding boxes is already available. Section II describes the architecture of AESBM, the synergy among different components, and the implementation of each component. Section III quantitatively evaluates the performance and discusses the major causes of errors. Conclusions are drawn in Section IV.

II. Description of AESBM

A. *System Architecture*

In AESBM, the input is a multi-page PDF-based book, and the output is the metadata contained in the title page of the book. This system consists of three main components: metadata segmentation, metadata labeling, and post-processing. In metadata segmentation, title pages are first detected and then the MRegions are segmented into blocks. Metadata labeling employs two methods: a rule-based method and a learning-based method (Support Vector Machine). Post-processing includes two parts: metadata re-segmentation using a threshold based on the labeling results, metadata re-labeling based on the metadata continuity. In addition, there is dependency relationship among those components. Figure 2 displays the architecture of the system. In the following sub-sections we will present those individual components according to their logic order in the system.

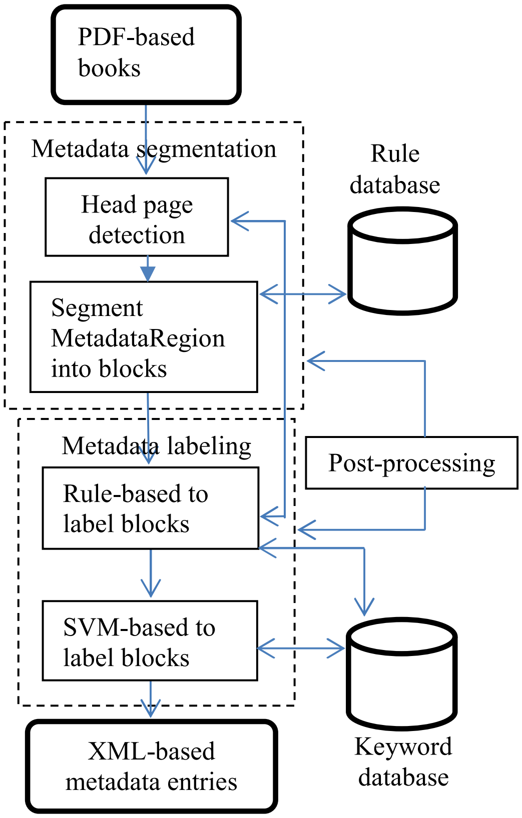


Figure 2. System architecture

B. *Metadata Segmentation*

1) *Title page detection*

To extract metadata from books, the first step is to find the title page of a book. We detect the title page in a multi-page book based on a number of heuristics. First, the title page is usually in the first several pages of a book. And in most Chinese books the second page is the title page. Second, title pages usually contain some special keywords such as "catalog in publication", "ISBN", "CIP", and the publisher name. Third, the MRegion is usually located at the bottom of the title page and there is a large white space between MRegion and the other regions.

2) *Segmenting the MRegion into blocks*

The goal of MRegion segmentation is to generate text blocks corresponding to individual metadata fields (author, publisher, price, ISBN, etc.) and their indicator cells. We segment the MRegion into blocks according to the following observations:

(1) A metadata entry usually occupies a single line, so we first divide the text of MRegion into lines through page projection in the horizontal direction.

(2) In a metadata entry, the indicator cell and the associated data cell are usually separated by special punctuations (e.g. ":") or spaces. Different entries are separated by spaces as well. Special punctuations can be directly obtained from the PDF text stream and spaces can be inferred by the distances between neighboring characters. A proper threshold is critical in detecting spaces. However, the threshold can be difficult to select because of the following factors:

- The space between the indicator cell and the data cell is usually larger than the space within each cell. However, a bigger space can also appear within an indicator cell when it contains very few characters. To deal with this corner case, we have compiled two keyword databases to record the frequently used indicator words and Chinese last names. The spaces in these words will be ignored.
- The spaces between metadata entries or cells usually vary with book styles. As a result, a threshold that works well in a book may be too large or too small for another book. To solve this problem, we have proposed a segmentation strategy using two thresholds: a smaller one and a larger one. If the distance between two neighboring characters in a line is smaller than the smaller threshold, we conclude that no space exists between the characters. If the distance is larger than the larger threshold, we conclude that a space exists between the characters and the text line is split up into blocks at the space. However, if the distance falls between the two thresholds, we first

assume that a space candidate exists between the characters, and then make the final decision in the following post-processing step.

3) Post-processing of Segmenting MRegion

In practice, the two thresholds mentioned above are set according to the width (*W*) of two neighboring characters (e.g. the smaller one is set as 0.25* W while the larger one is set to 0.4**W*). For a text line and the space candidates in it, if the number of candidates is ***n***, the text line may be divided in *2^n* ways. For each possible division, the blocks are generated, and then labeled using the methods introduced in Section II.C and II.D. Then a score is calculated for each division according to the labeling results. If a block is assigned a high confidence in the labeling step, the score increases by a certain amount. Finally, the division with the highest score is selected.

C. Rule-based Metadata Labeling

For most of Chinese books, their MRegions have regular layout and characteristic text clues. Thus, it is appropriate to use the rule-based method to label them. In our system, about 20 metadata types and their corresponding labeling rules are collected. The rules are created according to the following clues:

(1) Nearly all the indicator cells are the prefix of metadata entries. An indicator cell can directly tell us the type of the associated metadata data cell.

(2) Special verbs or nouns are usually added at the end of the metadata entries that have no indicator cells. For example, the word "millimeter" or "mm" appears at the end of the metadata type "page size".

(3) Some keywords, such as "publisher" and "ISBN", might be present to indicate the types of metadata entries.

(4) The text content of some metadata types, such as E-mail, ISBN, can be expressed in regular expressions.

There are two labeling rules:

***Is** Publisher:*

If** in the same line, the preceding block is indicator cell "Publisher" **or

*this block's text contains "publisher" **or***

there is a word like "published" or "issued" in the end of the block.

***Is** E-mail:*

If** in the same line, the previous block is an indication cell "E-mail", **or

the text content matches with regular expression "[\w-]+@([\w-]+\.)+[a-zA-Z]+"

A score is assigned to each rule based on its importance. And if a block is matched with a rule, the score of the rule is accumulated to the block's score.

After a text block is labeled by the rule-based method, there are three possible situations: (1) It has a higher score on only one metadata type. Then this metadata type is assigned to it. (2) It has no high confidence on any metadata type. (3) It has a high confidence score on more than a metadata type. For the last two cases, the metadata type of the text block is not obvious. So we use a leaning-based method to further label the text block.

TABLE I. FEATURES OF A TEXT BLOCK

NC	Number of characters
NL	Number of lines
LN	Line number
BN	Number of blocks in the same line
KW	Containing a keyword
FN	Containing a family name
RD	Ratio of digital number
RL	Ratio of Latin characters
RD	Ratio of special characters
FS	Font size
BL	At the beginning of a line
EL	In the end of a line

D. SVM-based Metadata Labeling

The rule-based metadata labeling method works well in most conditions. However, because of the wide range of Chinese book styles, they may fail to extract metadata in some cases. Besides, the expense and difficulty of creating and maintaining rules increase dramatically with more and more book styles added into the system. In order to overcome such limitations, we employ a learning-based method to determine metadata types of the text blocks. As with any pattern recognition problems, the key is to find discriminative and robust features. From extensive experiments, we select twelve features listed in Table I. Then an optimized implementation[1] of Support Vector Machines (SVM) is trained to label a text block. Furthermore, as we have observed from many books, several metadata types usually appear contiguously, so we feed not only the features of a block but also the features of its neighboring blocks into the SVM classifier.

If a metadata type is assigned to a text block by the SVM classifier, the score of the block on the metadata type increases by a certain amount on top of the rule-based method's score. Finally, the metadata type corresponding to the highest score is selected as the metadata type of the block.

E. Title and Author Identification

Title and author are two most important metadata types for a book. They usually appear in the cover page and/or the header/footer of body pages. We take advantage of this redundancy and go through the following steps to identify them:

(1) It is common that the text with the largest font belongs to the book title. Thus, when a book is processed, the first page is selected as the cover page, and the text with the largest font is selected as the title candidate. The text containing a family name is selected as the author candidate.

(2) Headers and footers are common formatting elements in book documents, and they surround a page's body text. Besides serving decoration purpose by making the page layout more balanced and more visually appealing, they also restate key information such as book titles, chapter titles, author names and page numbers. So we can utilize the information in header and footer to identify book title and author. After the headers and footers of a book are located by the method on physical structure extraction described in our

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

previous paper [5], the text in them is extracted to match with the title and author candidate detected in the above step. The matched text is determined as book title and author based on the rules mentioned earlier. For example, if the text contains family name, it is selected as the author field.

F. Post-processing of Metadata Labeling

The post-processing of metadata labeling is based on the following fact: the same kind of metadata type appears continuously in the MRegion. In the other words, the text of a metadata type is not separated by the text of other metadata types. Thus we first combine the neighboring blocks of the same type. Then if a block of Type 2 is "sandwiched" between two blocks of Type 1 and it has a high score of Type 1, the three blocks are combined together and assigned Type 1.

III. EVALUATION

The proposed system has already been used in a commercial E-book production system for over six months. In this section, we report the preliminary experimental results.

A. Dataset

The experiment data is obtained from Chinese digital libraries and contains 498 Chinese books, covering various title page styles and including 5392 metadata entries. We randomly select 150 books (or1768 metadata entries) as the training set and another 348 books (or 3624 metadata entries) for testing.

B. Performance Evaluation

In the title page detection and the MRegion segmentation, We provide both recall (the percentage of the true objects that the method finds), precision (the percentage of the objects that are in fact true). And we use the following equation to measure the metadata labeling accuracy:

$$\text{Accuracy} = \frac{\text{Number of correctly labeled entries}}{\text{Total number of entries}} \times 100\% \quad (1)$$

The results of title page detection and the MRegion segmentation are shown in Table II. 954 of the 5392 incorrectly-segmented metadata entries are recovered by the post-processing step. The labeling results of metadata entries are shown in Table III, where the rows correspond to the rule-based method, the SVM-based method, the combined method, and the impact of post-processing step. For the identification of titles and author names from cover pages and headers/footers, the results are shown in Table IV.

Our experiments are run on a standard PC (dual-core 1.4GHZ, 2GB RAM, Windows Server 2008), and the average processing time for a book is around 2 seconds.

TABLE II. HEAD PAGE DETECTION AND MREGION SEGMENTATION RESULTS

	Precision (%)	**Recall (%)**
Title page detection	96.1	99.4
MRegion segmentation	93.0	96.5

TABLE III. METADATA LABELING ACCURACY (%)

Metadata type	**Rule-based**	**SVM-based**	**Rule+ SVM**	**+Post-processing**
Publisher	93.5	73.9	94.2	94.7
Title	53.4	64.7	78.6	80.9
Author	97.1	99.3	99.4	99.8
Price	99.8	95.7	99.8	99.8
Word number	95.7	91.5	96.3	96.3
Address	84.8	45.7	88.7	91.6
Postcode	88.4	93.0	97.5	97.5
Telephone	93.8	90.6	94.4	95.1
E-mail	97.3	87.5	97.9	98.8
Issue	90.5	78.6	93.8	93.8
Page size	98.0	93.0	98.0	98.0
ISBN	96.5	95.5	96.8	97.2

TABLE IV. RESULTS OF IDENTIFICATION OF TITLES AND AUTHORS FROM COVER PAGE AND HEADERS/FOOTERS

Metadata type	**Precision (%)**	**Recall (%)**
Title	98.1	36.4
Author	93.9	34.2

C. Analysis of the Experimental Results

1) Metadata segmentation

In some books, the white spaces of title pages are represented with special characters in the PDF content stream, and thus the space cannot be detected through the distance between characters. Another common cause of segmentation errors is that some metadata entries on title pages have inter-character spaces even larger than the predefined high threshold of segmentation.

2) Metadata labeling

It can be seen from Table III that the rule-based method performs well on the metadata types with common distinguishing characteristics such as ISBN, E-mail, telephone, price, page size, and author. Its accuracy decreases in other metadata types that lack consistent characteristics. The performance of the SVM-based method complements that of the rule-based method. Thus, the combined method achieves better results in all types. In addition, because the metadata continuity is consistent in most books, the post-processing further boosts the accuracy, as shown in the last column of Table III. The majority of this task's errors are caused by metadata segmentation errors. Occasionally, the character codes of PDF files are illegible so that the text feature of blocks fails to identify their metadata types.

3) Title and author identification:

As can be seen from Table IV, the proposed method has a high precision rate of identifying title and author name from cover pages and headers/footers. However, it would fail when no enough text can be extracted from cover pages or headers/footers. Specially, many books do not contain the title or author information in their headers/footers and this method cannot process them. Consequently, its recall rate is low.

IV. Conclusions

This paper describes a complete system that can automatically extract metadata from Chinese books. The preliminary experiments demonstrate the effectiveness of the proposed methods. As far as we know, this is the first published attempt of extracting metadata from the title pages and building a complete solution around this idea. To handle various formats of title pages, we propose a number of effective techniques such as dual-threshold segmentation strategy, segmentation optimization based on labeling results, and post-processing of sandwiched labeling results. In the future, we will utilize the extracted metadata in several downstream document searching and management processing tasks.

References

[1] K. D. Bollacker, S. Lawrence, and C. L. Giles, "CiteSeer: an autonomous Web agent for automatic retrieval and identification of interesting publications," Proc. of the Agents '98, ACM Press, New York, NY, pp. 116-123.

[2] C. C. Chen, K.H. Yang, H. Y. Kao, and J. M. Ho, "BibPro: a citation parser based on sequence alignment techniques," Proc. of the IEEE International Conference on Advanced Information Networking and Applications (AINA 08), Mar, 2008, pp. 1175-1180.

[3] M. Y. Day, et al., "Reference metadata extraction using a hierarchical knowledge representation framework," Decision Support Systems, vol. 43, pp. 152-167, Feb. 2007.

[4] Eli, C., Altigran S. da Silva, Marcos A. G., Filipe M., and Edleno S. de Moura, "FLUX-CIM: flexible unsupervised extraction of citation metadata," Proc. of the Joint Conference on Digital Libraries (JCDL 07), ACM Press, Jun. 2007, pp. 215-224.

[5] L. Gao, and Z. Tang, "Comprehensive global typography extraction system for electronic book documents," Proc. of the Document Analysis Systems (DAS 08), Sep. 2008, pp. 615-621.

[6] H. Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox, "Automatic document metadata extraction using support vector machines," Proc. of the Joint Conference on Digital Libraries (JCDL 03), IEEE Computer Society, May. 2003, pp. 37-48.

[7] http://paracite.eprints.org/.

[8] A. Huang, J. M. Ho, H. Y. Kao, and S. H. Lin, "Extracting citation metadata from online publication lists using BLAST," Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 04), Springer, Berlin, May. 2004, pp. 539-548.

[9] A. K. Jain, M. N. Myrthy, and P. J. Flynn, "Data clustering: a survey," ACM Computing Survey, 1999, 31(3) pp. 264--323.

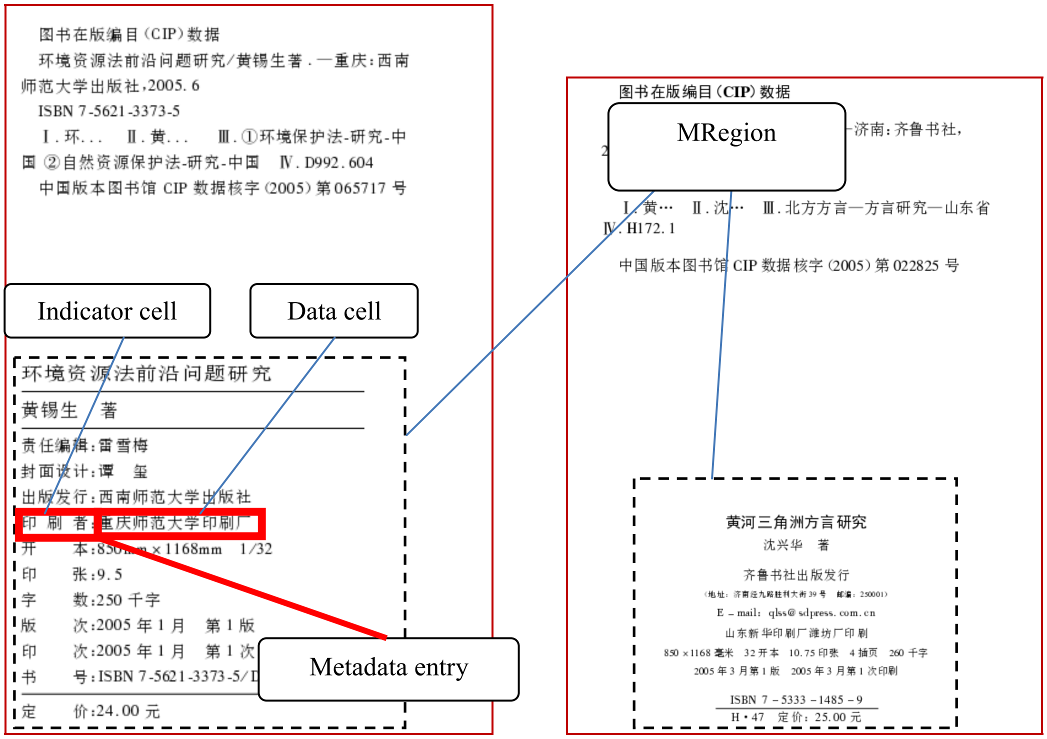
[10] C. Li, M. Zhang, Z. Deng, D. Yang, and S. Tang, "Automatic metadata extraction for scientific documents," Computer Engineering and Application, vol. 21, pp. 189-191.235, 2002.

[11] F. Peng, and A. McCallum, "Accurate information extraction from research papers using conditional random fields," Proc. of the HLT-NAACL '04, May. 2004, pp. 329-336.

[12] K. Seymore, A. McCallum, and R. Rosenfeld, "Learning hidden Markov model structure for information extraction," Proc. of the Conference on Artificial Intelligence (AAAI 99), Jul. 1999, pp. 37-42.

[13] A. Takasu, "Bibliographic attribute extraction from erroneous references based on a statistical model," Proc. of the Joint Conference on Digital Libraries (JCDL 03), IEEE Computer Society, Washington, DC, May.2003, pp. 49-60.

[14] W. Wei, I. King, and J.H.-M. Lee, "Bibliographic attributes extraction with Layer-upon-Layer tagging," Proc. of International Conference on Document Analysis and Recognition (ICDAR 07), Curitiba, Sep. 2007, pp. 804-808.



图书在版编目（CIP）数据

环境资源法前沿问题研究/黄锡生著．—重庆：西南师范大学出版社，2005.6

ISBN 7-5621-3373-5

Ⅰ.环...　Ⅱ.黄...　Ⅲ.①环境保护法-研究-中国 ②自然资源保护法-研究-中国　Ⅳ.D992.604

中国版本图书馆 CIP 数据核字（2005）第065717号

环境资源法前沿问题研究

黄锡生　著

责任编辑：雷雪梅
封面设计：谭　玺
出版发行：西南师范大学出版社
印 刷 者：重庆师范大学印刷厂
开　　本：850mm×1168mm　1/32
印　　张：9.5
字　　数：250千字
版　　次：2005年1月　第1版
印　　次：2005年1月　第1次
书　　号：ISBN 7-5621-3373-5/D

定　　价：24.00元

图书在版编目（CIP）数据

—济南：齐鲁书社，

2

Ⅰ.黄…　Ⅱ.沈…　Ⅲ.北方方言—方言研究—山东省
Ⅳ.H172.1

中国版本图书馆 CIP 数据核字（2005）第022825号

黄河三角洲方言研究

沈兴华　著

齐鲁书社出版发行

（地址：济南经九路胜利大街39号　邮编：250001）

E－mail：qlss@sdpress.com.cn

山东新华印刷厂潍坊厂印刷

850×1168毫米　32开本　10.75印张　4插页　260千字

2005年3月第1版　2005年3月第1次印刷

ISBN 7－5333－1485－9

H·47　定价：25.00元

Figure 1. Two typical title pages of Chinese books