ICIC Express Letters
Part B: Applications
Volume **12**, Number **2**, February **2021**

ICIC International ©2021 ISSN 2185-2766
pp. 121–132

A RULE-BASED INFORMATION EXTRACTION APPROACH FOR EXTRACTING METADATA FROM PDF BOOKS

Abrar Alamoudi, Amal Alomari, Sarah Alwarthan and Atta-ur-Rahman*

Department of Computer Science
College of Computer Science and Information Technology (CCSIT)
Imam Abdulrahman Bin Faisal University (IAU)
P.O. Box 1982, Dammam 31441, Saudi Arabia
{ 2200500066; aaalomari; saalwarthan }@iau.edu.sa; *Corresponding author: aaurrahman@iau.edu.sa

Received June 2020; accepted September 2020

Abstract. *Nowadays PDF books have become widely used because of their easily availability and accessibility through the Internet from anywhere and anytime. However, the process of extracting information structure from PDF sources automatically is considered as a challenging task. Nonetheless, it is necessary for effective knowledge representation, archiving and retrieval through creating the digital libraries in semantic web. In this work, an intelligent rule-based approach is proposed for extracting the logical metadata from PDF books accurately. Furthermore, a set of rules and underlying patterns are defined to achieve high performance in terms of accuracy. The experimental results indicate that the proposed approach is capable of extracting the metadata from PDF books successfully with an overall accuracy of 94.62% and 90.27% for both training and testing sets, respectively. Moreover, this study could benefit the publishing houses to enhance the indexing, citations, and usability of their electronic books. Also, it would help the researchers who are interested in improving the performance of extracting information from PDF documents.*

Keywords: Information extraction, Rule-based, Regular expression, Metadata, PDF books

1. **Introduction.** As a result of the evolution in technology nowadays, a huge number of resources become available in a digital form on the Internet. And Information Extraction (IE) has become an increasingly important field that has been growing very fast in the last few years, respectively with the massive amounts of generated data on the Internet. Information extraction refers to the process of extracting the information from unstructured or semi-structured text sources by detecting the rules of the document structure [1,2]. There are many useful applications for applying IE techniques such as books, scientific research, web intelligence, and search engines. To automate the extraction of structured information from the text documents, a set of patterns and rules need to be defined precisely. These patterns are determined based on the syntactic and semantic constraints in the natural language [1]. Most of the published scientific papers and books on the Internet are commonly available in Portable Document Format (PDF). Essentially, books are published in two forms, which are printed books and electronic books. The proposed work focuses on extracting information from PDF books that are known as one of the electronic books' formats. Besides, International Standard Book Number (ISBN) is used as a book identifier by the publisher. The main problem is that extracting structured data from unstructured data form is a challenging task that requires engaging the researchers to achieve more enhancement and improvement. Although there are some proposed studies, the problem of extracting metadata remains hard and complicated, primarily due to the huge diversity of the used layout and formatting style in PDF books. Moreover, it is

DOI: 10.24507/icicelb.12.02.121

difficult and not a direct task to access the targeted metadata of the PDF book as there is no clear keyword or tag that indicates metadata information. Therefore, there is an urgent need to develop an efficient approach to extracting information from PDF format correctly [3-7].

The objective of this work is to develop a method for automatic metadata extraction from PDF books to support the dramatic growing volume of books in PDF format. The underlying motive is that the process of extracting and entering metadata manually is time-consuming and the chance for errors is high, while automatic extracting and entering can speed up the process and grant accuracy. Moreover, it can perform the tasks of searching, retrieving, and querying metadata faster and more efficiently than the traditional manual way. In addition, many publishers have increasingly provided their published books through the Internet as PDFs since the high customer demands for this convent form of books. Recently, most publishers are keen to add the International Standard Book Number for electronic books (ISBN ebk) which indicates the popularity of using the PDF book formats. This massive amount of PDF books requires a sufficient, reliable, and accurate technique to enhance its usability. This research would help in meeting the demand of different sectors such as e-library of universities that offer PDF books for their researches, employees, and students. Also, it would assist in improving the performance of archiving systems and online stores that are mainly concerned with distributing and selling PDF books. To extract significant information from the PDF documents, we propose an automatic information extraction system that uses a PDFBox tool to convert the PDF book to a raw text. This text will be used as input for the proposed system with the aim of getting the book's metadata information such as Title, Authors, ISBN, Publisher, and Year of Publication. The existing information extraction solutions can be categorized as machine learning-based, and rule-based. Since rule-based approaches are robust and require less training [6], this paper presents a rule-based Information Extraction (IE) system from PDF books that were published by Wiley Publisher. The performance result of the proposed system to extract the book metadata got an overall accuracy of 94.62% for training and 90.27% for testing. The main contribution of this work is to implement an efficient information extraction system that can retrieve the information from digital PDF book sources accurately. In addition to adding a valuable contribution to those organizations that deal with PDF books documents such as electronic libraries, publishers, and e-booksellers.

The rest of the paper is organized as the following. Section 2 presented the related work. Section 3 introduced the proposed methodology. The experimental result is discussed in Section 4, while, Section 5 highlights the future work and concludes the paper.

2. **Literature Review.**

Many research works have been conducted about extracting metadata from digital documents. Different techniques and approaches were implemented and studied. In this section, we review the previous related literature for extracting information from source documents that are stored in a PDF format.

Paper in [3] has described the application of machine learning and data science pipeline as effective elements for structured information extraction from different papers and documents in metadata analysis. The method proposed to be able to extract data efficiently involves a procedural methodology approach of pipeline stages that are mentioned as follows: document acquisition and essential filtering; the payload extraction; the recipe step extraction that functions as an affiliation task in the process; the assembly of recipe and finally the information retrieval that is based on QA functionality. The pipeline uses a holistic document analysis where each step is essentially designed to pass on successive defined metadata features to the next stage. This includes the utilization of machine learning based extraction techniques applied to labeled as well as marked up corpora. The results of the pipeline approach are that it shares metadata standards from stage to

stage while providing a unified supporting framework for representations and algorithms of all the stages that are driven by machine learning. The advantage is that this system has incorporated information extraction into sequential and sub-sequential units which allows for specification and extraction of desired material only. In [4], authors proposed an information extraction system for the resumes in different types of files and diverse formats. The authors applied two steps to extracting the main resume information. In the first step, they used the raw text to determine the resume blocks. In the second step, the authors used semi-structured files to identify the main facts in the resume blocks. They started by extracting the raw text from the resume file by using Tika and they applied preprocessing for the files to removing any visual format information such as remove images, watermarks. Moreover, the authors identify the composition of each line in the file by using a multiple class to identify the label for each phrase such as date, university name, department name by applying the Naive Bayes classifier. Also, they created Writing Style features for each line that contain word index, punctuation index, word lexical attribute, and the results of classifiers, and all of this information will be in a semi-structure file that will be the input for the second step. In the second step, the authors used Writing Style to identify the main items in the block. This paper improved the accuracy of information extraction for resumes.

Authors in [5], focused on extracting information which is a relation extraction between two named entities from unstructured or semi-structured Wikipedia articles. The proposed system utilized DeepDive which is a powerful system to extract the value from a big amount of available data. DeepDive has been used to recognize the Name Entity Recognition (NER) and feature extraction. In this research the experiment was applied on 200 Wikipedia articles that have 113 articles containing spouse relationship and 87 articles without spouse relationship. Authors found that the proposed work extracts the information efficiently from unstructured Wikipedia articles by using DeepDive. DeepDive got a good accuracy which is around 79%. Authors in [6], proposed an approach to extracting the entities and their relation automatically from a PDF novel by using information extraction techniques. The proposed work consists of two steps which are, Name Entity Recognition (NER) that aims to extract and identify the entities from the novel and relation extraction that aims to find and classify the relation between those entities. These two processes utilized supervised machine learning and rule-based techniques. The proposed system identifies 5 types of entities but only person, organization and location entity will be part of the relation, as well as the relations. Six relations have been identified by the proposed system. In this research, the training data set is collected from 6 fiction novels. SVM, decision tree, and Naïve Bayes have been used to find combinations of features that optimize the process. SVM got the highest average evaluation score compared with others. Moreover, authors found that the proposed NER model is more appropriate to detect fictional words compared with the Stanford NER model that is appropriate to detect non-fiction sentences. In [7] the authors presented a crucial benchmark that can be used in the metadata extraction process with a focus on the PDF documents. The problem stated is that the PDF is essentially a layout-based format that puts more emphasis on the fonts and the positioning of the individual characters other than focusing on the semantic aspects of the text, which is more important. The proposed methodology involves TeX or PDF file parsing, identification of the logical text blocks, and serializing the logical text blocks to files. Through a benchmark analysis of the fourteen PDF extraction tools, they are still not performed inefficiently even including the authors' own tool they call "icecite" that outperforms the other popular tools investigated. They concluded that their tool was not perfect due to the rule-based approach. However, they propose that through their methodology for metadata extraction and by basing on the learning-based approach it would essentially fix the open problems. They claim that with the application of an effective logical block text isolation, then the

extraction will become more efficient with the required data easily isolated. The authors in [8], proposed a model to recognize the structure for the publication reports in PDF format for the Information Extraction (IE) system. They used PDFBox tools to extract the raw text from PDF format, and used Stanford's Named Entity Recognition to determine publication metadata such as author, and affiliation. After that, they applied the rule-based multi-pass sieve algorithm to classifying the PDF text into five categories, which are title, abstract, body text, etc. They found the proposed classification algorithm improved the performance of the IE system, and it gets 92.6% accuracy, which is higher than the machine learning classifier algorithm (logistic regression). Furthermore, the authors in [9] addressed the inefficiencies and difficulties encountered in the extraction and collection of metadata from existing literature and research papers. This is due to the challenging and diverse layouts which mostly yield ineffective data while using various data extraction software. The paper addresses this problem by recommending CERMINE, a complex and diverse tool that is efficient in the auto-extraction of metadata. The method involves page segmentation, metadata zone classification and reference parsing. The effective modular architecture and arrangement of CERMINE have made it more flexible as well as easy to adapt to diverse documents layouts in metadata extraction. Through its unique segmentation process, the software ability to obtain pages that have or contain characters that have been grouped into various zones, words and lines provides a more readable sequence that can be extracted in its entirety. A greater limitation of the methodology is that it does not have a process path that allows for the extraction of structured full text that contains the sections, subsections paragraphs and headers as part of the extraction. Authors in [10], proposed a Table of Content (TOC) recognition approach for large scale book documents with different TOC types and styles to recognize the three elements: title, page number and level. The proposed method for recognition and extraction of the TOC has three sub-tasks. The first sub-task is TOC detection, and it identifies the place or location of the TOC. Authors found that the TOC usually appears within the first 20 pages. The second sub-task is TOC parsing task that aims to parse the TOC by using different rules based on the TOC style and the third sub-task is TOC linking. The proposed method selects most suitable TOC parsing rules according to the table of contents' style. The proposed algorithm has been evaluated on two datasets that contain PDF and OCRed books. Finally, the authors found that the proposed algorithm is more outperformed compared with the two existing baselines on both datasets. The authors in [11] have presented and proposed an efficient automatic metadata extraction method that is specifically meant for the retrieval of the bibliographical information from the various digital research papers and academic documents in PDF formats. The used method is based on three important steps: first, extracting the font size and the text information by using PDFBox software. Then, the rule-based method is used in the identification of the titles of papers. Finally, the Hidden Markov Model (HMM) is applied for extracting the authors and the titles. The research obtained is then sent to the digital libraries to obtain the rest of the important metadata required. While using and basing the research on the HMM one-state model, which is the proposed model for the research, the results indicated a comparable performance with regard to the extraction of data but with a little more improved performance when compared to the HMM multi-state model. In addition, its advantage is that the system proves efficient in the extraction of bibliography data when tested with different formats and samples of digital documents.

In [12] the authors proposed a framework to extract the metadata from the scientific papers, which are title, authors, and abstract. The extraction system starts by using crawler to search for a scientific paper on the Internet in PDF format, and it records the URL of the paper and downloads this paper. After that, the system converted the PDF format to text file and XML file, and then identified if this paper is a scientific paper or not. If it is a scientific paper, then the metadata extraction will be extracted and stored in the

database. The extraction process in this paper is based on visual and spatial knowledge that people used while reading the document. At the end of the experiment, the authors get higher accuracy during the extraction of the titles and abstracts more than authors, according to difficulty for identifying the authors' names from affiliations. In [14], authors comprehensively presented the approaches for information extraction in the literature. Among these approaches, rule-based approaches gained a lot of attention due to their promising nature. For example, in [15] authors proposed a rule-based extraction method for metadata extraction from published PDF articles. The target articles were mainly focusing from a conference. Authors claimed an accuracy for above 70%. In [16-22], similar approaches were used to extract information from text for sake of automated text categorization/classification by means of extracting information from the unstructured text. Other than rule-based approaches, data mining-based approaches have been used for information extraction from structured and semi-structured data [23-25,31]. It is apparent that the ultimate purpose of information extraction from unstructured, semi-structured data is to provide a better way for knowledge representation [26]. So it can be efficiently retrieved by means of well-known retrieval algorithms and query languages [27,28] from the stored data in digital libraries [29,30]. It eventually contributes to equipping search engines by means of better indexing and searching through the semantic web. As a summary of the literature review and motivation behind the work, following points can be listed.

1) Rule based methods are more promising in terms of metadata extraction.

2) Rule based methods have not been investigated for the PDF books.

3) Information extraction from PDF books has its own significance due to emergence of digital libraries, indexing and efficient searching over the web.

4) A state-of-the-art rule-based metadata extraction method for PDF books is a need.

3. Research Methodology.

For the experimental research, a dataset of collected books from the specified domain was used and the PDFBox tool was employed to carry out the experiment following the computational standards. Besides, a set of rules was defined to extract the metadata (Title, Authors, ISBN, Publishers, and Year of Publication) from the PDF books automatically. As far as the computational complexity is concerned, it is linear with respect to the number of fields being extracted $O(n)$, where n is number of fields.

3.1. Description of dataset.

In this work, a dataset of collected books was used to carry out the experimental result. The dataset consists of 25 books about different computer science topics that were published by Wiley Publisher. The entire dataset was obtained for the examination where 80% of the dataset was used for training and 20% used for testing.

3.2. Converting PDF to text using PDFBox tool.

The Apache PDFBox is an open-source library based on the Java language. It provides the utilities to create, convert, and manipulate PDF documents. Also, it can extract text from PDF documents efficiently. In this study, the PDFBox tool was applied to extracting the raw text from PDF book content.

3.3. Books' metadata extraction model.

This section discusses the proposed model to extract the metadata from PDF books. The proposed model provides a solution to extract books' metadata automatically by using Information Extraction technology (IE). Detecting a book's metadata requires two stages. The first stage is working on converting the book from PDF to text format by using the PDFBox tool as shown in Figure 3. Then, the proposed model will extract the metadata for the book based on the proposed rules as illustrated in Figure 1.

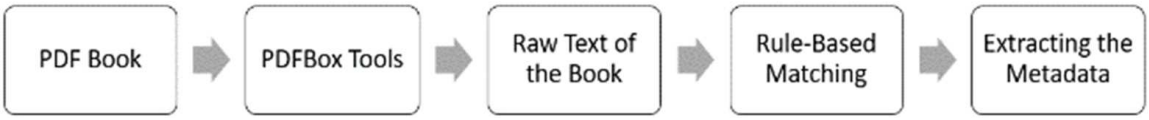


FIGURE 1. Books' metadata extraction stages

Extracting books metadata (Title, Authors, ISBN, Publisher Name, Year of Publication) from a PDF book needs to consider several constraints; for instance, the book's structure may differ from one to another since it depends on publisher writing styles. This research focuses on Wiley publisher which has special characteristics compared to other publishers as follows:

- All books are starting with a book cover which is an image and it has been ignored.
- Some of the book information (such as Title, Year of Publication, Publisher Name, ISBN) appear in the copyright page as shown in Figure 2.

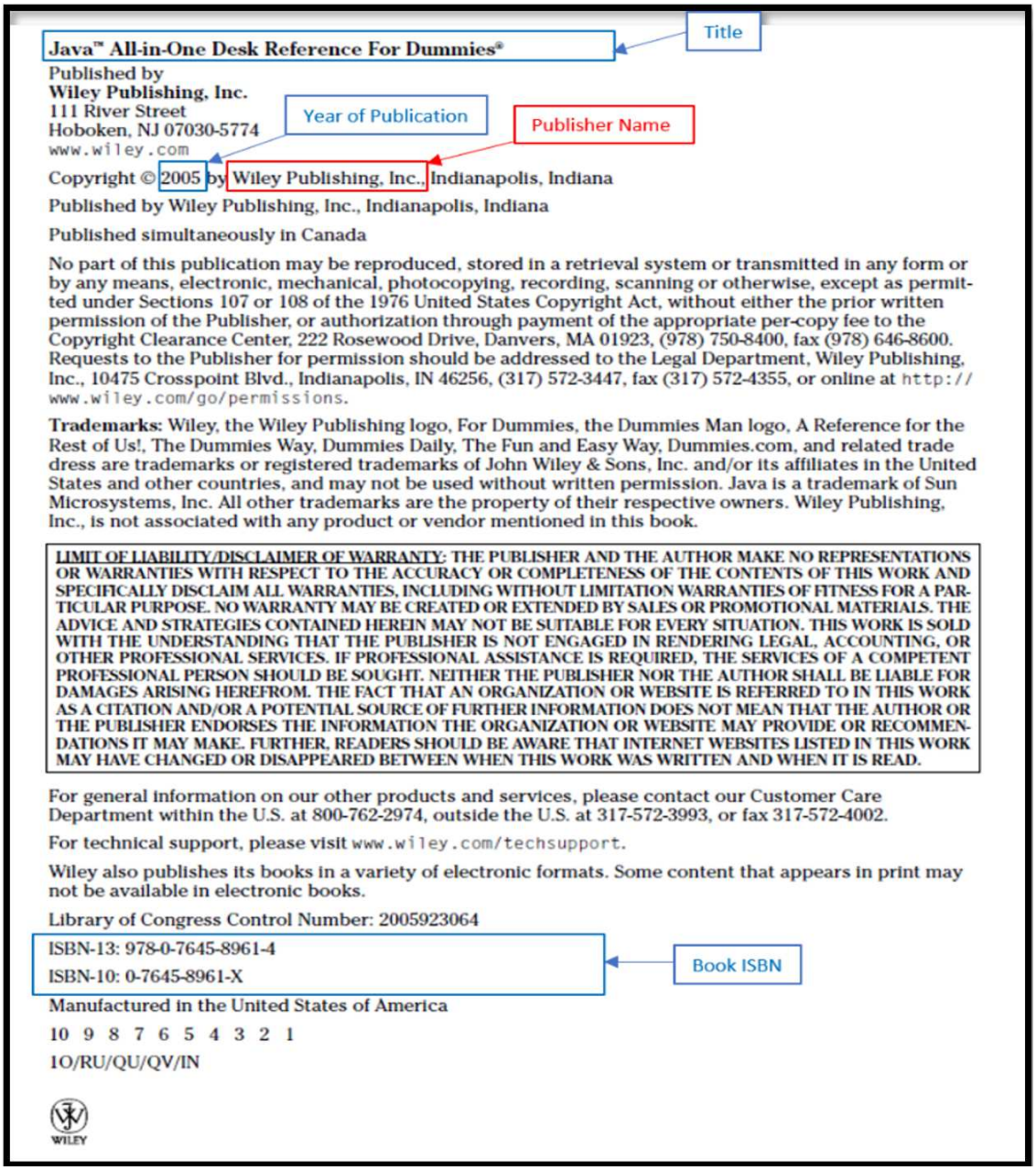
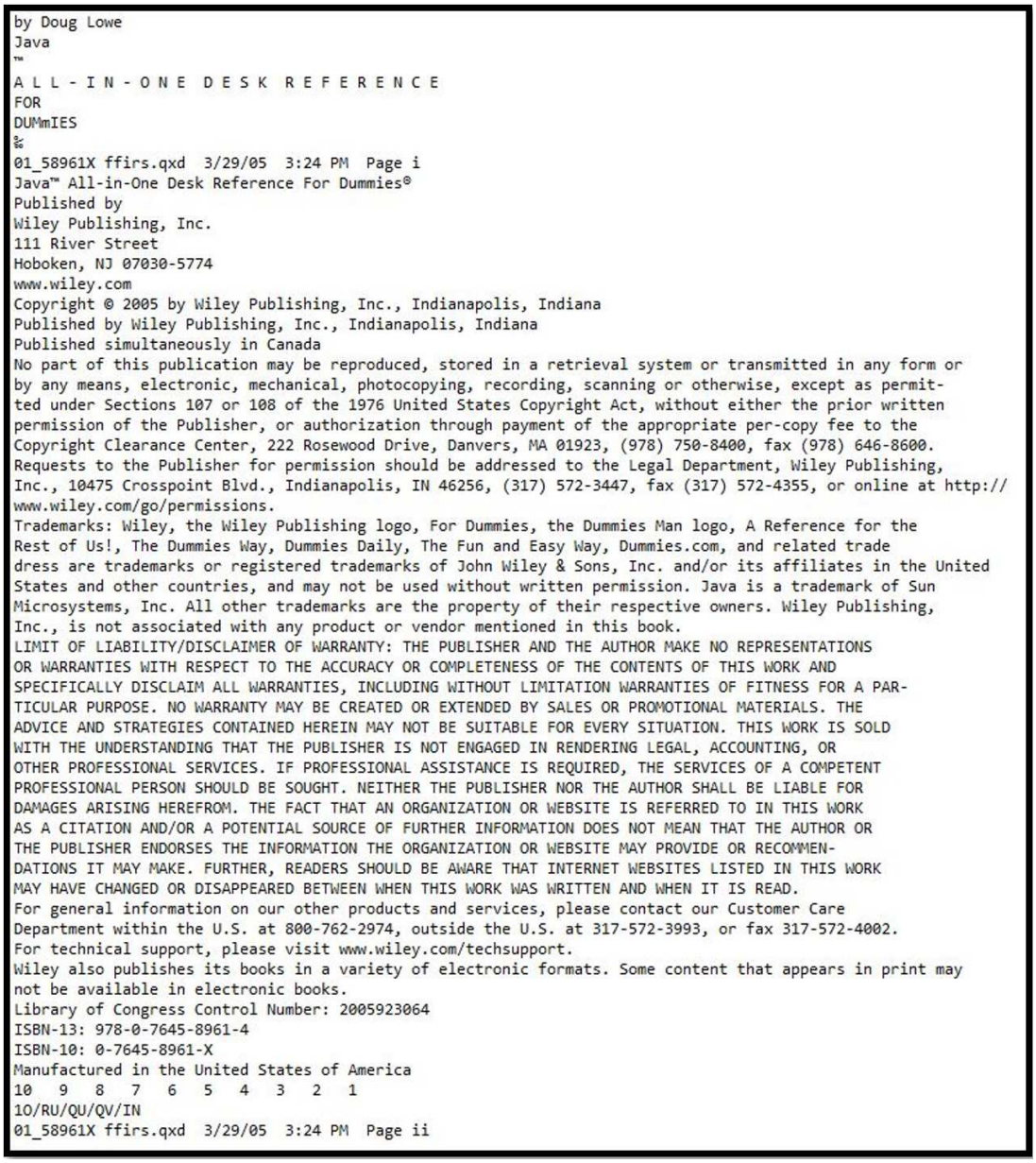


FIGURE 2. Copyright page from PDF book



```
by Doug Lowe
Java
™
A L L - I N - O N E  D E S K  R E F E R E N C E
FOR
DUMmIES
%
01_58961X ffirs.qxd  3/29/05  3:24 PM  Page i
Java™ All-in-One Desk Reference For Dummies®
Published by
Wiley Publishing, Inc.
111 River Street
Hoboken, NJ 07030-5774
www.wiley.com
Copyright © 2005 by Wiley Publishing, Inc., Indianapolis, Indiana
Published by Wiley Publishing, Inc., Indianapolis, Indiana
Published simultaneously in Canada
No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or
by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permit-
ted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written
permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the
Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600.
Requests to the Publisher for permission should be addressed to the Legal Department, Wiley Publishing,
Inc., 10475 Crosspoint Blvd., Indianapolis, IN 46256, (317) 572-3447, fax (317) 572-4355, or online at http://
www.wiley.com/go/permissions.
Trademarks: Wiley, the Wiley Publishing logo, For Dummies, the Dummies Man logo, A Reference for the
Rest of Us!, The Dummies Way, Dummies Daily, The Fun and Easy Way, Dummies.com, and related trade
dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United
States and other countries, and may not be used without written permission. Java is a trademark of Sun
Microsystems, Inc. All other trademarks are the property of their respective owners. Wiley Publishing,
Inc., is not associated with any product or vendor mentioned in this book.
LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS
OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND
SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PAR-
TICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE
ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD
WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR
OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT
PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR
DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK
AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR
THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMEN-
DATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK
MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.
For general information on our other products and services, please contact our Customer Care
Department within the U.S. at 800-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002.
For technical support, please visit www.wiley.com/techsupport.
Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may
not be available in electronic books.
Library of Congress Control Number: 2005923064
ISBN-13: 978-0-7645-8961-4
ISBN-10: 0-7645-8961-X
Manufactured in the United States of America
10   9   8   7   6   5   4   3   2   1
1O/RU/QU/QV/IN
01_58961X ffirs.qxd  3/29/05  3:24 PM  Page ii
```

FIGURE 3. Extracted raw text from copyright page using PDFBox tool

- The copyright page appears in different places in each PDF book.
- There is no specified page for the authors' names. Authors' names appear once or twice in the first 4 pages under the book title.
- Some authors' names are written across multi lines while others are listed in one line.
- Some authors' names have the prefix such as "by".
- An ISBN appears on the copyright page, but each book has different kinds of ISBNs. Some of the books have only one ISBN, and some books have two like ISBN 10 and 13.

3.3.1. *Title extraction.* To get the title from PDF book, a rule was built to match the book's title (Table 1). It has been noted that the book title always appears on the copyright page, where the title is the first line on the page followed by "Published by". The rule states that the text at the beginning of the page and before the "Published by" will be retrieved as the title of the book. To recognize the beginning of the page, it has

TABLE 1. Sample rules

#	Field	Rule
1	Title	"(Page\\s.|| Page\\s..||Page\\s...)(\\n) ([A-Za-z0-9\\s\\-\\.℗™'&%$#@!*()®,.⊡+]+)(\\n)(Published by)"
2	Publishing year and publisher name	"(Copyright ©\\s)([0-9][0-9][0-9][0-9])(\\sby\\s)(([A-Za-z ,&]+)(.))"
3	ISBN	"(ISBN:\\s)([0-9]+[-][0-9]+[-][0-9]+[-][0-9X]+)"
4	ISBN 10	"(ISBN-10:\\s)([0-9]+[-][0-9]+[-][0-9]+[-][0-9X]+)"
5	ISBN 13	"(ISBN-13:\\s)([0-9]+[-][0-9]+[-][0-9]+[-][0-9]+[-][0-9X]+)"
6	Electronic ISBN	"(ISBN\\s||ISBN:\\s)([0-9]+[-][0-9]+[-][0-9]+[-][0-9]+[-][0-9X]+)(\\(ebk\\))"
7	Print ISBN	"(ISBN\\s||ISBN:\\s)([0-9]+[-][0-9]+[-][0-9]+[-][0-9]+[-][0-9X]+)(\\(pbk\\))"
8	Authors' names	"(by)(?!Wiley)(?![a-z])([A-Z][a-zA-Z ,.]+)"

been noted that the beginning of the current page follows the page number of the previous page.

3.3.2. *Published year and publisher name extraction.* The publisher's information often appears on the copyright page. To identify this page, it needs to find out the specific word that distinguishes the publisher's information section which is "Copyright©". A simple way to extract the published year and publisher name is by finding the keyword "Copyright©" in the first 20 pages of the book. This keyword followed by 4 numbers that indicate year of publication, and the publisher name will appear after "by" as shown in Table 1.

3.3.3. *ISBN extraction.* An ISBN is used as a book identifier by the publisher, universities, libraries, and booksellers for the purpose of listing, ordering, and other uses. Various types of ISBN have been considered in this approach, which they commonly consist of many numbers separated by dash. Each part of the ISBN has its embedded meaning [13]. It has been noted that the book's ISBN number appears on the copyright page. The book usually contains one or more ISBNs that match the pattern. Mainly they are five types of ISBN like for e-Book (contains keyword ebk), and print book (contains keyword pbk) given in Table 1.

3.3.4. *Authors extraction.* Due to the lack of a unified style for the author's name, some difficulties were encountered during the author's name extraction process. It has been noted that the authors' names appear under the title in the first three pages. Unfortunately, the title that comes before the authors' names is in the form of a picture not text, so it was difficult to build a rule that reached the authors' names through the title. To solve this problem, the rule was built to match the author's name (Table 1). To extract the authors' names, the keyword "by" is used to identify the authors' names, because in most books' cases the keyword "by" appears before the authors' names. Also, rule was built to exclude the word Wiley and any words start with small letters with the aim of matching the names that start with capital letters.

4. **Results and Discussion.** The experimental result demonstrated that the implemented model in this study was able to automatically extract the metadata from the PDF books successfully with an overall accuracy of 94.62% for training and 90.27% for

testing as shown in Table 2. The original metadata was extracted manually and stored in an Excel file, while the output result of the implemented model is illustrated in Figure 4. As it is observed from Table 2, the majority of the extracted metadata achieved high performance in terms of accuracy in both phases of training and testing. The accuracy was calculated as ratio of correctly extracted characters to total characters. However, some metadata obtained lower accuracy regarding different reasons related to the input books' format. Conceding the ISBN, it is clear that the test accuracy has decreased to 97.10%, due to the new ISBN format that is followed in the new edition by Wiley Publisher in one of the testing samples. In this case, the tested book's ISBN consists of five parts while the applied format in this work consists of four parts. At the same time, ISBN 13 obtained 80% during the testing phase as a result of miss typing in the ISBN 13 format of the input book, since an expected single (-) typed as double (- -). On the other hand, the performance of extracting the author was low by comparing with other metadata. The achieved accuracy during training and testing was 51.57% and 35.37% respectively. The reason behind that is the defined rules in this study works on extracting the authors' names that are followed the word "by". Therefore, the authors' names that were written differently were not extracted correctly. Moreover, the applied pattern in the proposed work extracts the authors' names that are clearly written within a single line, not across multi-lines.

Table 2. Accuracy of extracting the metadata

Section	Accuracy	
	Training	Testing
Title	100%	100%
Publisher	100%	100%
Published year	100%	100%
ISBN	100%	97.10%
ISBN10	100%	100%
ISBN13	100%	80%
ISBN (pbk)	100%	100%
ISBN (ebk)	100%	100%
Authors	51.57%	35.37%
Overall	94.62%	90.27%

	A	B	C	D	E	F	G	H	I	J
1		Title	Publisher	Published year	ISBN	ISBN10	ISBN13	ISBN(PBK)	ISBN(EBK)	Authors
2	Training	ASP.NET 2.0 All-in-One Desk Reference For Dummies®	Wiley Publishing, Inc.	2006		0-471-78598-9	978-0-471-78598-9			Doug Lowe, Jeff Cogswell,
3		ASP.NET 2.0 Everyday Apps For Dummies®	Wiley Publishing, Inc.	2006		0-7645-9776-0	978-0-7645-9776-3			Doug Lowe
4		ASP.NET 2.0 Instant Results	Wiley Publishing, Inc.	2006		0-471-74951-6	978-0-471-74951-6			
5		Wrox's ASP.NET 2.0 Visual Web Developer™ 2005 Express Edition Starter Kit	Wiley Publishing, Inc.	2006		0-7645-8807-9	978-0-7645-8807-5			Step
6		AutoCAD® 2005 For Dummies®	Wiley Publishing, Inc.	2004	0-7645-7138-9					Mark Middlebrook
7		AutoCAD® 2007 For Dummies®	Wiley Publishing, Inc.	2006		0-471-78649-7	978-0-471-78649-8			David Byrnes and Mark Middlebrook
8		PHP & MySQL® For Dummies®, 2nd Edition	Wiley Publishing, Inc.	2004	0-7645-5589-8					Janet Valade
9		Visual Basic® 2005 For Dummies®	Wiley Publishing, Inc.	2006		0-7645-7728-X	978-0-7645-7728-4			Bill Sempf
10		Beginning Programming with Java® For Dummies®, 3rd Edition	John Wiley & Sons, Inc.	2012				978-0-470-37174-9		Barry Burd
11		Beginning Programming with Java® For Dummies®, 4th Edition	John Wiley & Sons, Inc.	2014				978-1-118-40781-3	978-1-118-41756-0	Barry Burd
12		Beginning Programming with Java™ For Dummies®, 2nd Edition	Wiley Publishing, Inc.	2005		0-7645-8874-5	978-0-7645-8874-7			Barry Burd
13		Beginning Visual Basic®2005	Wiley Publishing, Inc.	2006		0-7645-7401-9	978-0-7645-7401-6			
14		C++ For Dummies®, 5th Edition	Wiley Publishing, Inc.	2004	0-7645-6852-3					Stephen Randy Davis
15		C++ Timesaving Techniques™ For Dummies®	Wiley Publishing, Inc.	2005	0-7645-7986-X					Matthew Telles
16		C# 2005 For Dummies®	Wiley Publishing, Inc.	2006		0-7645-9704-3	978-0-7645-9704-6			Stephen Randy Davis
17		HTML 4 For Dummies, 5th Edition	Wiley Publishing, Inc.	2005		0-7645-8917-2	978-0-7645-8917-1			Ed Tittel and Mary C. Burmeister
18		Java™ 2 Enterprise Edition 1.4 Bible	Wiley Publishing, Inc.	2003	0-7645-3966-3					Valesky, Java How to Program by
19		Java™ All-in-One Desk Reference For Dummies®	Wiley Publishing, Inc.	2005		0-7645-8961-X	978-0-7645-8961-4			Doug Lowe
20		Java® eLearning Kit For Dummies®	John Wiley & Sons, Inc.	2014				978-1-118-09878-3	978-1-118-22370-3	John Paul Mueller
21		Linux® All-in-One For Dummies®, 5th Edition	John Wiley & Sons, Inc.	2014				978-1-118-84435-9	978-1-118-84431-1	Emmett Dulaney
22										
23	Testing	JavaScriptTM In 10 Simple Steps or Less	Wiley Publishing, Inc.	2004	0-7645-4241-9					
24		AutoCAD® & AutoCAD LT® All-in-One Desk Reference For Dummie	Wiley Publishing, Inc.	2006		0-471-75260-6				Lee Ambrosius and David Byrnes
25		PHP & MySQL® For Dummies,® 4th Edition	Wiley Publishing, Inc.	2010	978-0-470-52758					Janet Valade
26		Troubleshooting Your PC For Dummies®, 2nd Edition	Wiley Publishing, Inc.	2005	0-7645-7742-5					Dan Gookin
27		Beginning PHP5, Apache, and MySQL® Web Development	Wiley Publishing, Inc.	2005	0-7645-7966-5					

Figure 4. Output result of training and testing phases

Moreover, a comparison between the proposed approach and earlier published approaches is presented in Table 3 from the perspective of document type, proposed technique, and the information to be extracted. As it is clear from Table 3 below, there is a lake of the published works that focused on extracting the metadata from PDF books. Furthermore, the majority of the proposed techniques for extracting the metadata from PDF sources were rule-based, machine learning, or a combination of both.

Table 3. Qualitative comparison

Reference	Document type	Proposed technique	Information to be extracted
[3]	PDF scientific literature papers	Semi-supervised machine learning and data science	Payload extraction, step extraction, and recipe extraction
[4]	Resume in different types of files	Heuristic rules based with free text extraction method	Basic information (Name, Email, Other basic information, Self-evaluation), Work experience, Education information
[6]	PDF novels	Rule-based and machine learning techniques	Extracting the entities and their relation
[8]	PDF publication reports	Rule-based multi-pass sieve approach for classification	Title, Abstract, Body text, Semi-Structured (Figure, Table), Metadata (Header, Footer, Keyword, Author, Journal, Reference)
[9]	PDF scientific literature papers	Supervised and unsupervised machine learning techniques	Title, Author, Affiliations, Email addresses, Abstract, Keywords, Journal, Volume, Issue, Pages, Year, DOI, and References.
[10]	PDF books	Rule-based	Extracting three elements from the table of content: section number, title, and page number
[11]	PDF academic documents	Rule-based and Hidden Markov Model (HMM)	Titles and Authors
[12]	PDF scientific papers	Rule-based	Title, Authors, Abstract
Proposed approach	PDF books (Wiley Publishers)	Rule-based	Title, Authors, ISBN, Publishers, and Year of publication

5. **Conclusions.** In this paper, the problem of information extraction from PDF sources has been addressed. A rule-based approach has been proposed for extracting the information from electronic books in PDF format automatically. The proposed approach proved its capability of extracting the metadata information like Title, Authors, ISBN, Publishers, and Year of Publication from PDF books efficiently. The performance of the proposed method in this work showed an overall accuracy of 94.62% for training and 90.27% for testing. However, the performance of extracting the author still needs to be improved for more improvement. For future work, more investigation can be done to enhance the accuracy of extracting the author's name and other data like table of contents. Also, the research work can be expanded to cover more PDF books from different publishers.

REFERENCES

[1] S. G. Small and L. Medsker, Review of information extraction technologies and applications, *Neural Comput. Appl.*, vol.25, nos.3-4, pp.533-548, DOI: 10.1007/s00521-013-1516-6, 2014.

[2] S. Sarawagi, Information extraction, *Foundation and Trends in Databases*, vol.1, no.3, pp.261-377, 2008.

[3] H. Yang et al., Pipelines for procedural information extraction from scientific literature: Towards recipes using machine learning and data science, *2019 Int. Conf. Doc. Anal. Recognit. Work.*, vol.2, pp.41-46, 2019.
[4] J. Chen, C. Zhang and Z. Niu, A two-step resume information extraction algorithm, *Math. Probl. Eng.*, DOI: 10.1155/2018/5761287, 2018.
[5] D. Ameta and P. M. Jat, Information extraction from Wikipedia articles using DeepDive, *Proc. of 2018 Int. Conf. Commun. Inf. Comput. Technol. (ICCICT2018)*, pp.1-6, DOI: 10.1109/ICCICT.2018.8325869, 2018.
[6] R. Chaniago and M. Khodra, Information extraction on novel text using machine learning and rule-based system, *Int. Conf. Innov. Creat. Inf. Technol.*, pp.1-6, 2017.
[7] H. Bast and C. Korzen, A benchmark and evaluation for text extraction from PDF, *ACM/IEEE Jt. Conf. Digit. Libr.*, pp.1-10, DOI: 10.1109/JCDL.2017.7991564, 2017.
[8] D. D. A. Bui, G. Del Fiol and S. Jonnalagadda, PDF text classification to leverage information extraction from publication reports, *J. Biomed. Inform.*, vol.61, pp.141-148, DOI: 10.1016/j.jbi.2016.03.026, 2016.
[9] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek and Ł. Bolikowski, CERMINE: Automatic extraction of structured metadata from scientific literature, *Int. J. Doc. Anal. Recognit.*, vol.18, no.4, pp.317-335, DOI: 10.1007/s10032-015-0249-8, 2015.
[10] Z. Wu, P. Mitra and C. L. Giles, Table of contents recognition and extraction for heterogeneous book documents, *Proc. Int. Conf. Doc. Anal. Recognition (ICDAR)*, pp.1205-1209, DOI: 10.1109/ICDAR.2013.244, 2013.
[11] W. F. Hsiao, T. M. Chang and E. Thomas, Extracting bibliographical data for PDF documents with HMM and external resources, *Progr. Electron. Libr. Inf. Syst.*, vol.48, no.3, pp.293-313, DOI: 10.1108/PROG-12-2011-0059, 2014.
[12] Z. Guo and H. Jin, A rule-based framework of metadata extraction from scientific papers, *Proc. of the 10th Int. Symp. Distrib. Comput. Appl. to Business, Eng. Sci. (DCABES 2011)*, pp.400-404, DOI: 10.1109/DCABES.2011.14, 2011.
[13] *What Is an ISBN? | International ISBN Agency*, https://www.isbn-international.org/content/what-isbn, Accessed on April 11, 2020.
[14] G. Zaman, H. Mahdin, K. Hussain and Atta-ur-Rahman, Information extraction from semi and unstructured data sources: A systematic literature review, *ICIC Express Letters*, vol.14, no.6, pp.593-603, 2020.
[15] R. Ahmad, M. T. Afzal and M. A. Qadir, Information extraction from PDF sources based on rule-based system using integrated formats, *Semantic Web Evaluation Challenge*, pp.293-308, 2016.
[16] D. Musleh, R. Ahmed, Atta-ur-Rahman and F. Alhaidari, A novel approach to Arabic keyphrase extraction, *ICIC Express Letters, Part B: Applications*, vol.10, no.10, pp.875-884, 2019.
[17] N. Shahzadi, A. Rahman and M. J. Sawar, Semantic network-based classifier of Holy Quran, *International Journal of Computer Applications (IJCA)*, vol.39, no.5, pp.43-47, 2012.
[18] N. Shahzadi, A. Rahman and A. Shaheen, Semantic network based semantic search of religious repository, *International Journal of Computer Applications (IJCA)*, vol.36, no.9, 2011.
[19] A. Rahman, S. A. Alrashed and A. Abraham, User behavior classification and prediction using FRBS and linear regression, *Journal of Information Assurance and Security*, vol.12, no.3, pp.86-93, 2017.
[20] A. Rahman, D. N. Zaidi, M. H. Salam and S. Jamil, User behavior classification using fuzzy rule-based system, *The 13th International Conference on Hybrid Intelligent Systems*, Tunisia, pp.118-123, 2013.
[21] A. Rahman, S. Dash, A. K. Luhach, N. Chilamkurti, S. Baek and Y. Nam, A neuro-fuzzy approach for user behavior classification and prediction, *Journal of Cloud Computing*, vol.8, no.17, 2019.
[22] J. Alhiyafi, A. Rahman, F. Alhaidari and A. Alghamdi, Automatic text categorization using fuzzy semantic network, *SEAHF'19*, 2019.
[23] A. Rahman and S. Das, Big data analysis for teacher recommendation using data mining techniques, *International Journal of Control Theory and Applications*, vol.10, no.18, pp.95-105, 2017.
[24] A. Rahman and S. Das, Data mining for students' trends analysis using apriori algorithm, *International Journal of Control Theory and Applications*, vol.10, no.18, pp.107-115, 2017.
[25] A. Rahman, Teacher assessment and profiling using fuzzy rule based system and apriori algorithm, *International Journal of Computer Applications (IJCA)*, vol.65, no.5, pp.22-28, 2013.
[26] A. Rahman, *Knowledge Representation: A Semantic Network Approach, Handbook of Research on Computational Intelligence Applications in Bioinformatics*, 1st Edition, IGI Global, 2016.
[27] A. Rahman and F. A. Alhaidari, Querying RDF data, *J. of Theoretical and Applied Information Technology*, vol.26, no.22, pp.7599-7614, 2018.

[28] M. Ahmad, U. Farooq, A. Rahman, A. Alqatari, S. Dash and A. K. Luhach, Investigating TYPE constraint for frequent pattern mining, *Journal of Discrete Mathematical Sciences and Cryptography*, vol.22, no.4, pp.605-626, 2019.

[29] M. Ahmad, M. A. Qadir, A. Rahman, R. Zagrouba, F. Alhaidari, T. Ali and F. Zahid, Enhanced query processing over semantic cache for cloud based relational databases, *Journal of Ambient Intelligence and Humanized Computing*, DOI: 10.1007/s12652-020-01943-x, 2020.

[30] A. Rahman and F. A. Alhaidari, The digital library and the archiving system for educational institutes, *Pakistan Journal of Information Management and Libraries (PJIM&L)*, vol.20, no.1, pp.94-117, 2019.

[31] U. L. Yuhana, S. Rochimah, E. M. Yuniarno, A. Rysbekova, A. Tormasi, L. T. Koczy and M. H. Purnomo, A rule-based expert system for automatic question classification in mathematics adaptive assessment on Indonesian elementary school environment, *International Journal of Innovative Computing, Information and Control*, vol.15, no.1, pp.143-161, 2019.