PARDA: A Dataset for Scholarly PDF Document Metadata Extraction Evaluation

Tiantian Fan[1,2], Junming Liu[1,2], Yeliang Qiu[1,2],
Congfeng Jiang[1,2(✉)], Jilin Zhang[1,2], Wei Zhang[1,2], and Jian Wan[2,3]

[1] School of Computer Science and Technology, Hangzhou Dianzi University,
Hangzhou 310018, China
{ttfanx, jmliu, qiuyeliang, cjiang, jilin.zhang,
maghero}@hdu.edu.cn

[2] Key Laboratory of Complex Systems Modeling and Simulation,
Ministry of Education, Hangzhou 310018, China
wanjian@zust.edu.cn

[3] School of Information and Electronic Engineering,
Zhejiang University of Science and Technology, Hangzhou 310023, China

Abstract. Metadata extraction from scholarly PDF documents is the fundamental work of publishing, archiving, digital library construction, bibliometrics, and scientific competitiveness analysis and evaluations. However, different scholarly PDF documents have different layout and document elements, which make it impossible to compare different extract approaches since testers use different source of test documents even if the documents are from the same journal or conference. Therefore, standard datasets based performance evaluation of various extraction approaches can setup a fair and reproducible comparison. In this paper we present a dataset, namely, PARDA(Pdf Analysis and Recognition DAtaset), for performance evaluation and analysis of scholarly documents, especially on metadata extraction, such as title, authors, affiliation, author-affiliation-email matching, year, date, etc. The dataset covers computer science, physics, life science, management, mathematics, and humanities from various publishers including ACM, IEEE, Springer, Elsevier, arXiv, etc. And each document has distinct layouts and appearance in terms of formatting of metadata. We also construct the ground truth metadata in Dublin Core XML format and BibTex format file associated this dataset.

Keywords: Metadata extraction · Dataset · Performance evaluation · Document analysis

1 Introduction

Metadata of a scholarly document, including title, authors, affiliation, author-affiliation-email matching, publishing date, journal information(name, volume, issue) or conference information(conference name, location, date), is the key data for document analysis and digital library construction. Researchers proposed various approaches to extract metadata from different domains [1, 2]. However, different scholarly documents have different sections containing the metadata, with different formatting styles and

H. Gao et al. (Eds.): CollaborateCom 2018, LNICST 268, pp. 417–431, 2019.
https://doi.org/10.1007/978-3-030-12981-1_29

layouts. For example, some documents have headers, some have equations and special symbols in title, some have single-column authorship, some have multi-column authorship, others may have marks preceding or after the authors or affiliations. In some extreme cases, paper title may have different font type, font size and subtitle. This makes it impossible to compare different extraction approaches since they use different dataset with different formatting layouts [3].

As for the automatic metadata extraction, the extraction approaches usually aim to achieve the following goals:

(1) High accuracy. Accuracy is the basic goal and the most important performance metric for automatic metadata extraction;
(2) High adaptivity. Since massive scholarly documents have diverse formatting layouts, the extraction approaches must adapt to different documents and have universal high performance.

Therefore, a standard dataset can not only provide a real test data for performance evaluation, but also serve as a baseline to setup a fair and reproducible comparison among different extraction approaches.

Unfortunately, the PDF specification only defines the basic logical structure to describe the texts, paragraphs, and other layout objects. The accuracy and efficiency of metadata extraction are affected mainly by the following factors:

(1) Implementation variations of visual formatting in PDF documents from different computer programs. The structure and sometimes the source code of document elements is different although they have the same visual appearance, due to the implementation difference of formatting elements (such as paragraphs, segmentations, tables, columns, etc.) and complying with PDF specifications. The missing of standard for metadata and tags in PDF documents makes it very difficult even impossible to process and extract massive documents automatically without human manipulations.
(2) Individual style differences from different authors. For an instance, considering a paper with authors from two different affiliations, the double-column segment of author names and affiliations can be implemented by two columns, or two-column table in single column, or single column with manual alignment. Although they have the same visual appearance, the output documents have different binary source files. Even if a journal or publisher has a formatting template, different authors can meet the requiring formats by their own different means of typesetting and layouts.
(3) Source of PDF documents. Some PDF documents are born digital, some are generated by scanned images. The scanned-PDF documents must undergo the Optical Character Recognition processing before metadata extraction.
(4) Errors in PDF document itself. Some authors do not comply with the formatting template of the journal or publisher due to different writing behaviors, careless mistakes, personal styles or cultural background. Sometimes even the documents from the same journal issue have different formatting layouts or compilation errors due to mistakes in publishing process or digital preservation.

However, the creation of such standard datasets for layout analysis and metadata extraction is a costly process in terms of data selection, acquisition, and annotation [4–6]. Such cost of acquisition and annotation of real datasets sometimes forces the use of synthetic data. The research community devoted a big effort to the generation of public dataset for document analysis, such as document imaging, page analysis, and graphics recognition [7–10]. However, the existing datasets are for general purpose imaged document analysis, not suitable for metadata extraction.

It is ideal that the dataset contains real, representative, and comprehensive documents from different sources. However, in order to collect enough page layouts from different documents, researchers must browse documents as widely as possible to fulfill the realistic and comprehensive goal. Since different institutions have different accessibilities to different sources of scholarly documents due to institutional subscription or publicly access, such dataset must also be extended to make it more comprehensive when new layouts are found and added [11–14].

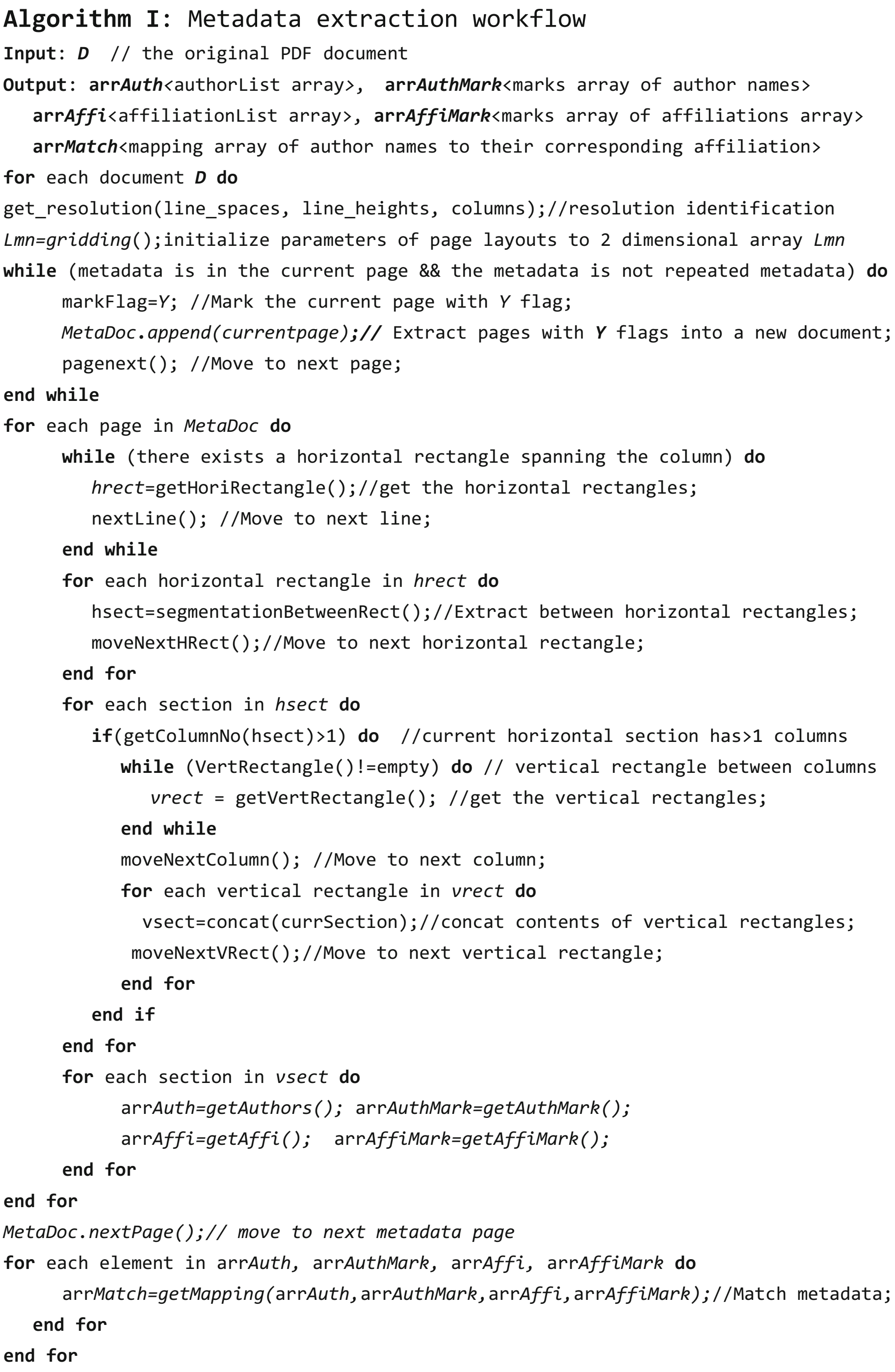
In this paper we present a dataset, namely, PARDA(Pdf Analysis and Recognition DAtaset), for performance evaluation and analysis of scholarly documents, especially on metadata extraction, such as title, authors, affiliation, author-affiliation-email matching, year, date, etc. We collect 147 real scholarly documents from the published journals, magazines, and conference proceedings. The dataset covers computer science, physics, life science, management, mathematics, and humanities. And each document has distinct layouts and appearance in terms of formatting of metadata. We also construct the ground truth metadata in xml file associated this dataset.

We identify different occurrences of formatting types in real corpus are summarized, including 6 types of titles, 14 types of author names, and 30 types of author-affiliation combinations. This summary can serve as a reference template database and baseline for general purpose metadata extraction and implementation. Moreover, various frequently occurred words in headers and affiliations are summarized to improve the accuracy of document segmentation and metadata extraction. These frequently occurred words can be used extensively if they are accompanied with domain-specific words and semantics from other sources.

The remainder of the paper is organized as follows: In Sect. 2 we give the metadata extraction workflow. In Sect. 3 we give the dataset description based on layout categorization of scholarly documents. In Sect. 4 we provide the ground truth of the dataset. We conclude the paper in Sect. 5.

2 Metadata Extraction Workflow

In order to extract metadata from scholarly documents, both character stream (i.e., pure texts) and formatting stream (line height, font type, font size, character location, etc.) are extracted in parallel. When a PDF document is being processed, explicit and implicit formatting semantics are used for segmentation and metadata extraction. The targeted pages are segmented into header, footnote, title, authors, affiliations, and address (if any). Each segment may have multiple elements. The element identification (images, tables, figures, etc.) and resolution identification including line spaces, line heights, and columns will be done before metadata extraction. The procedure is illustrated in Fig. 1 as algorithm I.



```
Algorithm I: Metadata extraction workflow
Input: D  // the original PDF document
Output: arrAuth<authorList array>,  arrAuthMark<marks array of author names>
   arrAffi<affiliationList array>, arrAffiMark<marks array of affiliations array>
   arrMatch<mapping array of author names to their corresponding affiliation>
for each document D do
get_resolution(line_spaces, line_heights, columns);//resolution identification
Lmn=gridding();initialize parameters of page layouts to 2 dimensional array Lmn
while (metadata is in the current page && the metadata is not repeated metadata) do
      markFlag=Y; //Mark the current page with Y flag;
      MetaDoc.append(currentpage);// Extract pages with Y flags into a new document;
      pagenext(); //Move to next page;
end while
for each page in MetaDoc do
      while (there exists a horizontal rectangle spanning the column) do
         hrect=getHoriRectangle();//get the horizontal rectangles;
         nextLine(); //Move to next line;
      end while
      for each horizontal rectangle in hrect do
         hsect=segmentationBetweenRect();//Extract between horizontal rectangles;
         moveNextHRect();//Move to next horizontal rectangle;
      end for
      for each section in hsect do
         if(getColumnNo(hsect)>1) do  //current horizontal section has>1 columns
            while (VertRectangle()!=empty) do // vertical rectangle between columns
               vrect = getVertRectangle(); //get the vertical rectangles;
            end while
            moveNextColumn(); //Move to next column;
            for each vertical rectangle in vrect do
              vsect=concat(currSection);//concat contents of vertical rectangles;
             moveNextVRect();//Move to next vertical rectangle;
            end for
         end if
      end for
      for each section in vsect do
            arrAuth=getAuthors(); arrAuthMark=getAuthMark();
            arrAffi=getAffi();  arrAffiMark=getAffiMark();
      end for
end for
MetaDoc.nextPage();// move to next metadata page
for each element in arrAuth, arrAuthMark, arrAffi, arrAffiMark do
      arrMatch=getMapping(arrAuth,arrAuthMark,arrAffi,arrAffiMark);//Match metadata;
   end for
end for
```

Fig. 1. Metadata extraction algorithm

In the preprocessing stage, the document is preprocessed, including when a document is loaded. Usually the first page (or multiple pages if the document has longer section of authors and affiliations) will be truncated into the targeted pages for post-processing, especially the metadata extraction. In our approach, the metadata are extracted based on formatting template database, explicit semantics, and implicit semantics. And the authors are matched with their affiliations according to the marks or implicit semantics. To improve the accuracy, the metadata are verified by redundant information in different segments. For example, author affiliation may appear after the paper title and before the abstract, and it may also be located in the footnote. Such redundant information can verify the author and affiliation matching automatically by extra verification process.

3 Dataset Description

3.1 Layout Categorization Methodology

There are many journals, magazines, and proceedings of different publishers that publish a large number of scientific articles, most of them have similar sections, such as headers, titles, authors, affiliations, abstract, footnote, body texts, the references, figures and tables, etc. However, different scholarly documents may have different formatting styles and layouts, although they have similar reading order and logical metadata. This makes it difficult to extract and parse metadata widely using the same formatting templates or by machine learning techniques unless the machine has exhaustive templates or training datasets. To make the dataset as pervasive as possible, we classify the formatting layouts of existing scholarly documents by sections and try to cover as many as possible layouts from our collections and observations. We list the categorization in Table 1.

Table 1. Layout categorization.

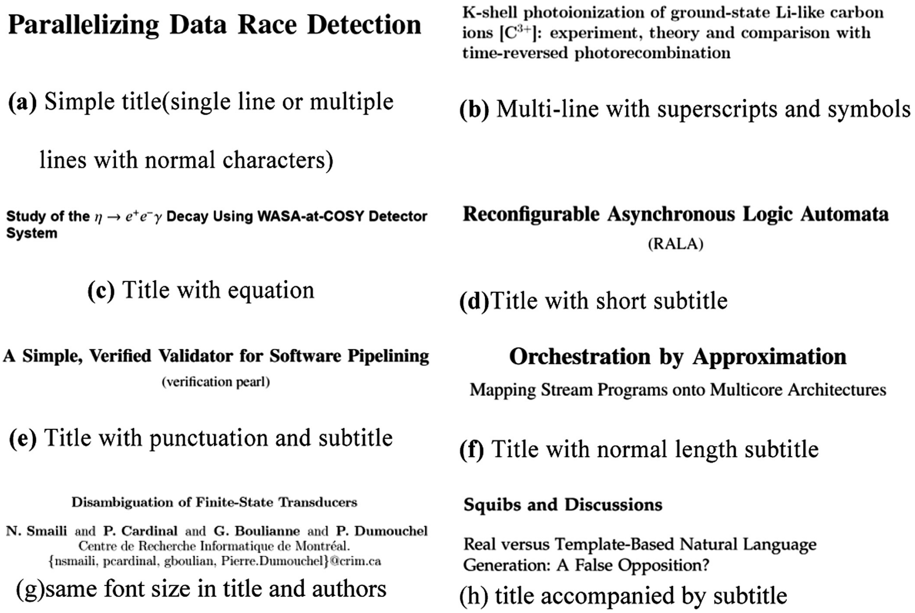
Section	Containing elements(if any)
Header	Publisher, journal type, journal name, publishing dada of papers, paper status, date, URL, DOI, special symbols, embedded image
Title	Main title, subtitle, equations, symbols, special characters, different font types, font sizes, line heights
Authorship	Multiple authors, single column, multiple columns
Author name	Different blocks of first name, middle name, and last name, special characters, diacritics and dialects
Affiliation	Full name, abbreviation, multiple lines, phone number, country, state/province, city, zip code
Email	{}, “lastname”, “firstname”, embedded image
Visual order	Relative order of Authorship, affiliation, abstract, and footnote
Footnote	Affiliation, conference location, date paper submission and processing details

We select the dataset documents according to the existence and formatting style of each element of each section and group them into different categories.

3.2 Templates of Title and First Page

Please note that although some researchers in the same academic domain tend to use similar formatting styles, there are no dominant or prevalent formatting styles or patterns in all the research fields. Actually the formatting patterns are scattered and occurred across different research fields. This requires the scalable and extensible approach to extract metadata from such papers other than universal and fixed templates. We select from the publicly accessed and our institutional subscribed scholarly documents to construct the dataset. The selected documents come from various academic fields, including computer science, physics, life science, management, mathematics, and humanities from publishers such as ACM, IEEE, Springer, Elsevier, arXiv, etc. These documents vary from journals, magazines to conference proceedings. We list some examples in the following figures. For example, some journals of Elsevier, like FGCS, JPDC, JSA, the journal name has larger font size than paper title. In order to adapt to more scholarly documents, we collect a comprehensive formatting templates from real published documents including title, author name affiliation, and address.

In Fig. 2 we list most used title templates.



Parallelizing Data Race Detection

(a) Simple title(single line or multiple lines with normal characters)

K-shell photoionization of ground-state Li-like carbon ions [C^{3+}]: experiment, theory and comparison with time-reversed photorecombination

(b) Multi-line with superscripts and symbols

Study of the $\eta \rightarrow e^+e^-\gamma$ Decay Using WASA-at-COSY Detector System

(c) Title with equation

Reconfigurable Asynchronous Logic Automata

(RALA)

(d) Title with short subtitle

A Simple, Verified Validator for Software Pipelining

(verification pearl)

(e) Title with punctuation and subtitle

Orchestration by Approximation

Mapping Stream Programs onto Multicore Architectures

(f) Title with normal length subtitle

Disambiguation of Finite-State Transducers

N. Smaili and P. Cardinal and G. Boulianne and P. Dumouchel
Centre de Recherche Informatique de Montréal.
{nsmaili, pcardinal, gboulian, Pierre.Dumouchel}@crim.ca

(g)same font size in title and authors

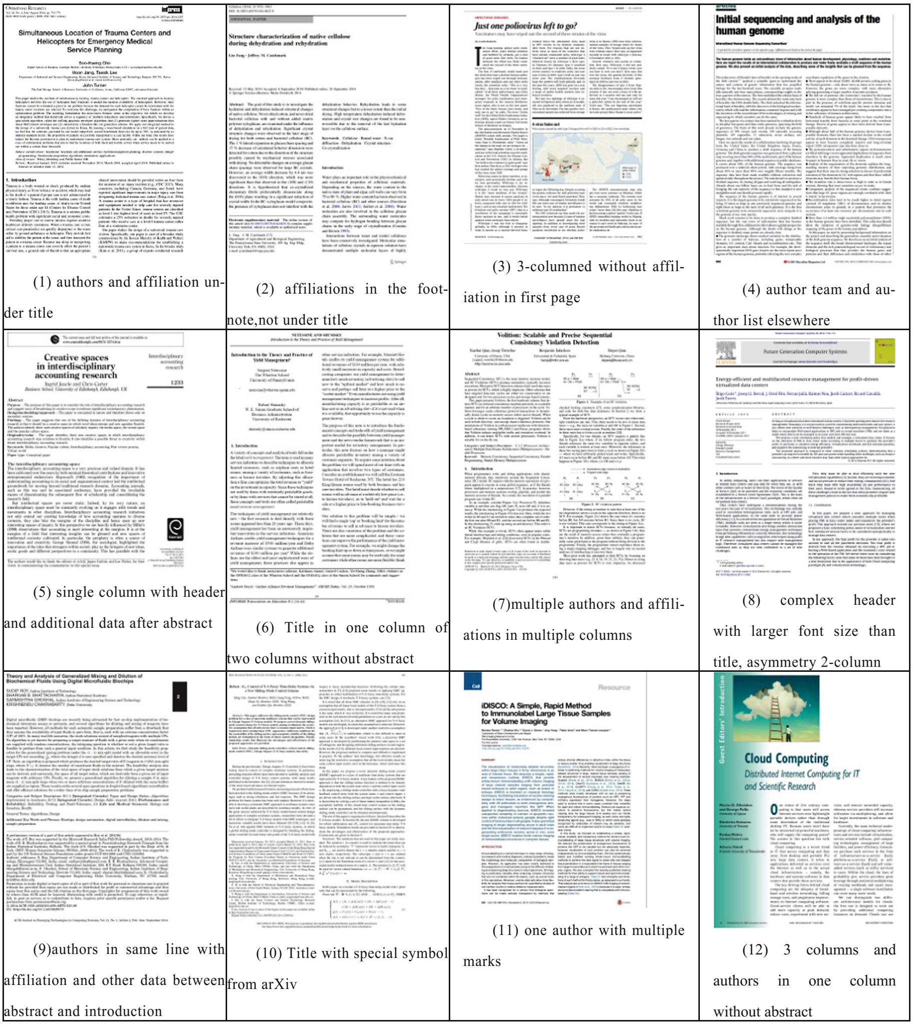
Squibs and Discussions

Real versus Template-Based Natural Language Generation: A False Opposition?

(h) title accompanied by subtitle

Fig. 2. Frequently used titles

In Fig. 3 we list some layout examples from real documents.



(1) authors and affiliation under title

(2) affiliations in the footnote,not under title

(3) 3-columned without affiliation in first page

(4) author team and author list elsewhere

(5) single column with header and additional data after abstract

(6) Title in one column of two columns without abstract

(7)multiple authors and affiliations in multiple columns

(8) complex header with larger font size than title, asymmetry 2-column

(9)authors in same line with affiliation and other data between abstract and introduction

(10) Title with special symbol from arXiv

(11) one author with multiple marks

(12) 3 columns and authors in one column without abstract

Fig. 3. Sample layouts from real documents.

We only list a handful of the real documents that have different layouts in Fig. 3. However, many documents have diverse layouts and may have combination of multiple layouts on header, title, authors, affiliation, abstract, footnote, etc.

For example, the title of a scientific paper may have the following formatting layouts except normal titles:

(a) The title has multiple lines with superscripts and symbols
(b) The title has equation in it

(c) The title has subtitle
(d) The title has special characters and punctuations

Moreover, the abstract section is easier to extract if it has preceding word like "abstract" and "summary" and the abstract section has uniform formatting layouts (at least 3 consecutive lines). Even if the paper has many authors and affiliations and the abstract is not on the first page, the abstract section can still be extracted if they have these preceding words. In some cases if the paper has no such preceding words but does have abstract section, we can still extract the abstract section according to the following two criteria: (1) There are at least 3 consecutive lines with uniform formatting layouts; and (2) In these lines there are words or phrases that merely appear in author names or affiliations, such as "we", "recently", "this paper", "proposed", review", etc. Such collection can be extended easily for further and newly added documents. In this paper we locate the beginning of the abstract if: (1) one line is not belonging to affiliation (according to feature words of affiliation), (2) its length is greater than 40, and (3) the following consecutive lines are not belonging to affiliation. Note that here the length 40 is empirically set by observations and can be tuned to other values. Since abstract is the concise summary of the whole paper using descriptive words and usually has no overlapping words with affiliation, the abstract can be identified by the formatting layouts and their changes.

3.3 Header Templates

The first line is title if a paper has no header. However, there are many papers that have headers (mostly in single column). Although the contents in headers vary in different papers, they can be divided into images and pure texts.

As for the images in header, it can be identified and filtered by the binary values of the parsed text stream. Usually they are not normal ASCII characters or the file has special pointer to the image objects. Therefore, the forgoing input can be regarded as header and then filtered until there are consecutive ASCII characters. When it comes to header of pure texts, the occurrence of feature words can be identified and the line can be regarded as header and filtered. Some feature words are summarized in Table 2.

Table 2. Frequently occurred feature words and symbols in pure text headers

Class	Examples
Publisher	IEEE, ACM, arXiv, Publishing, Company
Journal type	Journal, J., Proc., Proceedings, Conferences Workshop, Symposium, Letters, Lett., Transaction, Trans.
Journal name	Complete names and abbreviations of terms, Sci., Tech., Med., Com., Inf.
Publishing data	Volume, Vol., Issue, No., page, pp., article
Paper status	Manuscript, preprint, accepted, submitted, publication, pub.
Date and time	Year and month, January to December, Jan. to Dec.
Special format string	Partial or complete website URL, beginning and end of page number, DOI digits, date
function word	on, and, in, of, for
Special symbol	&, /, ©, ®

Although Table 2 is not a complete list of feature words in header, it can be easily extended and contain newly observed words. In this paper we use this list to identify and filter header texts and achieve high accuracy for most scholarly documents. Moreover, data in header and footnote can mutually be verified and improve the accuracy of metadata extraction.

3.4 Authorship Templates

In order to extract and parse the metadata, the fundamental work is to locate the header, title, authors, affiliation, abstract, and the footnote. Specifically, if the header and the abstract section are located, the metadata, such as title, author, affiliation, and address must be between the header and the abstract for normal document and can be extracted by their formatting templates respectively. Therefore, the content positioning and segmentation is the first step before metadata extraction. Since most of the metadata is located between the running header and the abstract, the metadata can be located as long as the running header and abstract is detected and identified.

The authorship section is another important source of metadata. But different publisher has different styles for authorship. For example, some journals require single-column authorship. Some conference proceedings requires multiple column authorship if the authors have more than one affiliation. In order to extract author names, the scholarly documents can be simply categorized into two classes: author names with symbols or marks, and author names without any symbol or mark. If all the author names have no symbol or mark, all the authors may belong to the same affiliation (if the paper has only one affiliation), or authors belong to multiple affiliations separated by columns. Or if at least one author has symbol or mark, her affiliation may also have symbol or mark.

If the authorship is single column, it still can have multiple formatting layouts like:

(a) Multiple separated authors belong to one affiliation and authors names in different line from affiliation;
(b) Multiple authors without separators belong to one affiliation and author names in different line from affiliation;
(c) Multiple separated authors belong to multiple affiliations, each author name in same line with its affiliation
(d) There are other symbols or marks in the authorship, like “Jr.”, comma, parenthesis, “and”, and email address in the same line or different line with affiliation
(e) Authors are only separated by white spaces.

However, in most scholarly documents, if the authorship section is formatted in multiple columns, it is indicating that the paper has more than one author and maybe multiple affiliations. In this situation, each author may belong to different affiliation, or all the authors belong to the same affiliation, or some authors belong to the same affiliation and the remaining authors belong to other affiliation. Then the authorship may have the following variations:

(a) Multiple authors belonging to one affiliation and they are in the same column;
(b) Multiple authors belonging to one affiliation and they are in the different columns, each column has its own affiliation name;
(c) Multiple authors belonging to one affiliation and they are in the different columns, all column sharing their affiliation name;
(d) Multiple authors belonging to multiple affiliations respectively.

In many cases, if one paper has multiple affiliations, it can also be achieved by single column formatting with marks in superscripts locations in author names and affiliations avoiding multiple columns formatting. The introduction of marks can result in many variations of authorship layouts of single column, such as:

(a) Authors are numbered by digital, lower case characters, upper case characters, special symbols, Greek symbols, or embedded images;
(b) One author has more than one marks and one of them indicating correspondence author or co-first-author;
(c) Multiple affiliations in different lines or the same line, and the author belonging to first affiliation has no mark;
(d) Multiple affiliations in one line and each affiliation has a mark, and author mark is preceding the comma separator;
(e) Affiliation name is after the author name and in a parenthesis;
(f) Five affiliations with marks of digitals, characters and symbols of correspondence author;
(g) Marks are preceding the author names.

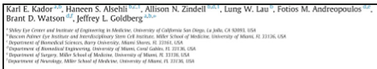
Sometimes papers jointly use multiple columns and marks for segmentation of author names and their affiliations. Such examples of multiple columns formatting with marks in author names and affiliations as:

(a) Two affiliations, two columns, one mark for the only one author, the remaining authors belong to the same affiliation without marks;
(b) Three affiliations, each affiliation has one mark, combination of single column and two columns;
(c) Four affiliations, only one author has mark for correspondence authors, combination of single column and two columns;
(d) Four affiliations, each author name separated in two lines.

We list some authorship examples in Fig. 4.

In some research fields such as life science and high energy physics, international cooperation results in a huge number of authors and affiliations. It's difficult for the existing approach to extract authors and their affiliations with high accuracy. We collect the dataset that covers the entire above mentioned formatting layout from real documents. It reflects the breadth of real documents and can serve as the basic testing data for in-depth performance evaluation.

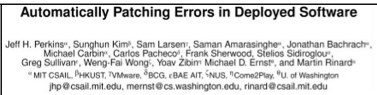
Due to copyright constraints, rewriting all the real documents as their appearance and distribute all the rewritten PDF files can make this dataset publicly accessible, the rewritten files may contain different compiling codes as the original documents. Therefore, the rewritten documents are useful for OCR based approaches after they are



Karl E. Kador[a,b], Haneen S. Alsehli[b,c,1], Allison N. Zindell[b,d,1], Lung W. Lau[b], Fotios M. Andreopoulos[d,e],
Brant D. Watson[d,f], Jeffrey L. Goldberg[a,b,*]

[a] Shiley Eye Center and Institute of Engineering in Medicine, University of California San Diego, La Jolla, CA 92093, USA
[b] Bascom Palmer Eye Institute and Interdisciplinary Stem Cell Institute, Miller School of Medicine, University of Miami, FL 33136, USA
[c] Department of Biomedical Sciences, Barry University, Miami Shores, FL 33161, USA
[d] Department of Biomedical Engineering, University of Miami, Coral Gables, FL 33136, USA
[e] Department of Surgery, Miller School of Medicine, University of Miami, FL 33136, USA
[f] Department of Neurology, Miller School of Medicine, University of Miami, FL 33136, USA

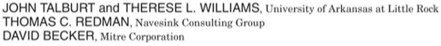
(1)Single column authorship with marks and separators,all authors first and all affiliations together



Automatically Patching Errors in Deployed Software

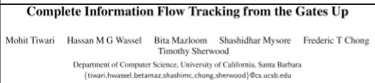
Jeff H. Perkins[α], Sunghun Kim[β], Sam Larsen[γ], Saman Amarasinghe[α], Jonathan Bachrach[α],
Michael Carbin[α], Carlos Pacheco[δ], Frank Sherwood, Stelios Sidiroglou[α],
Greg Sullivan[ε], Weng-Fai Wong[ζ], Yoav Zibin[η] Michael D. Ernst[θ], and Martin Rinard[α]
[α] MIT CSAIL, [β]HKUST, [γ]VMware, [δ]BCG, [ε]BAE AIT, [ζ]NUS, [η]Come2Play, [θ]U. of Washington
jhp@csail.mit.edu, mernst@cs.washington.edu, rinard@csail.mit.edu

(2) Numbering with Greek symbols and multiple affiliations in one line



JOHN TALBURT and THERESE L. WILLIAMS, University of Arkansas at Little Rock
THOMAS C. REDMAN, Navesink Consulting Group
DAVID BECKER, Mitre Corporation

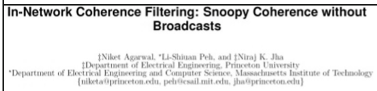
(3)Author in same line with affiliation



Complete Information Flow Tracking from the Gates Up

Mohit Tiwari Hassan M G Wassel Bita Mazloom Shashidhar Mysore Frederic T Chong
Timothy Sherwood
Department of Computer Science, University of California, Santa Barbara
{tiwari,hwassel,betamaz,shashimc,chong,sherwood}@cs.ucsb.edu

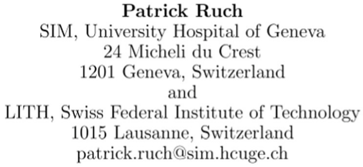
(4) Four-blocks authorname without separator, author name in different line from affiliation



In-Network Coherence Filtering: Snoopy Coherence without Broadcasts

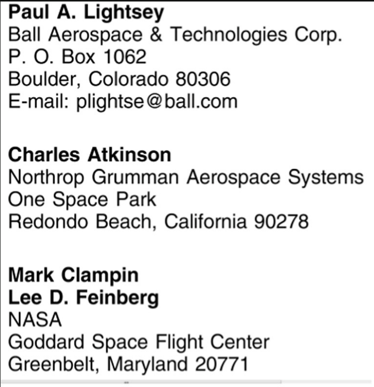
‡Niket Agarwal, *Li-Shiuan Peh, and ‡Niraj K. Jha
‡Department of Electrical Engineering, Princeton University
*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology
{niketa@princeton.edu, peh@csail.mit.edu, jha@princeton.edu}

(5) Mark preceding the author name



Patrick Ruch
SIM, University Hospital of Geneva
24 Micheli du Crest
1201 Geneva, Switzerland
and
LITH, Swiss Federal Institute of Technology
1015 Lausanne, Switzerland
patrick.ruch@sim.hcuge.ch

(6) single column authorship one author belonging to two affiliations without marks



Paul A. Lightsey
Ball Aerospace & Technologies Corp.
P. O. Box 1062
Boulder, Colorado 80306
E-mail: plightse@ball.com

Charles Atkinson
Northrop Grumman Aerospace Systems
One Space Park
Redondo Beach, California 90278

Mark Clampin
Lee D. Feinberg
NASA
Goddard Space Flight Center
Greenbelt, Maryland 20771

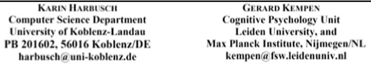
(7) single column authorship with author-affiliation block one by one

Kees van Deemter*
University of Aberdeen

Emiel Krahmer†
Tilburg University

Mariët Theune‡
University of Twente

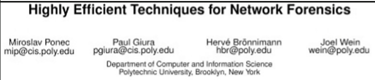
(8)Multi-column authorship with marks



KARIN HARBUSCH
Computer Science Department
University of Koblenz-Landau
PB 201602, 56016 Koblenz/DE
harbusch@uni-koblenz.de

GERARD KEMPEN
Cognitive Psychology Unit
Leiden University, and
Max Planck Institute, Nijmegen/NL
kempen@fsw.leidenuniv.nl

(9)multi-column authorsship with fully captitalized names



Highly Efficient Techniques for Network Forensics

Miroslav Ponec
mip@cis.poly.edu

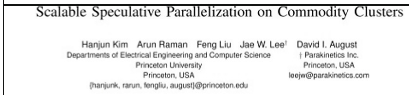
Paul Giura
pgiura@cis.poly.edu

Hervé Brönnimann
hbr@poly.edu

Joel Wein
wein@poly.edu

Department of Computer and Information Science
Polytechnic University, Brooklyn, New York

(10) Multiple authors belonging to one affiliation and they are in the different columns, all column sharing their affiliation name



Scalable Speculative Parallelization on Commodity Clusters

Hanjun Kim Arun Raman Feng Liu Jae W. Lee† David I. August
Departments of Electrical Engineering and Computer Science
Princeton University
Princeton, USA
{hanjunk, rarun, fengliu, august}@princeton.edu

† Parakinetics Inc.
Princeton, USA
leejw@parakinetics.com

(11) Two affiliations ,two columns, one mark for the only one author ,the remaining authors belong to the same affiliation without annotation

Fig. 4. Authorship examples of single and multiple columns formatting with/without marks

scanned as images for post-processing and extraction. In order to make this dataset equally standard for original compilation based metadata extraction, such as PDFBox [15], we write an alternative file containing all the URLs of the selected documents to be publicly downloaded.

The corresponding document can be downloaded according to the institutional accessibility of each individual. If some documents can't be downloaded or accessed, the performance evaluation can also be conducted as long as the missing document ID is stated so that the evaluation on the subset of PARDA can still be comparable.

4 Groundtruthing

The availability of high accuracy metadata archives of standard PDF documents becomes the key issue for successful implementation of metadata extraction software. It is a very difficult task for collecting a large number of high quality PDF documents by using traditional methods to meet all the requirements for automatic metadata extraction process. We select the scholarly documents from the publicly accessed and our institutional subscribed database to construct the dataset. We choose XML as the PARDA ground-truth format for metadata extraction performance evaluation and comparison. The XML format is compatible with other established metadata description formats. However, the existing metadata specification formats do not have any explicit metadata entry of connection of authors and their affiliations. The CERIF (Common European Research Information Format) contextual metadata covers persons, organizations, projects, products, publications, patents, facilities, equipment, funding, and – most importantly – the relationships between them [16]. For example, the de-facto metadata standard, the Dublin Core(DC) standard [17], has 15 elements in a metadata of simple DC(more on qualified DC), ranging from title, creator, publisher, subject to language. In DC, the metadata landscape is currently characterized in terms of four "levels" of interoperability,i.e., Level 1 (Shared term definitions), Level 2 (Formal semantic interoperability), Level 3 (Description Set syntactic interoperability) and Level 4 (Description Set Profile interoperability). But it does not have any direct entry for connections for authors and their affiliations. The BibTex does not have any entry or field for authors and their corresponding affiliations either, although it has an "institution" field. For a paper that has multiple authors affiliated to multiple affiliations, one "institution" is not sufficient to present the authorship correctly.

Therefore, in order to make the ground truth file compatible with the Dublin Core standard, we use an extended sub-element "affiliation" of creator provide the connection for authors and their affiliations. If an author has multiple affiliations, the creator can also have multiple "affiliation" entries as they appear in the original paper. Moreover, we use the description entry to store the abstract.

We give an example record of the ground true file(part) in DC format in Fig. 5.

```
<?xml version="1.0"?>
<metadata
  xmlns="http://example.org/myapp/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://example.org/myapp/
http://example.org/myapp/schema.xsd"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/">
  <dc:title>
    Automatically Patching Errors in Deployed Software
  </dc:title>
  <dc:subject>
    Error Handling and Recovery,Monitors
  </dc:subject>
  <dc:subject>
    Corrections,Enhancement
  </dc:subject>
  <dc:subject>
    Invasive Software
  </dc:subject>
<dc:creator>
Jeff H. Perkins
 <affiliation>
MIT, Cambridge, MA, USA
 </affiliation>
  </dc:creator>
<dc:creator>
Sunghun Kim
 <affiliation>
HKUST, Hong Kong
 </affiliation>
  </dc:creator>
<dc:creator>
Sam Larsen
 <affiliation> VMWare, Redwood, CA, USA </affiliation>
  </dc:creator>
  <dc:description>
    We present ClearView, a system for automatically……
  </dc:description>
  <dc:publisher>ACM </dc:publisher>
  <dc:identifier xsi:type="dcterms:URI">
    http://dl.acm.org/citation.cfm?id=1629575.1629585
  </dc:identifier>
</metadata>
```

Fig. 5. Sample ground truth file of metadata in DC format

Alternatively, we also provide a dedicated BibTex file with institution field for each document which containing the affiliations that corresponds to authors as they appear in the author field. In order to make sure that each author is correctly connected with their affiliation, every author in the author field must have a string in the institution field, even if the institution is null. We use semicolon(;), not "and" in the institution field to separate multiple affiliations, instead of "and" in author field. We give an example record of the ground true file(part) in BibTex format in Fig. 6.

```
@inproceedings{Paper001,
  author    = {Jeff H. Perkins and
               Sunghun Kim and
               Samuel Larsen......},
  institution={MIT,   Cambridge,   MA,   USA;HKUST,   Hong
Kong;VMWare, Redwood, CA, USA;......}
  title      = {Automatically patching errors in deployed
software},
  booktitle = {Proceedings of the 22nd {ACM} Symposium on
Operating Systems Principles 2009, {SOSP} 2009, Big Sky,
Montana, USA, October 11-14, 2009},
  pages     = {87--102},
  year      = {2009},
  url                                                      =
{http://doi.acm.org/10.1145/1629575.1629585},
  doi       = {10.1145/1629575.1629585}
}
```

Fig. 6. Sample ground truth file of metadata in BibTex format

5 Conclusion Remarks

In this paper we presented a new dataset, PARDA, for performance evaluation of metadata extraction from scholarly documents. Although there are some datasets for layout analysis and performance evaluation, they are not suitable for metadata extraction due to their coverage of variations in sections including title and authorship. We selected scholarly documents widely from publicly accessed and our university subscribed sources. PARDA provides comprehensive and accurate ground truth description file and associated metadata for a wide variety of layouts that have complex combinations of titles, authors, and affiliations, address, and emails (if any).

The original URL of each document is collected in independent files. The ground truth files of metadata are both in Dublin Core and BibTex format.

We are currently rewriting all the real dataset documents with forged stuff according to their layouts, especially focusing on the complex combinations of titles, authors, and affiliations, address, and emails so that the dataset can be freely downloaded without charge of copyright or permissions. Note that the newly rewritten PDF documents are more useful for OCR based metadata extraction approaches because our rewritten documents do not keep it exactly same with the original PDF documents as they are in the publisher databases.

Acknowledgment. The funding support of this work by Natural Science Fund of China (No. 61472109, No. 61572163, No. 61672200, and No. 61772165) is greatly appreciated.

References

1. Lipinski, M., Yao, K., Breitinger, C., Beel, J., Gipp, B.: Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In: JCDL 2013 Indianapolis, Indiana, USA, 22–26 July 2013, pp. 385–386 (2010)
2. Do, H.H.N., Chandrasekaran, M.K., Cho, P.S., Kan, M.Y.: Extracting and matching authors and affiliations in scholarly documents. In: JCDL 2013, Indianapolis, Indiana, USA, 22–26 July 2013, pp. 219–228 (2013)
3. Jiang, C., Liu, J., Ou, D., Wang, Y., Yu, L.: Implicit semantics based metadata extraction and matching of scholarly documents. J. Database Manag. (JDM) **29**, 1–22 (2018). https://doi.org/10.4018/JDM.2018040101
4. Tkaczyk, D., Szostek, P., Bolikowski, Ł.: GROTOAP2—the methodology of creating a large ground truth dataset of scientific articles. **20**(11/12) (2014)
5. Märgner, V., El Abed, H.: Tools and metrics for document analysis systems evaluation. In: Doermann, D., Tombre, K. (eds.) Handbook of Document Image Processing and Recognition, pp. 1011–1036
6. Antonacopoulos, A., Bridson, D., Papadopoulos, C., Pletschacher, S.: A realistic dataset for performance evaluation of document layout analysis. In: 10th International Conference on Document Analysis and Recognition, ICDAR 2005 (2005)
7. Nartker, T.A., Rice, S.V., Lumos, S.E.: Software tools and test data for research and testing of page-reading OCR systems. In: SPIE and IS&T (2005)
8. Todoran, L., Worring, M., Smeulders, A.W.M.: The UvA color document dataset. IJDAR **7**, 228–240 (2005)
9. Becker, C., Duretec, K.: Free benchmark corpora for preservation experiments: using model-driven engineering to generate data sets. In: JCDL 2013, pp. 349–358 (2013)
10. Caragea, C., et al.: CiteSeer^x: a scholarly big dataset. In: de Rijke, Maarten, et al. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 311–322. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06028-6_26
11. Antonacopoulos, A., Karatzas, D., Bridson, D.: Ground truth for layout analysis performance evaluation. In: IAPR International Workshop on Document Analysis Systems, DAS 2006 (2006)
12. Tkaczyk, D., Czeczko, A., Rusek, K., Bolikowski, L., Bogacewicz, R.: GROTOAP: ground truth for open access publications. In: JCDL 2012, pp. 381–382 (2012)
13. Tao, X., Tang, Z., Xu, C., Gao, L.: Ground-truth and performance evaluation for page layout analysis of born-digital documents. In: 2014 11th IAPR International Workshop on Document Analysis Systems, DAS 2014, pp. 247–251 (2014)
14. Valveny, E.: Datasets and annotations for document analysis and recognition. In: Doermann, D., Tombre, K. (eds.) Handbook of Document Image Processing and Recognition, pp. 983–1009
15. http://pdfbox.apache.org
16. Jeffery, K.G., Houssos, N., Jörg, B., Asserson, A.: Research information management: the CERIF approach. Int. J. Metadata Semant. Ontol. **9**, 5–14 (2014)
17. http://dublincore.org/