**Journal of Database Management**
Volume 29 • Issue 2 • April-June 2018

# Implicit Semantics Based Metadata Extraction and Matching of Scholarly Documents

Congfeng Jiang, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China
Junming Liu, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China
Dongyang Ou, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China
Yumei Wang, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China
Lifeng Yu, Hithink RoyalFlush Information Network Co., Ltd., Hangzhou, China

## ABSTRACT

The authors propose to use formatting templates and implicit formatting semantics information for automatic metadata identification and segmentation. The pure texts and their corresponding formatting information including line height, font type, and font size, are recognized in parallel to guide metadata identification. The authors use implicit formatting semantics, such as the change of formatting, formatting templates and implications, explicit formatting layouts, as well as predefined frequently occurred keywords database to increase the extraction accuracy. Unlike other OCR-based approaches, the authors use open source PDFBox package as the basic preprocessing tool to get pure texts and formatting values of the document contents. On top of PDFBox they built their own pipeline program, namely, PAXAT, to implement their approaches for metadata extraction. 10177 papers from arXiv, ACM, ACL and other publicly accessed and institution-subscribed sources are tested. The overall extraction accuracy of title, authors, affiliations, author-affiliation matching are 0.9798, 0.9425, 0.9298, and 0.9109, respectively.

## KEYWORDS

Formatting Semantics, Information Retrieval, Metadata Extraction, PDF Document, Template

## INTRODUCTION

With the advance in digital libraries and online publishing, the quantity of online scholarly documents and born-digital documents is increasing significantly, and the digital transition away from print continues (Ware & Mabe, 2015). Open-access initiatives and platforms such as publisher-owned websites and arXiv.org also make the personal digital library not only possible, but prevalent for researchers and scientists (Laakso & Björk, 2013). Portable Document Format (PDF) has become the de facto standard of producing, delivering, exchanging, and archiving scholarly documents because of its independence of visual information and source-file structure. Therefore, automatic extraction of

DOI: 10.4018/JDM.2018040101

metadata from such PDF documents is the fundamental work of digital preservation, bibliometrics, and scientific competitiveness analysis and evaluations (Suh & Lee, 2001, Zhao, 2010, Fiori et al., 2014).

Metadata is defined by some as data about data and is used by both humans and computers. Digital libraries must ensure that computer systems can both read and "understand" metadata (Lee, Kim, & Kim, 2001). This requires formal syntax and defined semantics—humans can overcome inconsistencies and vagueness, but computers cannot (Jeffery & Koskela, 2015). This chapter predominantly refers to scholarly documents from scientific literature rather than journalistic magazines, and refines the metadata to include title, author names, affiliations, and author-affiliation matching. We try to extract such metadata because they have versatile formatting styles and change frequently and significantly in different scholarly papers. The remaining metadata, such as publishing source, journal name, publication date, volume, and issue number, are outside the scope of this chapter, although they can be extracted similarly by approaches proposed here.

Unfortunately, the PDF specification only defines the basic logical structure to describe the texts, paragraphs, and other layout objects. The PDF specification is optimized for content presentation, but lacks structural information on the content, especially the structure in reading order. The absence of explicit tags or discernible labels for many elements in documents is the main obstacle to machines automatically identifying the metadata. Moreover, such absence of uniform formatting and layout standards makes it very hard, sometimes even impossible, to extract metadata from different scholarly documents appearing in different publishing sources. The accuracy and efficiency of metadata extraction are affected mainly by implementation variations of visual formatting in PDF documents from different computer programs; individual style differences from different authors; source compilation of PDF documents; and errors in the PDF document itself. The paper's title is an example. The title has obvious visual formatting features, such as location on the first page, largest font size, or centered text. Although it is traditionally believed that the title with its simple formatting semantics is easy to extract, the extraction accuracy is still affected by the existence of the following factors:

- The title is not located at the beginning of the first page;
- The title has multiple lines;
- The title has a subtitle besides the main title;
- The title text has multiple font types and sizes;
- The title has special characters, a digital string, or mathematical or physics equations;
- There is a page header before the title;
- Different documents from different sources have different typesetting templates for the title.

In order to extract the metadata, the different sections of the document first must be recognized, located, and segmented precisely, and then extracted. Fortunately, authors structure and order different sections of a scholarly document to deliver dedicated information to readers, and they have different explicit and implicit semantics. In this paper we call the visual formatting elements and contents explicit semantics, such as the title, author name, affiliations, running heads, footnotes, emails, address, abstract, keywords, section names, acknowledgments, and references. Implicit semantics means not only the formatting switch and change, but also redundant texts. Formatting switch and change means the change of line height, font type, font size, alignment, and other strictly physical characteristics. Some redundant texts, such as dates that appear in both the running header and footnotes, and author names in both an authorship section and footnotes, are also important sources for metadata extraction and mutual verifications.

The work described in this chapter uses both explicit and implicit formatting semantics to extract the metadata automatically. First, we construct an extensive template database for domain terms, affiliations, and complex combinations of typesetting elements. Second, both character stream from the original PDF file and formatting stream are extracted in parallel according to PDF specification. Third, different sections are segmented and trimmed according to templates and implicit semantics.