Erschienen in: Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries / J. Stephen Downie (Hrsg.). - New York : ACM, 2013. - S. 385-386. - ISBN 978-1-4503-2077-1

Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents

Mario Lipinski, Kevin Yao, Corinna Breitinger, Joeran Beel, Bela Gipp
University of California, Berkeley, CA, USA
mario@lipinski.tk, {kzyao, breitinger, jbeel, gipp}@berkeley.edu

ABSTRACT

This paper evaluates the performance of tools for the extraction of metadata from scientific articles. Accurate metadata extraction is an important task for automating the management of digital libraries. This comparative study is a guide for developers looking to integrate the most suitable and effective metadata extraction tool into their software. We shed light on the strengths and weaknesses of seven tools in common use. In our evaluation using papers from the arXiv collection, GROBID delivered the best results, followed by Mendeley Desktop. SciPlore Xtract, PDFMeat, and SVMHeaderParse also delivered good results depending on the metadata type to be extracted.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *performance evaluation (efficiency and effectiveness).*

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

Information Retrieval; Metadata Extraction; Evaluation; PDF

1. INTRODUCTION

Obtaining structured metadata from documents, including title, authors, and publication date, is important to support retrieval tasks, e.g., in digital libraries [2]. Various tools exist to automatically extract this information from PDF documents. However, the extraction is error prone given that no standards specify how metadata should be structured or formatted. Different style guides imposed by various publishers and venues, as well as the scope of metadata information provided by individual authors, increase the difficulty of automatic metadata extraction. In this paper we evaluate tools for metadata extraction from scientific articles.

A recent publication compared the metadata extraction capabilities of Mendeley and ParsCit, concluding that Mendeley's two-staged SVM solved the problem of metadata extraction for crowdsourced bibliographic metadata management [5]. Several approaches for extracting metadata have been proposed and independently examined. Since studies used different methods and data sources for their evaluation, a direct comparison is not possible. Currently, no comprehensive evaluation of tools for metadata extraction exists.

Fundamental methods used for metadata extraction are stylistic analysis, machine learning, and the use of knowledge bases. Metadata extraction tools using stylistic analysis extract titles using heuristics, e.g., font sizes and position information of examined elements. Machine learning techniques for metadata extraction use support vector machines (SVM), hidden Markov models (HMM), or conditional random fields (CRF). These approaches rely on previous training and natural language processing. Knowledge base approaches make use of databases, such as Google Scholar, or pronoun repositories, e.g., lists of common names, to act as a cross-reference to extracted entities. Software systems for metadata extraction combine these methods.

The JISC ConnectedWorks project published an overview of available software for processing PDF documents [6]. For our study, we focus on tools that are freely available and do not require user interaction so that the tools can be integrated into custom projects. Although Mendeley Desktop cannot be included in custom software, we include it in the evaluation as a widely used software with metadata extraction capabilities. We did not include Zotero, since its 25-document limit makes it unsuitable for bulk processing. Also the recently developed Docear's PDF Inspector [4] was not included since it was not yet available at the time of evaluation. Table 1 gives an overview of tools for extracting header information from PDF documents.

Table 1. Tools for Metadata Extraction from PDF Documents

Name of Tool Link	Approach used
Docear's PDF Inspector* http://docear.org	Style information analysis
GROBID https://github.com/kermitt2/grobid	CRF
Mendeley Desktop http://www.mendeley.com/	SVM, web-based look-up
ParsCit http://aye.comp.nus.edu.sg/parsCit/	CRF
PDFMeat http://code.google.com/p/pdfmeat/	Queries Google Scholar, pdftotext
PDFSSA4MET http://code.google.com/p/pdfssa4met/	Structure/Syntax analysis of XML
SciPlore Xtract http://sciplore.org/	Style information analysis of XML
SVMHeaderParse http://citeseerx.sourceforge.net/	SVM
Zotero* http://zotero.org	Queries Google Scholar

* not evaluated

2. METHODOLOGY

To meet the suitability requirement, extraction tools must fulfill three requirements. First, the tools had to provide an interface that allowed the integration into custom projects. They either had to provide a library for integration into other programs, or a stand-alone program that either accepted plain text or PDF as input. Second, they were not allowed to require user interaction to allow bulk processing. Third, the examined tools had to output machine-readable data, for example in XML format.

Table 2. Results (A_{100}: First evaluation setup with 100 articles, B_{100}: Second evaluation setup with 100 articles, B_{1153}: Second evaluation setup with 1,153 articles)

	Title			Authors			Authors' last names		Abstract			Year	
	A_{100}	B_{100}	B_{1153}	A_{100}	B_{100}	B_{1153}	B_{100}	B_{1153}	A_{100}	B_{100}	B_{1153}	B_{100}	B_{1153}
GROBID	N/A	0.92	0.92	N/A	0.83	0.83	0.90	0.91	N/A	0.75	0.74	0.64	0.69
Mendeley Desktop	N/A	0.84	0.82	N/A	0.72	0.70	0.78	0.77	N/A	N/A	N/A	0.23	0.26
ParsCit	0.59	0.52	0.54	0.47	0.29	0.31	0.36	0.37	0.49	0.31	0.26	0.06	0.07
PDFSSA4MET	0.13	0.21	0.18	0.05	0.02	0.01	0.20	0.18	N/A	N/A	N/A	N/A	N/A
PDFMeat	0.60	N/A	N/A	0.6	N/A	N/A	N/A	N/A	0.14	N/A	N/A	N/A	N/A
SciPlore Xtract	0.76	0.81	0.78	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SVMHeaderParse	0.50	0.57	0.61	0.64	0.70	0.73	0.74	0.76	0.37	0.64	0.64	0.21	0.20

Since the tools are written in different languages and have different output formats, we created a Java framework to provide a uniform interface for the variety of tools. The framework requires a PDF file as input and converts it, if required by the tool, to plain text using pdftotext. The output is wrapped into a unified Java data structure that stores the extracted fields.

To evaluate the tools, we compiled a test collection from arXiv.org, a scientific publication archive, which contains articles from various disciplines with various document formatting. Given the diverse document styles in the arXiv collection, it provides a good data source to test the performance of metadata extraction tools on articles in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics. In examining publications only from these fields, the results might not apply to other fields. By using arXiv's API, we obtained 1,153 random PDF articles including their metadata, dated from 2006 to 2010.

We performed three evaluations with two independent test setups. For the first setup, we randomly chose 100 articles from the test collection and the metadata extracted from the PDF documents was manually compared against the metadata from arXiv. We created a scoring scheme to assess the performance of the individual tools on the following metadata types: title, authors, and abstract. Each field was scored individually. For every field a score of 1 was given if the extracted metadata matched the reference data. A score of 0 was assigned if the field was extracted incorrectly. For title, authors, and abstracts a reduced score of 0.5 was allotted if the data was retrieved, but some characters, such as accents or ligatures produced problems. If only a fraction of the correct data was detected, a score of 0.25 was allotted.

For the second test setup, we performed two evaluations. First, we used the 100 documents from the first setup and second we used all 1,153 documents. In this setup, we developed a program to automatically determine the scores for extraction of title, author full names, author last names, abstract, and publication year. The program used the Levenshtein distance and normalized it using the length of the reference value. The resulting score approximately fits the percent match. The scores for all documents from the test collection were averaged.

3. RESULTS

Table 2 shows the results of the evaluations. In the following, we point out noteworthy results and give their accuracy scores. GROBID performs best; 0.92 for titles, 0.83 for authors, 0.91 for authors' last names, 0.74 for abstracts, and 0.69 for publication date. We believe that by working directly with PDF files without losing information in preprocessing, and by accurately engineered models, GROBID has an advantage over other methods. Mendeley Desktop ranks second for all extracted metadata types (0.82 for titles, 0.70 for author first names, and 0.77 for author last names). Mendeley leverages its extensive online database to enrich the extracted metadata. While the authors in [5] claim that Mendeley's two-stage SVM performs best, our results show that GROBID's CRF implementation can deliver better metadata extraction without consulting external resources.

SciPlore Xtract shows a good accuracy of 0.78 in extraction of the title. By taking into account the article's style information (font sizes and layout information), SciPlore Xtract can gain an advantage over other tools that ignore this information [3]. Nevertheless, the low score of 0.18 for PDFSSA4MET demonstrates that relying solely on font size is insufficient. Looking at the data revealed that PDFSSA4MET often extracted arXiv's document ID banner in the left margin as the title, so the tool may be at a disadvantage for the chosen test collection.

SVMHeaderParse delivered good results for extracting author names (0.73) and abstracts (0.64). The results for SVMHeaderParse and ParsCit for extracting titles and abstracts slightly differed between our two test setups. We believe that different versions of the tool for transforming PDFs into plain text can affect the performance of these tools. PDFMeat delivered relatively good results of 0.60 on title and author extraction. The relative good quality of the extracted data may result from incorporating results from Google Scholar [1].

The evaluation framework, including test collection, the ground truth and the test-software, is available from the authors by request.

4. ACKNOWLEDGEMENTS

We greatly acknowledge support by the VIGRE Program to work on this project at the Department of Statistics at the University of California, Berkeley.

5. REFERENCES

[1] Aumueller, D. 2009. Retrieving metadata for your local scholarly papers.

[2] Beel, J., Gipp, B., Langer, S., Genzmehr, M., Wilde, E., Nürnberger, A. and Pitman, J. 2011. Introducing Mr. DLib, a Machine-readable Digital Library. *JCDL'11*.

[3] Beel, J., Gipp, B., Shaker, A. and Friedrich, N. 2010. SciPlore Xtract: Extracting Titles from Scientific PDF Documents by Analyzing Style Information (Font Size). *ECDL'10*.

[4] Beel, J., Langer, S., Genzmehr, M. and Müller, C. 2013. Docears PDF Inspector: Title Extraction from PDF files. *JCDL'*13.

[5] Granitzer, M., Hristakeva, M., Knight, R. and Jack, K. 2012. A Comparison of Metadata Extraction Techniques for Crowdsourced Bibliographic Metadata Management. *SAC'*12.

[6] JISC ConnectedWorks Project 2010. *Research on existing PDF processors*. University of Cambridge.