2020

A Heuristic Baseline Method for Metadata Extraction from Scanned Electronic Theses and Dissertations

Muntabir H. Choudhury
Old Dominion University

Jian Wu
Old Dominion University

William A. Ingam

Edward A. Fox

Original Publication Citation

Choudhury, M. H., Wu, J., Ingram, W. A., & Fox, E. A. (2020). *A heuristic baseline method for metadata extraction from scanned electronic theses and dissertations.* Paper presented at the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020), Virtual, August 1-5, 2020.

A Heuristic Baseline Method for Metadata Extraction from Scanned Electronic Theses and Dissertations

Muntabir Hasan Choudhury
Jian Wu
Old Dominion University
Norfolk, VA
{mchou001,j1wu}@odu.edu

William A. Ingram
Edward A. Fox
Virginia Polytechnic Institute and State University
Blacksburg, VA
{waingram,fox}@vt.edu

ABSTRACT

Extracting metadata from scholarly papers is an important text mining problem. Widely used open-source tools such as GROBID are designed for born-digital scholarly papers but often fail for scanned documents, such as Electronic Theses and Dissertations (ETDs). Here we present a preliminary baseline work with a heuristic model to extract metadata from the cover pages of scanned ETDs. The process started with converting scanned pages into images and then text files by applying OCR tools. Then a series of carefully designed regular expressions for each field is applied, capturing patterns for seven metadata fields: titles, authors, years, degrees, academic programs, institutions, and advisors. The method is evaluated on a ground truth dataset comprised of rectified metadata provided by the Virginia Tech and MIT libraries. Our heuristic method achieves an accuracy of up to 97% on the fields of the ETD text files. Our method poses a strong baseline for machine learning based methods. To our best knowledge, this is the first work attempting to extract metadata from non-born-digital ETDs.

KEYWORDS

Digital Libraries, Optical Character Recognition (OCR), Text Mining, Metadata Extraction, Heuristic Method

ACM Reference Format:
Muntabir Hasan Choudhury, Jian Wu, William A. Ingram, and Edward A. Fox. 2020. A Heuristic Baseline Method for Metadata Extraction from Scanned Electronic Theses and Dissertations. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3383583.3398590

1 INTRODUCTION

Automatic metadata extraction from PDF documents is key to building a scalable document processing system for digital library search engines. Many AI-based methods have been proposed to extract metadata from scholarly papers, such as SVMHeaderParse [3], GROBID [5], Mendeley Desktop, and ParsCit [4]. However, most of these tools are built for relatively short, born-digital documents, such as articles in conference proceedings and journals published in recent years. They do poorly with scanned book-length documents. That applies for the millions of electronic theses and dissertations (ETDs), which may contain rich domain knowledge, but are underrepresented in academic search engines [2].

Since 1997, starting with Virginia Tech, more and more universities have started supporting ETD submissions. However, ETDs before then, and many ETDs since then, are still published in non-born-digital formats, usually generated by scanning physical copies. Many of these ETDs are accompanied with incomplete, little, or no metadata, posing great challenges for accessibility through search engine interfaces. Although many state-of-the-art open access tools exhibit satisfactory performance with certain types of documents, experiments indicate that they tend to produce unacceptable errors or fail for scanned ETDs. Extracting metadata from scanned ETDs is challenging due to poor image resolution, imperfections with OCR techniques, and typewritten text. Although commercially-based OCR tools such as OmniPage, ABBYY, and CuneiForm could be used, we chose Tesseract OCR. It is a widely adopted open source tool that takes any printed or scanned fonts, supports more than 100 languages, and returns output in text, hOCR, PDF, and other formats. Tesseract OCR also has been used in combination with Open-CV to extract text from smartphone screenshots [1].

Although many complicated learning-based models can be built, e.g., Conditional Random Field (CRF; [5]) or Support Vector Machine (SVM; [3]), there has not been dedicated effort and evaluation of heuristic methods with the ETD task. Heuristic methods are generally faster, suitable for capturing evident patterns, and do not require training data. In this paper, we attempt to build a heuristic baseline method to extract metadata from cover pages of ETDs. Heuristic methods are suitable here because a majority of ETDs follow similar templates. The heuristic method provides a strong baseline for development of learning-based methods.

2 METHOD

The ground truth was compiled by selecting 100 ETDs, with 50 each from the Virginia Tech and MIT digital libraries. Of these, 50 were published between 1945 and 1975, while the rest were between 1986 and 1990. They cover 41 majors includes STEM majors such as Biology and Chemistry, and non-STEM majors such as Education and Marketing. The combined corpus includes 5 bachelors, 70 doctoral, and 25 masters ETDs. We also downloaded metadata files in XML (MIT) or JSON (Virginia Tech) formats. We derive 6 datasets based on the raw dataset as intermediate files or for evaluation purposes.

(1) The first page of each ETD in PDF.
(2) TIFF images of (1). We found TIFF tends to produce significantly fewer misspellings than JPEG as the input of Tesseract.

JCDL '20, August 1–5, 2020, Virtual Event, China

ACM ISBN 978-1-4503-7585-6/20/06.
https://doi.org/10.1145/3383583.3398590

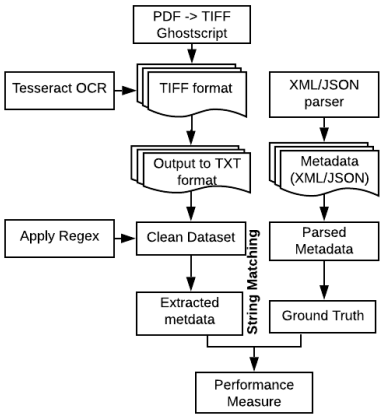

Figure 1: Metadata Extraction Flow Chart

(3) TXT-OCR. The text file converted from (2) by Tesseract.
(4) TXT-clean: We rectified the TXT-OCR dataset by correcting misspellings and missed text produced by OCR in (3).
(5) GT-meta: ground truth from metadata provided by libraries.
(6) GT-rev: there could be discrepancies between library provided metadata and the actual PDF documents. For example, the "department of chemistry" appearing on the cover page was called "analytical chemistry, polymers, and chemistry" in the metadata. Another example is that the advisor names appearing on the cover pages may not be in the metadata. To be consistent with the actual data, we use values printed on PDFs in lieu of library provided metadata *for these fields*.

The pipeline to extract metadata from ETD cover pages is illustrated in Figure 1. The names of these fields, the rules applied, and their accuracy values appear in Table 1. The regular expressions can be found in the GitHub repository[1].

3 EVALUATION AND RESULTS

The evaluation is conducted by comparing the metadata of each field extracted from TXT-clean, against the corresponding GT-rev data. See Section 2 and the $\mathbf{A}_{\text{cln}}$ column in Table 1. The accuracies are computed by dividing the number of correctly extracted samples by the total number of samples for a particular field. For title, degree, program, and institution, we compare the lowercased strings. In many cases, the names of authors and advisors on the cover page may be written in different ways in the ground truth. For example, "Inrique I. Kilayko" is spelled as "Inrique Kilayko". Therefore, for names, instead of performing the whole string comparison, we decompose the full names into prefix, first name, middle name, last name, and suffix, and perform lowercased string comparisons of each field. In the ground truth, most degrees are expressed as abbreviations, such as "Ph.D." or "M.Arch." but the cover page usually prints the full names, such as "Doctor of Philosophy" or "Master of Architecture". We map the acronyms to the full names by incorporating an external dictionary from Wikipedia[2].

Since our method involves analysis of text strings, it is essential that the strings studied be correct. When starting with image files, very high quality OCR methods should be employed. We present upper bound results for such OCR methods by testing with manually corrected/rectified OCR data. The problems resulting from using noisy OCR results can be seen by comparing the accuracies in the last two columns of Table 1.

Field	Rules	$\mathbf{A}_{\text{cln}}$%	$\mathbf{A}_{\text{OCR}}$%
Title	The first 4-5 lines preceded with 'by'	81%	56%
Author	The string followed after 'by' but started in a new line	78%	33%
Degree	The string after 'degree of' but before a space or starting in a new line	81%	55%
Program	The string preceded with 'department of' or 'in' but followed after space or started in a new line	97%	35%
Institution	The string after 'at the', 'faculty of the', 'at', or 'faculty of' and followed after space or started in a new line	94%	66%
Year	The 4 digits before a 'month'	65%	61%
Advisor	The string after 'certified of' or 'approved' and in a new line	36%	1%

Table 1: Rules for extracting each metadata field and accuracy. $\mathbf{A}_{\text{cln}}$% and $\mathbf{A}_{\text{OCR}}$% are accuracies based on TXT-clean and TXT-OCR datasets, respectively.

4 CONCLUSION

We applied a heuristic model to extract metadata from 7 fields from ETDs and achieved 36%-97% accuracy measures. This work provides a relatively strong baseline for developing learning based methods. The results indicate the necessity to clean text directly output by Tesseract. Future evaluations will incorporate more and diverse sources of ETDs, e.g., more universities and academic programs. We will investigate efficient approaches to automatically correct text directly generated by OCR tools.

ACKNOWLEDGMENTS

Support was made in part by the Institute of Museum and Library Services for grant LG-37-19-0078-198.

REFERENCES

[1] Chiatti, A., Cho, M. J., Gagneja, A., Yang, X., Brinberg, M., Roehrick, K., Choudhury, S. R., Ram, N., Reeves, B., and Giles, C. L. Text extraction and retrieval from smartphone screenshots: Building a repository for life in media. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (2018).
[2] Fox, E. A., McMillan, G., and Srinivasan, V. *Electronic Theses and Dissertations: Progress, Issues, and Prospects.* Center for Digital Discourse and Culture, 10th Anniversary Book Blacksburg, VA, 2009, pp. 126–148.
[3] Hui Han, Giles, C. L., Manavoglu, E., Hongyuan Zha, Zhenyue Zhang, and Fox, E. A. Automatic document metadata extraction using support vector machines. In *Proceedings of JCDL* (2003).
[4] Lipinski, M., Yao, K., Breitinger, C., Beel, J., and Gipp, B. Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In *Proceedings of the 13th JCDL Conference* (2013).
[5] Lopez, P. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries* (2009), ECDL'09.

[1] https://github.com/lamps-lab/ETDMiner
[2] https://en.wiktionary.org/wiki/Appendix:Academic_degrees