Named Entity Extraction and Classification in Digital Publications

Chuan-Yu Wu*, Bom Yi Lee*, Jing Sun†, Yin Yin Latt‡, Kim Shepherd‡ and Jared Watts‡

*Department of Electrical and Computer Engineering
†Department of Computer Science
‡Library Digital Services
The University of Auckland, New Zealand
Emails: *{cwu323, blee660}@aucklanduni.ac.nz, †j.sun@cs.auckland.ac.nz, ‡{y.latt, k.shepherd, j.watts}@auckland.ac.nz

***Abstract*—This paper describes the design and implementation of a PDF extraction tool, which provides the functionalities of meta-data creation, bibliography extraction, HTML conversion and key phrase classification. Evaluation results showed high accuracy rates and good performance measurements.**
***Index Terms*—Text Extraction, Automatic Classification**

I. Introduction

The University of Auckland is New Zealand's largest research organisation, with over 13,000 staff and postgraduate students involved in fundamental and applied research. Its General Library holds thousands of publications and dissertations within its repository. The standard file format for publications is in Portable Document Format (PDF), which is not content-accessible [1]. This means that, although the library repository holds a large amount of information, this available information is not being utilised to its full potential.

II. Design and Implementation

To enable automated extraction of information from the digital scholarly publications in the library repository, it involves analysing and extending existing tools that fit our purpose. Our project has explored four areas of information extraction and classification from publications in PDF format. We have developed separate tools for each feature, which will be used by the library. The general structure of our overall solution is shown in Figure 1.

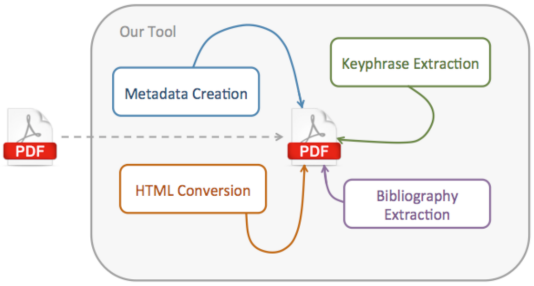

Fig. 1. Overall Structure of the PDF extraction tool

DOI reference number: 10.18293/SEKE2017-132

Four main components included in our project are metadata creation, key phrase extraction, HTML conversion and bibliography extraction. These features were implemented by utilising and extending the existing tools, such as the Apache PDFBox, KEA, PDF2Dom and Freecite. Our keyword identification uses the ACM Computing Classification System SKOS file. The front-end GUI is a means of displaying the resulting output of each tool. It was implemented using JavaFX with SceneBuilder following a model-view-controller (MVC) architecture.

III. Evaluation and Conclusion

The developed tool was evaluated against quantitative and qualitative analyses to measure its efficiency and viability for use within the library repository. We have carried out our evaluation on a set of 30 open-source publications that were provided by the library team. The results showed high accuracy rates, and fast processing time in data extraction and key phrase classification. A comparison table was drawn up to compare the features provided by several existing tools in relation to our solution, as shown in Figure 2.

	Open source	PDF conversion	Keyword Extraction	HTML conversion	Bibliography Extraction	Standalone	Cross-platform
pdftotext	✓	✓				✓	
PDFlib (TET)		✓				✓	✓
PDFMiner	✓	✓				✓	✓
KEA	✓		✓			✓	✓
PDFBox	✓	✓		✓		✓	✓
FreeCite	✓				✓	✓	✓
Our tool	✓	✓	✓	✓	✓	✓	✓

Fig. 2. Feature comparison between our tool and exisiting tools

Our project has taken a breadth over depth approach to explore the different solutions. Further extensions can be made by incorporating: (1) Machine learning process during key phrase extraction; (2) Deriving sets of characteristics for disciplines other than Computer Science to allow extraction of entities from any types of scholarly publications.

References

[1] D. D. A. Bui, G. Del Fiol, and S. Jonnalagadda, "Pdf text classification to leverage information extraction from publication reports," *J. of Biomedical Informatics*, vol. 61, no. C, pp. 141–148, Jun. 2016.