Information Technology
Thesis
Prof. Dr. Josef Fink

Exposé Master Thesis

Data Extraction from PDF Reports

Abbad Zafar

1360066

Supervisors:

Prof. Dr. Josef Fink

Fabian Winkel, MSc.

at

Frankfurt University of Applied Sciences

Faculty 2: Computer Science and Engineering

Information Technology

Table of Contents

1. Motivation 3
2. Research Scope 3
3. Approach 5
   1) Data Collection 5
   2) Data Extraction 5
4. Evaluation 6
5. Literature 6
6. References 7

1. Motivation

The widespread use of portable document format (PDF) documents has sparked interest in research on automated data extraction from these files [1]. Text extraction from a PDF document is an important yet unexpectedly tough task. The reason for this is because PDF is a layout-based format in which the fonts and placements of individual characters are specified rather than the semantic units of text (e.g., words or paragraphs) and their role in the document (e.g., body text or caption). There are several extraction tools available, but determining their quality and length of capability is difficult [2].

With each successful extraction, hidden patterns, significant statistics, and critical trends are revealed. Our drive to perfecting the art of data extraction from PDFs enables us to overcome complicated obstacles, allowing us to innovate and be more efficient. The retrieved data are relevant in a wide range of sectors, including academic research and student and teacher support [1].

2. Research Scope

For this thesis, the research scope aims to develop an innovative and versatile data extraction technique capable of extracting valuable data from various types of reports, such as product reports, business reports, and more. The primary objective is to design a method that can effectively identify and extract relevant information from these reports, thereby facilitating data analysis and decision-making processes.

To accomplish this, the research will involve exploring existing data extraction techniques and methodologies commonly used in the field. A comprehensive literature review will be conducted to gain insights into the different approaches employed for data extraction in various domains. This will help in identifying the strengths and limitations of current methods and formulating a novel technique that can overcome existing challenges.

The proposed data extraction technique will be designed to be adaptable and scalable, allowing it to be applied to a wide range of report formats and structures. It will incorporate advanced algorithms and machine learning models to automatically detect and extract relevant data points, such as numerical figures, textual information, and categorical data.

Furthermore, the research will emphasize the development of robust data validation and error-checking mechanisms to ensure the accuracy and reliability of the extracted data. This will involve implementing quality control measures and statistical analyses to verify the integrity of the extracted information.

The research will also involve designing and implementing a software prototype or tool to demonstrate the practical application of the developed technique. This tool will serve as a proof of concept and provide a user-friendly interface for users to input their reports and extract the desired data effortlessly.

To evaluate the effectiveness of the proposed technique, comparative studies will be conducted, comparing the performance of the developed approach with existing methods. Real-world report datasets from different domains will be utilized to validate the technique's capability to handle diverse types of reports and extract useful information accurately.

The ultimate goal of this thesis research is to contribute to the field of data extraction by introducing a flexible and efficient technique that can be utilized across various industries and domains. By enabling automated data extraction from reports, organizations will be able to streamline their data analysis processes, make informed decisions, and enhance overall operational efficiency.

Furthermore, few details we specifically looking for to extract data in reports are:

1. How can we extract details about author, title, last modification date,

creation date, subject?

2. Finding technique to extract data from images inside the pdf reports?
3. Extraction method for tables and forms?
4. Extraction technique for headers and paragraphs in pdf reports?

As a result, we will be able to generate a useful data extraction method for the pdf reports.

3. Approach

1) Data Collection

To develop an effective data extraction technique, obtaining sample reports from diverse companies or products is crucial. These reports will serve as the basis for understanding the variations in report formats, structures, and data types. By analyzing and extracting data from these samples, we can design a technique that is capable of handling the complexities and nuances present in real-world reports, ensuring its applicability and accuracy across different contexts.

2) Data Extraction

• Once we are able to collect reports, we will start working on different data extraction techniques. We will have to apply different techniques to extract different parts of data in the report pdf files.

• For extracting author, title etc. we would need to extract metadata and get the title, author detail from there.

• For extract table we would need to specify the pdf page we need to extract data from so far, I have used “tabula” library for extraction of tables from pdf.

- For extraction of images data, first we would need to download images and then we can apply OCR techniques to extract text from pdf.
- Next task would be to extraction of header, footers, and paragraphs.

4. Evaluation

After implementing the data extraction techniques, it is crucial to evaluate the accuracy and correctness of the extracted data. This evaluation ensures that the chosen extraction method is reliable and effectively captures the desired information. By comparing the extracted data with ground truth or manually validated data, we can assess the precision, recall, and overall performance of the extraction technique.

Once the correctness of the extracted data is confirmed, it can be utilized for further analysis or predictions. Machine Learning algorithms can be applied to extracted data to uncover patterns, relationships, and insights. These algorithms can perform tasks such as classification, regression, clustering, or predictive modeling, depending on the specific objectives of the research or application.

5. Literature

Data extraction from PDF reports has been extensively studied in the literature. Notable research papers include "A Study on Information Extraction from PDF Files " by Yuan and Fang [3] which proposes techniques for extracting information from PDF reports. Another significant paper is " A Benchmark and Evaluation for Text Extraction from PDF" by Bast, Hannah & Korzen, [5] providing a comprehensive survey of different techniques. Additionally, "PDFX: Fully-automated PDF-to-XML Conversion of Scientific Literature" by Alexandru

Constantin, Steve Pettifer, and Andrei Voronkov [4] presents an automated approach for converting PDFs to XML.

6. References

[1] Wiechork, Karina & Charão, Andrea. (2021). Automated Data Extraction from PDF Documents: Application to Large Sets of Educational Tests. 10.5220/0010524503590366.

[2] Bast, H. and Korzen, C. (2017). A Benchmark andEvaluation for Text Extraction from PDF. In 2017ACM/IEEE Joint Conference on Digital Libraries(JCDL), pages 1–10.

[3] Yuan, Fang et al. "A Study on Information Extraction from PDF Files." International Conference on Machine Learning and Computing (2005). 10.1109/ACCESS.2021.3107975.

[4] Alexandru Constantin, Steve Pettifer, and Andrei Voronkov. 2013. PDFX: fully-automated PDF-to-XML conversion of scientific literature. In Proceedings of the 2013 ACM symposium on Document engineering (DocEng '13). Association for Computing Machinery, New York, NY, USA, 177–180. https://doi.org/10.1145/2494266.2494271

[5] Bast, Hannah & Korzen, Claudius. (2017). A Benchmark and Evaluation for Text Extraction from PDF. 1-10. 10.1109/JCDL.2017.7991564.