**NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS**

**MOSCOW INSTITUTE OF ELECTRONICS AND MATHEMATICS A.N. TIKHONOV**

**DEPARTMENT OF COMPUTER ENGINEERING**

**PROJECT PROPOSAL**

**AN AUTOMATIC SYSTEM FOR TAGGING OF SCIENTIFIC ARTICLES**

Tatiana Yudina, BIV-164

Advisors:

Eduard Klyshinskiy, Associate Professor

***Abstract*—With the increase in the amount of information, the need for automatic text analysis becomes more and more obvious. For scientific publications, the most common PDF format, which allows you to correctly display both text and various non-text elements that are a feature of this style of speech. One of the most important problems encountered when trying to convert PDF format to text is the problem of correctly displaying non-text elements. This article presents a study that aims to develop a new approach to solving this problem and to implement this system on a client-server application.**

***Keywords—scientiﬁc information extraction; client-server application; document structure; object detection***

Introduction

**Background.** Today, the value of automatic text analysis is especially great due to the increase in the amount of information and the impossibility of its manual processing. The problem of automatic analysis is relevant for different styles of speech, but the situation is especially acute in the case of scientific publications due to the existence of various fields of science with its own characteristics, which can complicate the automatic analysis of the text, as well as due to the presence of a large number of non-text elements. In addition, an important factor is that most scientific articles are publicly available in PDF format, which means that a tool is needed for high-quality conversion from this format to text with the correct transfer of non-text elements.

**Problem Statement.** The main goal of future research is to study existing approaches, algorithms and solutions in the field of automatic processing and markup of documents that are used today in production, and to study the structure of scientific articles for the subsequent development of an automatic system that extracts and marks semantic fragments of scientific articles using open software. Given this, the study will present a new general approach to converting scientific articles from a PDF format to a docx format while preserving the original hierarchy. To achieve this goal, it is reasonable to follow these steps: to study the text block extraction algorithms, used today, and the algorithms for extracting various non-text elements, such as images, tables, flowcharts, etc., in order to develop our solution based on existing ones, and then test it on the basis of the data and check whether this solution meets our requirements for speed and accuracy. Another important step is the implementation of the client part of the system for the distribution of computing resources created during the research of algorithms.

**Professional Significance.** The solution described in the proposed study provides the opportunity to work with scientific articles with the correct extraction of structural information from PDF: saving the hierarchy of text blocks and all non-text elements in the document and converting non-text elements to images for correct visual display.

**Research stages.** It is assumed that the project will begin by analyzing the existing approaches to converting PDF documents into text format and analyzing the structural parts of the scientific style of speech. The study will provide a new approach to highlighting structural blocks, and then to extracting textual information from those blocks where it is required. It is supposed that the study will demonstrate the practical implementation of this algorithm in the form of a client-server application that will allow you to distribute computing power. The final step will be to test the software to make sure that it works correctly with the various inputs.

Literature review

During the past decade, separation of semantic and logical structure of documents in the PDF has been a relevant problem, taken into account the ubiquitous use of the “Portable Document Format” (PDF) for the dissemination of scientific publications. To solve this problem, two approaches are most often used: a rule-based approach and machine learning algorithms. In this section, I am going to review several articles that contain these approaches to extracting information from PDF documents. The basic concepts for my research are derived from these publications.

In their article [1], Ramakrishnan et al. present the “Layout-Aware PDF Text Extraction” (LA-PDFText) system to facilitate extraction of text from PDF documents of research articles and the subsequent use of the extracted text in text analysis applications. This system detects adjacent text blocks using the JPedal library, written in the Java language and distributed under the GPL, and classifies them into logical units based on the rules that characterize specific sections. At the last stage of text extraction, the system connects the grouped text blocks in the correct order. Precision, recall, and F1-score metrics are used to assess the accuracy of text block identification and subsequent classification. The authors themselves emphasize that the LA-PDFText system focuses only on the text content of documents and is viewed as the basis for further experiments with more advanced extraction methods for processing various non-text content such as tables, images, graphics, forms and mathematical equation.

Ronzano et al. in article [2] propose a platform for automatically extracting of the structural and semantic aspects of scientific publications, presenting them as the RDF datasets. This platform uses a framework developed in the context of the European project Dr. Inventor. The first version of this project is presented by Ronzano et al. in their article [3]. Dr. Inventor Framework can analyze scientific publications both in plain text and in the PDF. Tools implemented in this framework include extracting structured text content, identifying structural elements of headings and bibliographic records of documents, and creating graph-based representations of text excerpts. The framework is distributed as the Java library. The main disadvantage of this framework is that it also allocates only textual information.

The browser system described by Abekawa et al. in article [4] provides methods for extracting sentences from PDF files along with their logical structures, as well as for matching arbitrary text intervals with the corresponding image areas. To highlight textual information, they wrote their own patch for pdftotext, which is included in the Poppler library. The main advantage of this system is that it selects pictures and tables using the PDFFigures tool. However, at the same time, the main disadvantage of this system is that it loses the location of non-textual information in the original structure of the PDF document and suggests interacting with it separately from the main text.

The framework described by Rahman et al. in article [5] uses layout information and text content extracted from PDF documents. PDFLib is used to extract metadata and text content. After that, modern machine learning models are developed, including a deep learning architecture for classification and semantic annotation. There are experiments with Dirichlet's hidden distribution models, TextRank and Tensor Flow in Textsum. This work is primarily focused on understanding the text of PDF. The purpose of this work is to section large and complex PDF documents automatically and annotate each section with semantic, human-understandable labels. The main advantage of this work is the use of machine learning instead of a rule-based approach. However, at the same time, they also lose non-textual information.

In general, the analyzed articles make a great contribution to solving the problem of highlighting the semantic and logical structure of documents presented in PDF format. Most of the articles discussed above use rule-based approaches, which is outdated, given the quality shown by methods based on machine learning algorithms. In work [4], unlike the others, not only textual information is extracted, but also the original structure of the document is lost, which, of course, is a significant drawback. Thus, our work will be to create a system that combines the best ideas of these works: the use of modern machine learning methods to highlight textual and modal information, while preserving the original structure of the document presented in PDF format.

Methods

This part of the proposal explains the methods that are going to be used in carrying out the study. To solve the problem of understanding the structure of a document presented in PDF format, we will use an approach based on computer vision - the search for objects in an image. To do this, we will generate synthetic data and teach them the algorithm to recognize four types of objects: text blocks, pictures, which also include flowcharts and diagrams, formulas and tables. This algorithm will first select a region and then classify it. Precision, recall, and F1-score metrics will be used to evaluate classification accuracy. After finding the objects in the image, a rule-based approach will be used to preserve the original hierarchy of the document, after which the remaining processing steps will be performed: extracting text from the image and bringing all the extracted information to the required docx format.

The client part of the system, which is responsible for the exchange of information between the user and the server, is necessary for the distribution of computing data resources of the algorithm. To implement the web interface of the client part of the system, Python version 3.7 and library Flask will be used. Using the client part, the source PDF document is downloaded, after which a REST API request to the server will be generated and sent, on which all the algorithms for converting the document from PDF to docx format will be executed. After receiving a response from the server, through the client part, the user will be able to download the created docx file and get an additional description of the document structure.

Results anticipated

This part of the project proposal is organized to report on the results which are likely to be obtained with the method mentioned. It is expected that the following results will be obtained within the framework of the project: analysis of existing approaches to converting PDF documents to any text format will show significant shortcomings, mainly due to the fact that in most cases approaches based on rules are applied to highlight areas and their classification, without the use of machine learning algorithms. The algorithm described above will presumably show the best result due to an increase in the classification accuracy, therefore, it will be more efficient and more suitable for use, although the use of neural networks will probably make it difficult to calculate the result on a PC. In order for the algorithm based on neural networks to work faster, a client-server approach is used that allows computing on a remote server with greater computing power than a standard PC. It is expected that within the framework of this project, the average time spent on processing one page, as well as a qualitative assessment of conversion will be calculated. This estimate will be the percentage of correctly allocated blocks to all blocks of each type.

Conclusions

In the proposed study, an attempt will be made to implement a new algorithm for extracting and classifying semantic structural blocks from PDF documents of a scientific style of speech for subsequent work with these blocks to convert the original document to the docx format. It is assumed that the created automatic system will be able to accurately identify the main types of information blocks, and will also be fast, reliable and easy to use. Preliminary experimental results will probably show that the created software implementation preserves the original hierarchy of the document and at the same time correctly displays non-text elements.

References

1. Ramakrishnan, C., Patnia, A., Hovy, E., Burns, G. A. Layout-aware text extraction from full-text PDF of scientific articles. *Source code for biology and medicine*, 2012, vol. 7, no. 7. Available at: https://link.springer.com/article/10.1186/1751-0473-7-7 (Accessed 16 February 2020).
2. Ronzano, F., Saggion, H. Knowledge extraction and modeling from scientific publications. *International workshop on semantic, analytics, visualization*, 2016, pp. 11-25.
3. Ronzano, F., Saggion, H. Dr. inventor framework: Extracting structured information from scientific publications. *International Conference on Discovery Science,* 2015, pp. 209-220.
4. Abekawa, T., Aizawa, A. SideNoter: scholarly paper browsing system based on PDF restructuring and text annotation. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations,* 2016, pp. 136-140.
5. Rahman, M. M., Finin, T. Unfolding the Structure of a Document using Deep Learning. *arXiv preprint arXiv:1910.03678*, 2019.