Insights to the State-of-the-Art PDF Extraction Techniques

Hashmi, Ahmer Maqsood; Qayyum, Faiza; and Afzal, Muhammad Tanvir

Abstract: ***Digitized documents have become the omnipresent medium of information. A plethora of scholarly documents on the web is excessively being increased. Various digital libraries such as Google scholar, Citeseer, MAS etc. store this plethora in different formats. Most of the scientific literature is stored in Portal Document Format (PDF). PDF documents hold a complex structure due to which their comprehension and extraction of useful information from them is a challenging task. In this regard, research community has been proposing different rule based and machine learning based techniques in the past several years. We believe that accurate and efficient information extraction form the PDF files is an important issue as major portion of scholarly literature is stored in PDF. This study presents a rigorous analysis of the contemporary state-of-the-art in PDF data extraction. The contemporary approaches from the window of past few years are recapitulated with the primary objective to assist the scientific community by providing them knowledge about current trend in PDF extraction techniques. The study also presents critical analysis and suggests future dimensions of some of the approaches.***

Index Terms: ***Key Information extraction, Research papers, PDF parser, Regular expression, XML and plain-text formats***

1. INTRODUCTION

For past several years, the paradigm of modern world has been shifted to the digital world in almost all spheres of life. Due to rapid growth of inventions in Science, huge plethora of research documents is produced by the scholarly community. According to [1] quantity of research papers is getting doubled after every five years. Digital libraries such as Google Scholar, Citeseer, Microsoft Academic search (MAS) store this bulk of data in different formats. These libraries contain around 114 million scholarly documents in the form of Portable Document Format (PDF) [2]. This excessive increase has posed a challenge for the scientific community in terms of finding best solutions to extract relevant information from this Big data [2]. Research community has proposed several techniques to extract the information from different formats of digital corpora and remained successful to some extent. However, the issue of extracting desired information from PDF has not been addressed with adequate accuracy. This is due to the lack of proper apprehension of PDF structure [3]. Dating back from the late 90's to date, researchers have proposed various techniques [6–11, 13, 14, 16–22] to extract information available inside PDF documents. Major portion of scientific documents available on the web is in the form of PDF, therefore, researchers must focus on proposing techniques to extract information from those documents in a coherent manner. This study presents a comprehensive literature review of the research conducted in this area and provides a critical overview of all the proposed approaches. The primary objective of this study is to assist the scientific community by providing rigorous analysis of state-of-the-art PDF extraction techniques so that they can evade the hassle of finding the relevant literature and then, have longer time to reach out to the important issue.

Some portion of PDF extraction techniques have focused on extracting data by considering logical sections of research papers. Logical extraction is the identification of the document into the logical sections, such as header, footer, abstract etc. Research community has proposed several techniques to evaluate the logical structure of the PDF. The approach proposed by Ramakrishnan et al. [4] performs layout analysis of the document and converts the PDF document into simple text file by identifying the text blocks. These text blocks are then categorized using a set of rules prepared by detailed analysis of documents. The system provides layout extraction with high precision; however, their system does not include extraction from graphs, tables, citations and figures. Dr. Inventor [5] is another framework to extract the logical structure of the document. It

Manuscript received August 1, 2016.
Ahmer Maqsood Hashmi, Faiza Qayyum, and Muhammad Tanvir Afzal are at the Department of Computer Science, Capital University of Science and Technology, Islamabad. Contact mail: mafzal@cust.edu.pk

converts PDF document into XML format. This conversion from PDF to XML provides feasibility in section identification. PDFX [6] converts PDF document into XML format by generating output in the form of XML. The output XML is generated by performing layout analysis of PDF document, followed by the identification of each extracted layout. The converted XML format contains geometrical features of the sections as well as the sections labelled according to the type of information that the tag contains. D´ejean and Meunier [7] proposed an approach to convert the PDF document into XML format. Their approach converts streams available in the form of PDF to structured XML. Identified heuristic are applied on the streams to logically evaluate the extracted structured XML document. Converting digital documents in the form of XML can help in extracting the information in a coherent manner. Riaz et al. [3] proposed an approach that works by converting PDF document into XML format using PDFX. The identified XML tags are then passed from multiple heuristics to extract the metadata of research documents. This approach was considered best metadata extraction approach among all the approaches presented in Extended Semantic Web Conference (ESWC) [8], securing an f-score of 0.77.

Researchers have proposed various machine learning approaches [16–19, 21] that extract information from PDF documents. GROBID [9] extracts metadata using machine learning approach and then generates a web request to extract the bibliographic information from PDF documents. Another approach CERMINE [10] extracts information from PDF documents in multiple steps. It performs the layout analysis and then classifies the type of metadata. SectLabel [11] performs metadata extraction and content classification using CRF [12], a machine learning approach that performs a probabilistic structure prediction using large set of input features. Klampfl and Kern proposed an unsupervised approach [13] to extract metadata from PDF documents. The approach performs logical structure analysis of a PDF document and extracts information from that logical structure analysis using unsupervised learning.

Hybrid approaches have always been employed by various researchers to extract metadata from PDF files. PDFMEF [14], combines open-source frameworks such as GROBID [9], CERMINE [10], ParsCit [15] to extract the information from a PDF document. Tuarob et al. proposed an approach [16] that identifies the section boundaries using machine learning approach and then labels these identified sections as standard identification (Abstract, Introduction, Background, and Experiment, etc.) using heuristics prepared by analyzing these PDF documents. Sateli and Witte [17] have combined the LOD-based Named Entity Recognition (NER) tool with the rule-based approach to extract information from PDF documents. This approach was considered second best information extraction approach in ESWC [8], by securing an f-score of 0.61.

So far, we have briefly recapitulated the trend of contemporary state-of-the-art in PDF extraction. The following section presents rigorous analysis of existing PDF extraction techniques.

2. Analysis of State-of-Art Techniques

The existing PDF extraction techniques can broadly be categorized into three types: (1) Rule/Heuristic based (2) Machine Learning based and (3) Hybrid approaches.

2.1 Rule Based Approaches

Rule based approaches are based on common patterns identified in the documents after a critical analysis of the dataset. Based on these identified patterns, multiple rules or heuristics are created to extract the information available inside the PDF.

Jahongir and Jumabek [18] extracted metadata from PDF documents. This approach works in three steps (1) Classification of the PDF files, (2) Metadata extraction, and (3) Storing PDF files in the form of XML or JSON. First step performs the classification of the document as either scientific or non-scientific. If the document contains keywords such as ‘Abstract’, ‘Introduction’, ‘Reference’ and ‘Conclusion’ etc. then these documents are termed as scientific documents and are processed further to extract metadata. The second step performs the metadata extraction and outputs the extracted metadata. The textual and font features are extracted using the Apache PDFBox [19]. The identified rules are applied on the extracted text to extract ‘Abstract’, ’Keywords’, ’Body text’, ‘Conclusion’ and ‘References’. The rules proposed by this approach are illustrated in Table 2.1. The final step of their methodology stores the extracted information in the form of XML or JSON. The approach achieved and accuracy of 97.71% for document classification and 96.31% for metadata extraction. Although the approach achieved a very high value of accuracy on the evaluation dataset, the formed rules are not generalized, thus, the results could not be similar for all the data sets.

Riaz et al. [3] proposed a rule-based

approach for extracting information from PDF documents. Their approach works by converting a PDF document into XML and plain text format. Each metadata extraction is performed by the respective metadata unit, and each unit consists of mainly three parts (1) Metadata identifier, (2) Metadata refiner, and (3) Metadata splitter. Metadata identifier identifies the metadata from the XML, followed by the metadata refiner that cleanses the identified text. Metadata splitter splits each extracted metadata and outputs actual extraction information. Firstly, it converts PDF document into XML format using PDFX [6]. PDFX converts the PDF document into the tagged XML. Further processing is performed on that tagged XML to extract the actual metadata and to output metadata information in the form of RDF triples.

The metadata information of author, affiliation and country is extracted by 'Author parts extractor'. This unit finds the title of the PDF document from the converted XML file and extracts the text between the title and 'Abstract' key phrase for further processing. Authors and affiliations are extracted using heuristics. Once these are identified, the country is extracted from the affiliation part using a predefined country list that contains names of all countries in the world. After the extraction of the author and affiliation, the author is affiliated with the respective institution, generating an output that contains the author, the affiliation, and the country. Information regarding figures, tables, supplementary material links, and funding agency is also extracted using the XML format. Regular expressions are developed for extraction of figure and table information using XML tags. Tables are extracted by using "> (Table|TABLE) [A-Za-z0-9\s\.:,\(\)*\%/-]{4,} </caption>" regular expression. The extracted information is then cleaned by removing the extra characters from the extracted text. The figures are extracted using multiple regular expressions developed from the XML format. If one regex does not return any output, then another regex is applied for the extraction of figures. Once figures are extracted, each figure is separated and extra characters that are not part of the figure are removed by the refiner. Supplementary material links are identified by following regular expression "http [A-Za-z0-9\\.#\%,:\/_\-\]*", which afterwards cleans its output. Same as figures, funding agency is also extracted using multiple regular expressions, which are formed by critically analyzing the text of the PDF document in the text viewer tool. Each unit output is passed through the content cleaner phase that removes the extra characters from extracted text and forwards it to the splitter, which outputs the metadata id along with the metadata text. The section identification is performed using both the XML and plain text formats. PDFX tool outputs the section headings as '<h1>' tag. After critically analyzing the sections headings in both plain text and XML formats, multiple heuristics are applied for the extraction of section headings. After the extraction of headings, these headings are separated with their number and passed for further processing. After the completion of extraction phase, they store the metadata in the form of triples using SPARQL. This component consists of two parts: the first part collects all the collected information extracted by using all the extraction units and the second part stores this extracted information in the form of RDF triples. This approach was developed using the training dataset of ESWC [8] consisting of 45 research articles, having different formatting styles and features. This approach was considered as the best performing approach in the ESWC, securing an f-score of 0.77, followed by the approach proposed by Sateli and Witte [17], that will be discussed in later sections.

Riaz et al. approach used PDFX (an open-source tool) [6] for the conversion of PDF document into XML format. PDFX performs the reconstruction of logical structure of the PDF document and identifies each block in terms of title, section, table, references etc. This tool works in two phases: first stage constructs a geometrical model using the content of the article and the second phase identifies the logical structure using the geometrical model generated in step 1. Multiple font features and geometrical features such as orientation, textual context, boundary, and font information are used by this tool for identification of different logical units. The most basic logical separation is performed using font size, whenever font size changes, a new logical unit starts. Furthermore, font frequency graphs are used to separate common text (section text) from the rare text (title, heading text, tables/ figures text etc.). The tool converts PDF documents into small text blocks and merges these small blocks afterwards, using the font and geometrical features. After merging the textual blocks, reading order-based rules are applied to label each logical unit as title, author, email, section, figure, reference, body etc. This approach was tested on Elsevier and PMC dataset, securing an f-score of 0.77 for the extraction of metadata and identification of logical units.

Another approach proposed by Klink and Kieninger [20] also incorporates the textual and physical features of PDF for the extraction of information from PDF documents. The

approach constructs the logical structure of a PDF document and identifies the header, footer, body text, table, and listings. Header section is identified by reading the text from the top of the page until a very large gap than usual is found. In the same manner, footer is identified. Lists (bulleted, numbered or dashed) are identified using the heuristic that the first character will be number, enumeration, dash, bullet or dot. Body text is also identified using the geometrical features such as start of the block, spacing between blocks and change of font features. To identify the tables, authors have used the algorithm proposed in T-Recs [21]. This approach was evaluated on the corpora of University of Washington received by the German Research Center for Artificial Intelligence. This approach achieved precision of 0.98 for 90% documents. However, the approach has proposed only one rule to extract information from PDF. It could further be enhanced by using a set of rules that can make extraction more diverse.

2.2 Machine Learning Approaches

Machine learning based approaches facilitate in terms of making a supervised system through learning different formats and features. Using this learning, the system can extract information from the PDF document in an automated manner. This section presents in-depth analysis of the machine learning approaches.

GROBID [9] is an open source ML library that performs extraction, parsing and reconstruction of a PDF document into a structured text. The system works by extracting the title, author, abstract etc. using the Conditional Random Field algorithm. After identification of the information, the system generates a web request that generates full metadata of the publisher. The approach has achieved an accuracy of 83.2%; however, the results may possibly be accurate only if the title and first author information is identified correctly by the system. This system is now available as an open-source tool and is in process of constant development.

CERMINE [10] is also an open-source ML tool that extracts metadata and content from a PDF document and generates an output in the form of XML or plain text. It performs the layout analysis in which character extraction, page segmentation, and reading order is resolved. Character extraction identifies characters along with their position on the page, whereas page segmentation stores the hierarchical structure of the document content in the form of zones, lines, words and characters. Reading order is used to maintain the right order in which the structure should be read. After layout analysis, content classification is performed in two steps: firstly, initial zone classification is performed which labels each zone as metadata, reference, body or other. After initial zone classification, metadata zone classification is performed that classifies each zone into specific metadata (title, author, affiliation etc.).

Layout analysis is performed in three steps: (1) Character extraction, (2) Page segmentation, and (3) Reading order resolving. Character extractor extracts each individual character from the PDF stream along with their position on the page, width and height. Page segmentation creates a geometric hierarchical structure storing the document's content that results in representation of document as a list of pages, where each page contains a set of zones, each zone containing a set of text lines, each line contains a set of words, and finally each word representing a set of individual characters. In the final step, reading order is resolved to determine the right sequence of the elements, in which they should be read. Resolving reading order helps in zone classification to extract full text of the document in right order.

Content Classification performs the labelling and determines the role of each identified zone. This phase works in two steps, first labelling each zone in one of the four classifications: (1) Metadata, (2) Body, (3) Reference, and (4) Other. After initial zone classification, multiple classifiers such as K-means clustering, CRF, or SVM are applied for metadata and bibliographic extraction. The system achieved F score of 0.95 while classifying zones and F score of 0.775 on metadata extraction.

SectLabel [11] is ML approach that also uses CRF to extract information from a PDF. The system uses 13 different types of metadata to tag extracted information: abstract, categories, general terms, keywords, introduction, background, related work, methodology, evaluation, discussion, conclusions, acknowledgments, and references. The approach works in two steps: logical structure classification and generic section classification. Logical structure classification tags each line as one of the 23 categories proposed by Loung et al. i.e. address, affiliation, author, body text, etc. This classification is identified by features such as location, number, punctuation, and length. The second step performs the identification of the generic sections (Abstract, Methodology, Results etc.) from the PDF document. This approach focuses on finding the type of generic section from the section heading. The generic section information (such as position, first word,

and second words), header information is used by this system. The approach was evaluated on a dataset consisting of 40 research articles. The results yielded an f-score of 0.84 by using the maximum set of font features.

Klampfl and Kern [13] published a study in ESWC, that performs the reconstruction of logical structure and extracts metadata using supervised and unsupervised learning. This approach uses Apache PDFBox [19] to obtain the low-level PDF streams. These streams are then combined using Merge and Splits. Merge performs horizontal and vertical clustering, whereas Split removes the merging of the text across the columns. Using these techniques, characters are merged to form a word. These words are combined to form a line and finally lines are combined to create a complete block. The approach uses supervised learning to extract the information related to header section (Author, Affiliation, Email etc.). Maximum Entropy in combination with Beam Search is used to extract and classify results and to avoid incorrect label sequencing. Key words like 'Table', 'Fig.', 'Figure' etc. were searched below/above the tables and figures to identify the captions. Sections headings were identified by using labelled text blocks in combination with the geometrical features. Multiple heuristics were applied after the extraction of the section heading, to make the section heading identification more accurate. Once all the information is extracted, it is stored in the form of RDF triples. This technique was prepared by using the training dataset provided by ESWC, consisting of 45 papers. The approach achieved an f- score of 0.592.

2.3 Hybrid Approaches

Hybrid approaches work by combination of multiple approaches. These approaches incorporate rule-based approaches with ML approaches, and they also combine several other data warehousing techniques with machine learning or rule-based approaches to extract metadata information.

Sateli and Witte [17] published a study in ESWC, that combines the LOD- based NER tool with rule-based approach to extract metadata information from PDF documents. The approach works by converting a PDF document into textual format and tags each part of each sentence as a part of speech. After tagging, each word is stored in its base format, to remove the likeliness of morphological variations. After performing the syntactic processing, the approach performs semantic processing in iterative phases, adding more and more annotations in each phase. Based on this tagged information from the semantic processing, manually developed rules are applied to extract information from the PDF. Authors' names are extracted by using gazetter, which helps in recognizing common first names and tags them as 'Author'. Affiliation and Country extraction is performed by annotating lines of metadata section (part of research article between title and abstract) using the LOD cloud. Afterwards, the annotated information is passed through a set of rules to extract the affiliation of the research article. Information regarding tables, figures, and section headings is extracted in syntactic phase wherein terms are annotated as the metadata information. If any of this information is not found, then, for tables and figures, a set of trigger words is used, and section headings are checked against gazetter to find conventional research article headings (Introduction, Conclusion, Experiments etc.). This approach was evaluated on the training dataset of ESWC, consisting of 45 research articles, and achieved an f-score of 0.63.

Another hybrid approach proposed by Tuarob et al. recognizes hierarchical sections from the PDF document [16]. The system automatically recognizes section boundaries and standard sections of the research articles. The approach proposes 22 different features that can be used to identify section boundaries. These identified features can mainly be characterized into: (1) Pattern based, (2) Style based, and (3) Structure based. Pattern based features are used for finding standard sections of the PDF document. Style features help in removing lines that are not part of a section, such as tables, figures or captions. The structure features are used to identify the location of the section in the PDF document helping in the identification of the section more accurately. Multiple classifiers like SVM, RIPPER, RF and Naïve Bayes are used to identify section boundaries. A proposed set of rules is applied on the sections, to identify them as Abstract, Introduction, Background, Conclusion, and Acknowledgment. The approach was evaluated on the dataset comprising of over 200 PDF documents, selected from CiteseerX. The approach achieved an accuracy of 92.38% and 96% for section boundary recognition and section identification respectively. However, it focuses on extracting the textual content of the PDF document, ignoring figures, tables, and listings etc.

PDFMEF [7] is an open-source multi-knowledge extraction framework that performs extraction of metadata by incorporating multiple open-source systems. The open source

systems are used for the identification of metadata. GROBID is used for header information (author, email, affiliation etc.), whereas PDF Figure is used for table, figures and algorithm extraction, and ParsCit is used for extracting the information regarding citation. The performance of PDFMEF is based on the underlying open-source software used for the extraction. The f-score of header section is same as the f-score obtained by the GROBID. In the same manner, the accuracy and f-score of extracting figures, tables, algorithm and citation depends on PDFFigure and ParsCit.

4. ANALYSIS

The critical review delineated in the above section states that contemporary PDF data extraction techniques belong to three broad categories: (1) Rule based approaches, (2) Machine learning approaches, and (3) Hybrid approaches. All of these techniques have their merits and demerits. There does not exist any approach that could be deemed as generalized and applicable in all the scenarios. The following table concretely recapitulates the existing PDF approaches.

Table 1 Recapitulation of state-of-the-art PDF data extraction techniques

Paper	Approach	Category	Accuracy	Limitation
Ramakrishnan et al. [4]	Performs layout analysis	Rule-based approach		The system does not include extraction from graphs, tables, citations and figures.
Jahongir and Jumabek [18]	Extracts metadata of research papers	Rule-based approach	Achieved accuracy of 0.96	The formed rules are not generalized and approach does not behave similar for different data sets.
Riaz et al. [3]	Extracts metadata of PDF documents	Rule-based approach	F-score of 0.77	The extraction of metadata is dependent on the output generated by PDFX. The approach has been evaluated on a very small data set.
Klink and Kieninger [20]	Extracts metadata of PDF documents using textual and physical features of the PDF		Precision of 0.96	The approach follows only one rule for each information it extracts. There should be more rules for diverse extractions.
GROBID [9]	Extracts metadata using Conditional Random Field Algorithm.	Machine-learning based approach	Accuracy of 83.2%,	The results may possibly be right only if the title and the first author information is correctly identified by the system.
CERMINE [10]	Extracts metadata and content from PDF files using K-means clustering, CRF, and SVM classifiers	Machine-learning based approach	F score of 0.775	Text extraction is dependent on libraries (iText). Headings, tables, and figures have not been extracted.
SectLabel [11]	Performs logical structure classification and generic section classification of PDF files	Machine-learning based approach	F-score of 0.84	The results have been evaluated on a very small data set.
Klamp and Kern [13]	Performs reconstruction of logical structure and extracts metadata using supervised and	Machine-learning based approach	F-score of 0.592.	The data set contains only 45 research articles, which is very small to be used as a training dataset for ML approach.

	unsupervised learning.			
Sateli and Witte [17]	Combines the LOD-based NER tool with rule-based approach	Hybrid approach	F-score of 0.63.	The rules crafted in this approach were specific to dataset. The size of dataset to develop technique was also small.
Tuarob et al. [16]	The approach uses SVM, RIPPER, RF and NaiveBayes classifiers and Pattern based, Style based, and Structure based features to extract PDF data	Hybrid approach	Accuracy of 92.38%	The approach only focuses on extracting the textual content of the PDF document, and ignores figures, tables, and listings etc.

Most of the rule-based approaches convert PDF document into the XML or plain text format. Applying rules on the converted XML or plain text document is much easier than on the PDF itself. Although the development of rules/ heuristics become much easier, most of the tools that convert PDF into XML format or plain text format, do not fully support all the characters and information gets removed from the converted text, which results in incorrect extraction of the information. Another problem with the rule-based approaches is that, they are not generalized and perform well only on the data set for which the rules have been designed. Moreover, the preparation of rules/ heuristics is also a challenging task and involves cognitive process. As the dataset becomes large, the rules to extract the information becomes more complex and requires more effort to identify different formats to cater all the format in the rules.

Machine learning PDF extraction approaches are more dependent on the obtained feature set from the dataset and a large dataset. Large tagged datasets help in training the system more effectively and extracts the information more accurately. With more training data, the model built by the ML system becomes more effective and accurate. The second challenging task in ML approach is features extraction. The feature extraction methodology must provide correct feature description, as features are main building block in ML approaches to extract and identify the information.

Hybrid approaches incorporate multiple approaches and provide a solution to identify the logical sections or metadata of the PDF document. The problem with the hybrid approaches is that they inherit problems from their parent approaches. Generally, these approaches require a large tagged dataset to train the model more effectively. Also, the feature extractor needs to extract the feature with high precision so that the model could be trained effectively. The required rules are more generalized and complex to create. With the large tagged dataset, the created rules are more complex and demand more critical analysis of the PDF documents.

5. CONCLUSION

In this study, we have presented an in-depth review of existing PDF data extraction techniques. Most of the scholarly literature on the web is stored in the form of PDF files. However, PDF has a more complex structure than other formats. Due to lack of proper understating about the PDF structure, PDF data extraction systems fail to extract information in a cohesive manner. Researchers have proposed various techniques to extract data from PDF files. These techniques are broadly categorized into three categories: (1) Rule-based approaches, (2) Machine learning based approaches, and (3) hybrid approaches. In rule-based approaches, different rules are crafted according to data sets. These rules are then used to extract metadata or content of documents. Similarly, in machine learning PDF extraction approaches, features are formed and supervised and unsupervised learning is applied on them to extract data. Hybrid approaches are combination of rule-based and machine learning based approaches. The analysis revealed that there does not exist any approach that could be deemed as an optimal for all the scenarios. Most of them depend on certain aspects due to which they fail to maintain similar behavior in different scenarios.

For instance, rule-based approaches perform well only for those data sets according to which the rules are formed. Similarly, other categories also have their own limitations and parameters to perform well. In future, scientific community should focus on proposing such PDF data extraction system that rarely depends on features and is generalized enough to maintain similar patterns for different scenarios and data sets.

REFERENCES

[1] Larsen, P. & Von Ins, M., "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index", Scientometrics, 84(3), 2010, pp. 575-603.

[2] Nissim, N., Cohen, A., Wu, J., Lanzi, A., Rokach, L., Elovici, Y. & Giles, L., "Sec-Lib: Protecting Scholarly Digital Libraries from Infected Papers Using Active Machine Learning Framework". IEEE Access, 7, 2019, 110050-110073.

[3] Ahmad, R., Afzal, M. T. & Qadir, M. A., "Information extraction for PDF sources based on rule-based system using integrated formats", In Communications in Computer and Information Science, Springer, 2016, pp. 293–308.

[4] Ramakrishnan, C., Patnia, A., Hovy, E., APC Burns, G., "Layout-aware text extraction from full text PDF of scientific articles. Source Code for Biology and Medicine", In Source code for biology and medicine, Springer, 7(1), 2012, pp. 1–7.

[5] O' Donoghue, DP., Saggion, H., Dong, F., Hurley, D., Abgaz, Y., Zheng, X., Corcho, O., Zhang, J. J., Careil, J-M., Mahdian, B. et al., "Towards Dr Inventor: A Tool for Promoting Scientific Creativity", In Proceedings of the Fifth International Conference on Computational Creativity ICCC, 2014, pp. 268–271.

[6] Constantin, A., Pettifer, S., and A. Voronkov, "PDFX: fully-automated pdf-to-xml conversion of scientific literature", In DocEng' 13, 2013, pp. 177–180.

[7] D´ejean, H., Meunier, J-L., "A system for converting PDF documents into structured XML format", In Lecture Notes in Computer Science, Springer, 2006, pp. 129–140.

[8] "Extended Semantic Web Conference Task 2", https://github.com/ceurws/lod/wiki/SemPub16 Task2/, accessed on 30 October 2018.

[9] Lopez, P., "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications", In Research and Advanced Technology for Digital Libraries, Springer, 2009, pp. 473–474, 2009.

[10] Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., Bolikowski, L., "CERMINE: automatic extraction of structured metadata from scientific literature", In International Journal on Document Analysis and Recognition (IJDAR), Springer, 18(4), 2015, pp. 317–335.

[11] Luong, M-T., Nguyen, T.D., Kan, M-Y., "Logical Structure Recovery in Scholarly Articles with Rich Document Features", In International Journal of Digital Library Systems (IJDLS), 1(4), 2012, pp. 1-23.

[12] Sutton, C., and McCallum, A., "An introduction to conditional random fields for relational learning", In Introduction to statistical relational learning, 4(4), 2010, pp. 267–37.

[13] Klampfl, S., Kern, R., "Reconstructing the Logical Structure of a Scientific Publication Using Machine Learning", In Communications in Computer and Information Science, Springer, 2016, pp. 255–268.

[14] Wu, J., Killian, J., Yang, H., Williams, K., Choudhury, S. R., Tuarob, S., Caragea, C., and Giles, C. L., "Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search", In Proceedings of the 8th International Conference on Knowledge Capture, 2015, pp. 1–8.

[15] Councill, I.G., Giles, C.L., Kan, M.Y., "Parscit: An open-source CRF reference string parsing package", In Language Resources and Evaluation Conference, 2008, pp. 661–667.

[16] Tuarob, S., Mitra, P., Giles, C.L., "A Hybrid Approach to Discover Semantic Hierarchical Sections in Scholarly Documents", In 13th International Conference on Document Analysis and Recognition (ICDAR), Springer, 2015, pp. 1081–1085.

[17] Sateli, B., Witte, R., "An Automatic Workflow for the Formalization of Scholarly Articles' Structural and Semantic Elements," In Communications in Computer and Information Science, Springer, 2016, pp. 309–320.

[18] Jahongir, A., Jumabek, A., "Rule Based Metadata Extraction Framework from Academic Articles", In 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, 2018, pp. 400–404.

[19] "Apache PDFBox – Java library", https://pdfbox.apache.org/, accessed on 12 November 2018.

[20] Klink, S. and Kieninger, T., "Rule-based document structure understanding with a fuzzy combination of layout and textual features", In International Journal on Document Analysis and Recognition, Springer, 2001, pp. 18–26.

[21] Tkaczyk, D. et al. "A modular metadata extraction system for born-digital articles", In 10th IAPR International Workshop on Document Analysis Systems, 2012, pp. 11–16.

Ahmer Maqsood Hashmi has earned his MS (Computer Science) from Capital University of Science and Technology (CUST), Islamabad, Pakistan. He has completed his BS (Computer Science) from CUST in 2016. His research area is web and Information systems. He is working as a Software Engineer in Teradata, Pakistan.

Faiza Qayyum is PhD (Computer Science) scholar at Capital University of Science and Technology (CUST), Islamabad, Pakistan. She has completed her MS (Computer Science) from CUST in 2017. She works as a Research Associate in the Department of Computer Science at CUST. Her research area is web and Information systems.

Muhammad Tanvir Afzal received the PhD degree with high distinction in Computer Science from the Graz University of Technology, Austria, secured Gold medal in his M.Sc. Computer Science from Quiad-i-Azam University, Islamabad, Pakistan. He has been associated with academia and industry at various levels for the last 20 years, and currently he is serving as Professor in the Department of Computer Science at Capital University of Science and Technology, Islamabad. He is also serving as editor-in-chief for reputed impact factor journal known as: Journal of Universal Computer Science. Dr. Afzal authored more than 90 research papers in the field of Digital Libraries, Information retrieval and visualization, Semantics, and Scientometrics including two books. His ISI impact factor is 50+. With citations over 400. He played pivotal role in making collaborations between MAJU-JUCS, MAJU-IICM, and TUG-UNIMAS. He served as PhD symposium chair, session chair, finance chair, committee member, and editor of several IEEE, ACM, Springer, Elsevier international conferences and journals. Dr. Afzal conducted more than 100 curricular, co-curricular, and extra- curricular activities in the last 5 years including seminars, workshops, national competitions (ExclTeCup) and invited international and national speakers from Google, Oracle, IICM, IFIS, SEGA Europe etc. Under his supervision, more than 50 post grad students (MS and PhD) have defended their research theses successfully and a number of PhD and MS students are pursuing their research with him.