CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Rule Based Approach to Extract Metadata from Scientific Papers

by

Ahmer Maqsood Hashmi

A thesis submitted in partial fulfillment for the degree of Master of Science

in the
Faculty of Computing
Department of Computer Science

2019

CERTIFICATE OF APPROVAL

Rule Based Approach to Extract Metadata from Scientific Papers

by

Ahmer Maqsood Hashmi

MCS171014

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Waseem Shahzad	FAST, Islamabad
(b)	Internal Examiner	Dr. Abdul Basit	CUST, Islamabad
(c)	Supervisor	Dr. Muhammad Tanvir Afzal	CUST, Islamabad

Dr. Muhammad Tanvir Afzal
Thesis Supervisor
October, 2019

Dr. Nayyer Masood Head Dept. of Computer Science October, 2019	Dr. Muhammad Abdul Qadir Dean Faculty of Computing October, 2019

Author's Declaration

I, **Ahmer Maqsood Hashmi** hereby state that my MS thesis titled "**Rule Based Approach to Extract Metadata from Scientific Papers**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Ahmer Maqsood Hashmi)

Registration No: MCS171014

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Rule Based Approach to Extract Metadata from Scientific Papers**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been dully acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Ahmer Maqsood Hashmi)

Registration No: MCS171014

Acknowledgements

I would like to dedicate this thesis to my parents, because of whom I was able to do it. I would like to thank all my friends and family for always being on my side and support me. I would espacially like to thank my supervisor, who always encouraged me and was always there for my support.

(Ahmer Maqsood Hashmi)

Registration No: MCS171014

Abstract

Most of the research articles are available in the form of PDF. Information extraction from the documents is very time consuming task and requires a lot of human effort. Multiple approaches have been developed to extract information from the PDF documents, that performs extraction using the document text or font features. These techniques can be classified into three major categories: (1) Rule/ Heuristic based, (2) Machine Learning Based, and (3) Hybrid. The rule based approaches have high accuracy and better results than other approaches, however these approaches tend to be dependent on dataset. Machine learning approaches on the other hand are not dependent on dataset and provide a generalized solution, however they requires a large tagged dataset for training. The hybrid approaches combine rule based and machine learning approaches to extract information from the PDFs, but they inherit problems from their parent approaches .i.e. a large tagged dataset and generalized rules. In this research thesis, we propose a generalized rule-based approach that combines the textual features with font and geometrical features of PDF document to extract metadata. Most of the previous rule based approaches extract information from PDF document by converting it in the form of XML. Our approach incorporates XML output with geometrical features for creation of rules. Our approach receive an f-score of 0.93 on training dataset and 0.85 on test dataset. Experimental results shows that our approach performs 12 times better than GROBID and 68 times better than CERMINE on evaluation dataset.

Contents

List of Figures

List of Tables

Abbreviations

CERMINE	Content ExtRactor and MINEr
CEUR	Central Europe operated under the Umbrella of RWTH Aachen University
CRF	Conditional Random Fields
ESWC	Extended Semantic Web Confernece
F1	F Score
GROBID	GeneRation Of BIbliographic Data
LOD	Linked Open Data
ML	Machine Learning
NER	Named Entity Recognition
PDF	Portable Document Format
PDFMEF	PDF Multi-Entity knowledge Extraction Framework
SVM	Support Vector Machine
JSON	JavaScript Object Notation
XML	eXtensible Markup Language

Chapter 1

Introduction

Millions of documents are available for researchers and these documents are increasing rapidly each year. According to research conducted by Jinha [1], approximately 2.5 million scientific articles are published each year. These documents are usually available in the form of PDF. These PDFs are stored in digital libraries and citation indexes. Some of the digital libraries/ citation indexes store these PDFs by indexing them using their metadata, whereas most of the digital libraries/ citation indexes do not have explicitly stated metadata and store PDFs without metadata. When a user wants to query a document, user searches these documents using the metadata; The metadata means: Title, Author, Author's Affiliation, Country, Abstract, Funding Agency, Tables and Figures. Some of such information is available inside the text of PDF document and can help in performing document indexing according to metadata and constructing complex queries. Examples of such queries are: (1) Provide all the papers that are written by author "Edsger Wybe Dijkstra", affiliated with the institute "Eindhoven University of Technology", and has European funding agency (2) Provide all papers whose title contains "k-Map", written by the authors affiliated with the institutes in "Netherland", and are not funded by European agency. To answer such queries, it is required to extract information available inside PDF document and store such information as metadata.

Extracting information from PDF has been addressed by many researchers in the past and multiple approaches have been identified [6–11, 13, 14, 16–22]. These approaches can be categorized into three major categories (1) Rule/ Heuristic based (2) Machine Learning based (3) Hybrid. These approaches work in different ways to regenerate/ identify the structure of the document. Heuristics/ Rule based approaches apply rules on the document and identifies the common patterns, based on which heuristics are made to extract information from PDFs. Whereas ML approaches works by training the system on the tagged dataset, that has been manually tagged by the humans. Once the machine is trained, it is used for extraction of metadata. Hybrid approaches works by incorporating the heuristics and machine learning together for the extraction of metadata from the PDF document.

ESWC (Extended Semantic Web Conference) [3] in 2016 addressed the problem [4] of extracting information from the PDF. The extracted information included (1) Author (2) Affiliation (3) Country (4) Section heading/ title (5) Table number/ caption (6) Figure number/ caption (7) Supplementary material and (8) Funding agency. Researchers from all over the world, participated and proposed approaches that used heuristics, machine learning, data mining, or combination of multiple approaches for metadata extraction. The best performing approach was Riaz et. al. approach[8] that used heuristics for metadata extraction and attained an f-score of 0.771 [5]. This approach was made by critically analysing the research articles, provided in the dataset of ESWC, and analysing the common patters in those papers. The heuristics/ rules proposed in this approach were developed for ESWC dataset and were not generalized.

Multiple approaches have been proposed by researchers for extracting metadata from the PDF documents and much more research is being performed in this area. The most challenging task in this area is large variety of scientific documents and formats available world wide. Researchers have proposed many approaches but, these cannot guarantee to extract the exact information from the PDF document. The approaches that are rule based have the problem that they are not generalized and most of the times works only on the dataset, they are prepared from. ML approaches are generalized in nature and require very small changes if dataset is changed, but these approaches require many annotated dataset for training, which is not practically possible because it requires a lot of human effort for data annotation. The problem with the hybrid approaches is complexity of combining these approaches together and the problems inherited from their parent approaches. A detailed background knowledge and previous work is provided in Chapter 2 of this thesis.

Physically separating metadata from scholarly PDF documents is very challenging task and requires a great degree of effort. As indicated by Chrystal [23], it would take around 60 representative years to make a metadata gathering for 1 million reports. To provide solutions for this problem, there needs to be a mechanism by which extraction from PDF can be performed in an automated way. Due to the large variety of different formats and documents, creating generalized rules or creating a ML approach that can extract information by its learning is very difficult task.

1.1 Purpose

The purpose of this research is to make use of physical and textual features of the PDF document. Physical features like (1) Page number (2) Top (3) Height (4) Width (5) Left (6) Font Size (7) Bold (8) Italic can be used in much more accurate extraction of data. In our approach, we incorporated physical features, layout features and textual features of the document to extract the metadata information available inside PDF document. We converted the PDF documents into XML format for extracting the metadata, as done by the Riaz et. al. [8]. We performed critical analysis of the xml format and combined textual features with physical features of PDF document. We created our approach by using training dataset of ESWC [4]. Our approach achieved a 20% increase in results than Riaz et. al. approach[8] for extracting metadata information from PDF documents. In this research thesis, we created a generalized rule-based approach for extraction of metadata and evaluated our approach on multiple datasets having different formats having different physical features achieving an f-score of 0.97 on training dataset and 0.87 on evaluation dataset.

1.2 Problem Statemnet

This thesis focuses on building a rule-based approach for metadata extraction from PDF documents.

1.3 Scope

Our approach makes use of the physical features while extracting information from the PDF. The proposed approach was evaluated on the dataset provided by ESWC as well as a dataset prepared in-house. The dataset was composed of research articles from different fields of study and had different formats. Complete statistics of datasets are provided in later sections of this thesis.

1.4 Significance of the Solution

This research thesis creates a generalized rule-based for extracting metadata information from the research articles. We used textual features, in combination with the physical features of a PDF document for creating generalized. After performing critical analysis of the research articles from ESWC training dataset, were able to identify the common patterns in the text and created rules based on these identified patterns. This approach is benefited from textual, as well as physical features of a PDF document, creating rules as much generalized as possible.

Chapter 2

Literature Review

There are many approaches [6–11, 13, 14, 16–22] proposed by the researchers in the past to extract information available inside the PDF documents. These approaches can be categorized into three types: (1) Rule/ Heuristic based (2) Machine Learning based (3) Hybrid. This section provides a comprehensive literature review of the research conducted in this area and provides a critical overview of all the proposed approaches.

The logical extraction is the identification of the document into to the logical sections, such as header, footer, abstract, etc. There are multiple techniques provided by researchers to evaluate the logical structure of the PDF. The approach proposed by Ramakrishnan *et al.* [11] 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 [10] is another framework to extract the logical structure of the document, by converting the PDF document into XML format. This conversion from PDF to XML helps in extra processing on the document, which provides easiness in identifying the section. A detailed information is provide in comming section, on how converting PDF to XML helps in crafting of rules. PDFX [6] converts PDF document into XML format 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 as what type of information that tag contains. Déjean and Meunier [9] proposed an approach to convert the PDF document into XML format. Their approach converts the streams available in the form of PDF to structured XML. The 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 much more effective and easier. Riaz *et al.* approach [8] proposed in ESWC [4], works by converting PDF document into XML format using PDFX. The identified XML tags are then passed from multiple heuristics, that results in extraction of metadata. This approach was considered best approach for extracting metadata in ESWC [4], securing an f-score of 0.77.

Researchers have proposed many machine learning approaches [16–19, 21] that extracts information from the PDF document. GROBID [18] extracts metadata using machine learning approach and then generates a web request to extract the bibliographic information of PDF document. Another approach CERMINE [17] extracts information from pdf document in multiple steps. It performs the layout analysis and then based on that layout analysis, classifies the type of metadata. SectLabel [16] performs metadata extraction and content classification using CRF [25], a machine learning approach that performs a probabilistic structure prediction using large set of input features. Klampfl and Kern proposed an approach [19] that works by using unsupervised learning to extract metadata from the PDF documents. Their approach performs logical structure analysis of the PDF document and extracts information from that logical structure analysis using unsupervised learning.

The hybrid approaches have always been in consideration by many researchers in the past. PDFMEF [7], combines the opensource frameworks such as GROBID [18], CERMINE [17], ParsCit [24] to extract the information from the PDF document. Tuarob *et al.* proposed an approach [22] that identifies the section boundaries using machine learning approach and then label these identified sections as standard identification (Abstract, Introduction, Background, and Experiment, etc.) using heuristics prepared by analysing these PDF documents. Sateli and Witte [20] combines the LOD-based Named Entity Recognition (NER) tool with the rule-based approach to extract information from the PDF document. This approach was considered second best approach in ESWC [4], securing an f-score of 0.61 for extraction of information from the PDF document.

2.1 Rule-based Approaches

Rule based approaches are created by finding out the common patterns in the documents after a critical analysis of the dataset. Based on these identified patterns, multiple rules or heuristics are made to extract the information available inside the PDF. This section discusses multiple rule-based approaches proposed by the researchers that are most related to this research thesis.

Jahongir and Jumabek [13] performs the extraction of metadata from the PDF document. Their approach works in three steps (1) Classification of the PDF files, (2) Metadata extraction, and (3) Storing of PDF files in the form of XML or JSON. Fist 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 passed on to the second step for metadata extraction. The second step performs the metadata extraction and outputs the extracted metadata. The textual and font features are extracted using the Apache PDFBox [12]. 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 given in Table 2.1. The final step of their methodology stores the extracted information in the form of XML or JSON. Their approach achieved and accuracy of 97.71% for document classification and 96.31% for metadata extraction. Although the approach achieved a very high accuracy on the evaluation dataset, however

Table 2.1: Metadata extraction rules proposed by Jahongir & Jumabek [13]

Metadata Property	Starting key phrase	Ending key phrase
ABSTRACT	Abstract or ABSTRACT	KEYWORDS or Keywords or INDEX TERMS or Index Terms
KEYWORDS	KEYWORDS or Keywords or INDEX TERMS or Index Terms	I. Intro, 1. Intro or Intro
BODY TEXT	I. Intro, 1. Intro or Intro	Conclusion or CONCLUSION
CONCLUSION	Conclusion or CONCLUSION	Reference or REFERENCES or ACKNOWLADG-MENT or Acknowledgement
REFERNCE	REFERENCE or Reference	till the end of file

the rules provided are not genialized and do not work on different datasets. Our result section provides a detailed analysis of the rules proposed in this approach.

Riaz *et al.* approach proposed in ESWC [4] proposed a rule-based approach for extracting information from the PDF documents. Their approach works by converting the 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 cleans the identified text. Metadata splitter splits each extracted metadata and outputs the actual extraction information. The approach starts its working by converting the 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 by the Riaz *et al.* approach to extract the actual metadata. The paper converts the PDF document and output the 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 by the identified heuristics. Once these are identified, Country is extracted from the affiliation part using a predefined country list that contains the names of all the countries in the world. After the extraction of author and affiliation, the author is affiliated with the respective author, generating an output that contains the author, affiliation and country.

The 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 the XML tags. Tables are extracted by “`>(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 figure information is extracted using multiple regular expressions developed from the XML format. If one regex do not return any output, then another regex is applied for the extraction of figures. Once figure/s are extracted, then 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 made by critically analysing 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 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 analysing 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 then 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 that is extracted by 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 [4] 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 [20], that will be discussed in later sections.

Riaz *et al.* approach used PDFX (an opensource 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: (1) First stage constructs the geometrical model using the content of the article, and (2) 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 the identification of different of logical units.

The most basic logical separation is performed using the font size, whenever a font size is changed, a new logical unit has started. Furthermore font frequency graphs are used, that separates the common text (section text) with the rare text (title, heading text, tables/ figures text etc.). The tool converts the PDF document into small text blocks and merges these small blocks afterwards, using the font and geometrical features. After the merging of the textual blocks, with reading order in consideration, multiple rules are applied to label each logical unit as title, author, email, section, figure, refence, 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 [14] also incorporates the textual and physical features of the PDF for the extraction of information from the PDF document. The proposed approach constructs the logical structure of PDF document and identifies the header, footer, body text, table and listings. Header section is identified by start reading from the top of the page until a very large gap than usual is found in the reading. In the same manner footer is identified. Lists (bulleted, numbered or dashed) are identified by using the heuristic that the first character will be number, enumeration, dash, bullet or dot. Body text is also identified by using the geometrical features such as start of the block, spacing between blocks and change of font features. The identification of table performed by this approach uses the algorithm proposed in T-Recs [15]. This approach was also evaluated on the University of Washington document corpus and the letterheads received by the German Research Center for Artificial Intelligence. They achieved precision of 0.98 for 90% documents. This approach proposes only one rule for each information it extracts. It can be further enhanced by using a set of rules that can make extraction more diverse.

Most of the rule-based approaches we have studied, converts the PDF document in to 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, however most of the tools that converts the PDF into XML format or plain text format, do not fully support all the character 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 works on the dataset they are prepared from. The preparation of rules/ heuristics is also a challenging task. As the dataset grow bigger, the rules to extract the information becomes more complex and requires more effort to identify different formats and cater all the format in the rules.

2.2 Machine Learning Approaches

Machine learning approaches provide a way to make system learn the different formats and features. Using this learning, the system can extract the information from the PDF document in an automated manner. This section provides the machine learning approaches and a comprehensive overview of how these approaches work.

GROBID [18] is an opensource ML library that performs the extraction, parsing and reconstruction of the PDF document into structured text. The system works by extracting the title, author, abstract etc. using the Conditional Random Field algorithm. After the identification of the information, the system generates a web request that generates full metadata of the publisher. The approach achieves an accuracy of 83.2%, however the results may possibly be right only if the title and first author information is identified correctly by the system. This system is now available as an opensource tool and is in process of constant development.

CERMINE [17] is also an opensource ML tool that extracts the metadata and content from the PDF document and generate the 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 the 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 structure should be read. After layout analysis, content classification is performed in two steps. First initial zone classification is performed which label 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 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 the full text of the document in right order.

Content Classification performs the labelling and determine the role of each identified zone. This phase works in two steps, first labelling each zone in one of the four classification: (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 received F score of 0.95 while classifying zones and an F score of 0.775 on metadata extraction.

SectLabel [16] is ML approach that also uses CRF to extract the information form the PDF. The system uses 13 different types of metadata to tag the 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.) form the PDF document. This approach focuses on finding the type of generic section from the section heading. To identify the type of generic section features such as position, first and second words, and whole header information is used. The approach was evaluated on a dataset consisting of 40 research articles, receiving an f-score of 0.84 by using the maximum set of font features. Klampfl and Kern [19] proposed an approach in ESWC, that performs the reconstruction of logical structure and extracts metadata using supervised and unsupervised learning. This approach uses Apache PDFBox [12] 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 column. 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 for extracting and classifying the results and avoid the 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, the extracted information is stored in the form of RDF triples. This technique was prepared by using the training dataset provide by ESWC, consisting of 45 papers. The approach achieved an f-score of 0.592.

Machine learning approaches are more dependent on the obtained feature set from the dataset and a large dataset. Large tagged dataset helps in training the system more effectively and extract the information more accurately. With more training data, the model build by the ML system will be more effective and accurate. The second challenging task in ML approach is extraction of the features. The methodology to extract features should work correctly, to provide correct feature description, as features are main building block in ML approaches to extract and identify the information.

2.3 Hybrid Approaches

Hybrid approaches work by combination of multiple approaches. These approaches incorporate rule-based approaches with ML approaches, as well as, they combine several other data warehousing techniques with machine learning or rule-based approaches to extract the metadata information. This section will provide the related work preceded in this area of research.

Sateli and Witte [20] proposed an approach in ESWC, that combines the LOD-based NER tool with rule-based approach to extract the metadata information from the PDF documents. The approach works by converting the PDF document into textual format and tags each part of sentence as a part of speech. After tagging each word is stored in their base format, to remove the likeliness of morphological variations. After performing the syntactic processing, the approach performs semantic processing in iterative phase, 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. Author were extracted by using the gazetter, that helps in recognizing the common first names and tag them as 'Author'. Affiliation and Country extraction was performed by annotating the lines of metadata section (part of research article between title and abstract) using the LOD cloud. Afterwards the annotated information is passed from a set of rules to extract the affiliation of the research article. Information regarding tables, figures and section headings are extracted in syntactic phase where terms are annotated as the metadata information they are. If any of these 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, achieving an f-score of 0.63.

Another hybrid approach proposed by Tuarob *et al.* [22] recognizes the hierarchical sections from the PDF document. The system automatically recognizes the section boundaries and recognize the standard sections of the research article. The approach proposes 22 different features that can be used to identify the 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 the standard sections of the PDF document. Style features helps in removing the lines that are not part of 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 section more accurately. Multiple classifiers like SVM, RIPPER, RF and NaiveBayes are used to identify the section boundaries. A proposed set of rules are applied on the sections, to identify them as Abstract, Introduction, Background, Conclusion, and Acknowledgment. The approach was created and evaluated on the dataset comprising of over 200 PDF documents, selected from CiteseerX. They achieved an accuracy of 92.38% and 96% for section boundary recognition and section identification respectively. The focuses on extracting the textual content of the PDF document, ignoring the figures, tables, and listings etc.

PDFMEF [7] is an opensource multi-knowledge extraction framework, that performs the extraction of metadata by incorporating multiple opensource systems. The opensource systems are used for the identification of metadata. GROBID is used for header information (author, email, affiliation etc.), whereas PDFFigure for table, figures and algorithm extraction, and ParsCit for extracting the information regarding citation. The performance of PDFMEF is based on the underlying opensource 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. Hybrid approaches tend to 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 tend to inherit problems from their parent approaches. Generally, these approaches require a large tagged dataset to train the model more effectively. Also, the feature extractor to extract the features needs to extract the feature with high precision to train the model correctly. The rules required are more generalized and complex to create. With the large tagged dataset, the rules created for the extraction are more complex and require more critical analysis of the PDF documents.

Chapter 3

Proposed Methodology

We have discussed multiple techniques provided by researchers in the past: (1) Rule Based, (2) Machine Learning Approach, (3) Hybrid. Rule based approaches requires less human effort and can be made easily by using a small dataset, however the problem with these approaches is that rules created by these approaches are not generalized and are not generic in nature. Machine learning approaches require a large training dataset having variety of font features and a system to extract these font features correctly. Providing a large tagged diverse dataset is very challenging task, as it will require gathering a dataset and after that tagging all those extracted features. These tasks are very time consuming and requires a large amount of human effort. Hybrid approaches provide solution for extraction of metadata from PDF's by combining multiple techniques. To provide system that is a combination of rule-based and machine learning approach, will require a (1) Diverse dataset, (2) System to extract PDF features from the dataset, (3) Generalized rules, and (4) Tagging the dataset. As already discussed these are very timeous and time confusing tasks and requires large human effort. In this research thesis, we focus on rule-based approaches and creating this approach more generalized. Our approach provides a set of generalized rules to extract information from the PDF documents. Fig. 3.1 shows the research methodology of proposed approach. This approach works by converting the PDF document into XML format and This approach uses an already available dataset and a dataset prepared in-house for evaluation purposes. This approach consists of three phases: (1) Training phase, (2) Extraction phase, and (3) Evaluation phase. Training phase identifies the generalized rules and store them in knowledge base, which are then later used in Extraction phase to extract metadata. Evaluation phase performs the comparison of precision, recall and f-score of this proposed approach with CERMINE [17] and GROBID [18]. Fig. 3.2 shows the detailed methodology diagram of our approach.

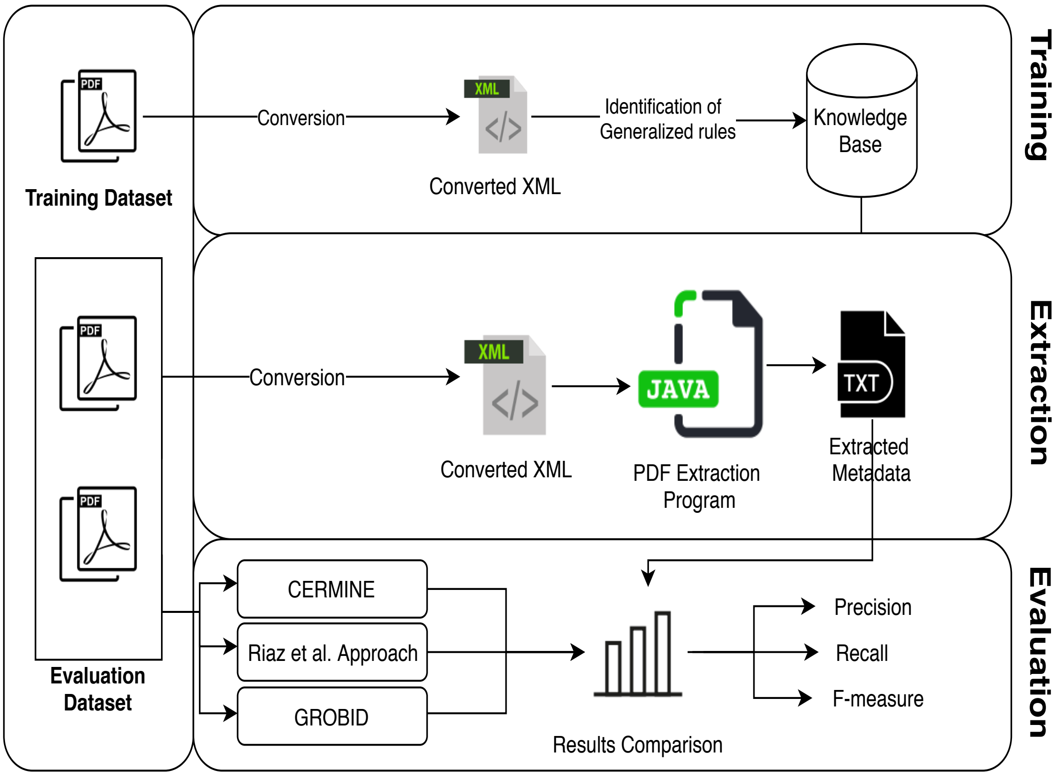


FIGURE 3.1: Research methodology diagram

3.1 Dataset Selection

This research thesis focuses on providing a set of generalized rules to extract information from the PDF documents. Creating generalized rules is only possible if dataset is diverse and contains multiple formats. Either we had to gather a dataset or select an already available dataset that was diverse enough, to make generalized rules for our approach. There were variety of datasets available for PDF extraction, however most of the datasets were focused and pointed to a specific domain.

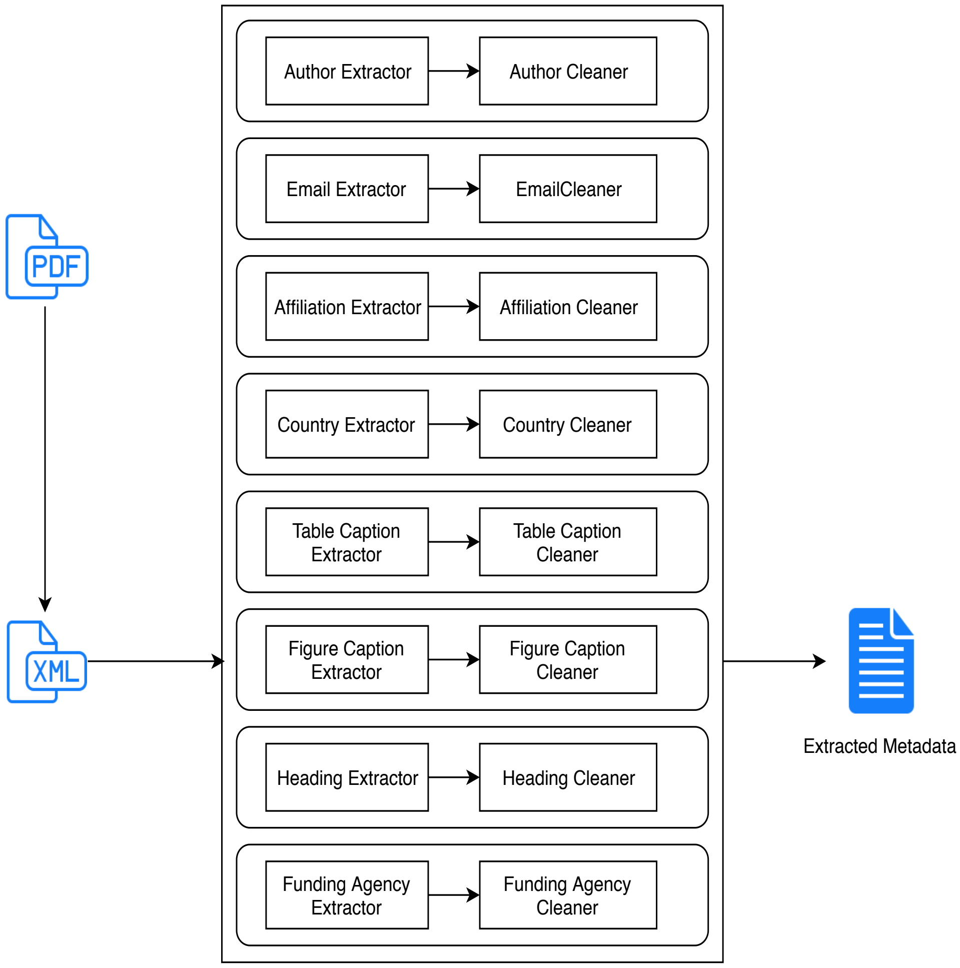


Figure 3.2: Detailed Research methodology diagram

This approach was not focused to any specific domain and was being created to provide a generalized set of rules. The selected dataset was required to have following characteristics: (1) Research articles from different journals/ conferences, (2) Research articles having rich document features, (3) Research articles with different formats, and (4) Research articles not focusing to any specific domain.

CEUR Workshop [2] is very famous open-access publication service that provides different volumes for researchers. These volumes contain research articles from the proceedings of these workshops. Research articles published in these workshops have variety of formats and features. A dataset containing research articles from CEURWS was available when this research was started. This dataset was provided in ESWC conference [3] for extracting metadata information from the research articles. The CEUR dataset contained all the characteristics that were essential to developed generalized rules. We used CEUR dataset for developing generalized rules and extracting metadata information from scientific documents. Second dataset used by this approach was prepared in-house by our research group. This dataset created by selecting research articles from different open access journals and conferences. The selected research articles belonged to different fields of science containing variety of formats and font features. This dataset was used in evaluation phase of our approach, providing a dataset to evaluate the proposed set of generalized rules.

3.2 PDF to XML Conversion

Reading the PDF documents in their base format is a timeous work. Managing the data streams and extracting document features while reading the PDF document in it's base format adds extra effort in formulation of rules. Many approaches have been proposed to extract information from the PDF document by converting it into XML format [8, 16, 26, 27]. Converting PDF document into XML format helps in decreasing the overhead of reading the PDF document from its base format and combining all its information together to regenerate the text. Also, conversion tools extract the feature information of the PDF document, which helps in training the systems or in identification of metadata.

PDFX [6] an online tool, converts the PDF document into XML format, and identify the tag names of the document. Riaz *et al.* approach used PDFX to convert PDF document into XML and then utilize its capabilities to extract information available from the PDF document. In our approach we were unable to use PDFX because of its unavailability. We used a free pdf to xml converter tool [28] for converting PDF document into XML format. This tool provides the conversion of PDF document into XML format, along with document geometrical features. Using the output of the XML format, we were able to create generalized rules by combining the text with the geometrical features of the document.

3.3 Training Phase Methodology

In this phase, we perform the identification of different font and textual features for creation of generalized rules. This approach uses CEURWS dataset for creation of generalized rules. This phase was focuses on finding following metadata properties: (1) Title, (2) Author, (3) Affiliation, (4) Country, (5) Headings level 1, (6) Table captions, (7) Figure captions, and (8) Funding agency. We converted the PDF document into XML formats and critically analysed the output. Analysis of the XML documents helped us in creation of generalized rules and extraction of metadata information much more accurate.

3.3.1 Title Extraction Methodology

Identification of title is very difficult if performed using textual features, as we have no information that how title will start and end. In our approach we have used font features to identify the title of research articles. Font features help in identification of the font used by the research articles. Most of the ML approaches rely on font features, such as font name, font family, font size for the extraction of title. In our proposed approach, wo also extracted title by using the font information gathered from XML output. Using the identified heuristic, we achieved an f-score of 1 for title identification in training dataset.

3.3.2 Email Extraction Methodology

Critical analyses of research articles helped us in identification of number of formats in which email are provided in a research article. A variety of email formats were identified and extracted from the research articles. Email identification works by identifying the header section of a research article. After identification of header section, lines containing email addresses are extracted. These lines are then passed to a set of different rules to resolve email addresses, providing all the email addresses present in research article.

3.3.3 Affiliation & Country Extraction Methodology

Affiliation part, also like email part is extracted from the header section of the research article. We identified a set keywords which were used in the affiliation part. These keywords along with some heuristics are applied together to extract the affiliation part. After the extraction of affiliation part, extra characters that are not part of the affiliation are removed, providing the affiliations in research articles. Once affiliation has been extracted, our system takes affiliation as an input, and extracts country from the extracted affiliation. Our system can extract some of the famous countries using abbreviations also. Extracting all the countries with abbreviation was not possible, because that resulted in the results that were not country.

3.3.4 Author Extraction Methodology

Author extraction along with the other header metadata properties is also extracted from header section. In our proposed system, extraction of author greatly relies on other metadata properties extracted from the header section. The relevant information of authors is extracted by author extraction unit, which afterwards is cleaned by its respective cleaner and generates authors as the output of the author extractor.

3.3.5 Figure Caption Extraction Methodology

Figure caption extractor extracts the caption of the figure. To perform the extraction of the figure caption, a set of heuristics have been identified that combine textual features with geometrical features of the PDF document. We critically analysed all the PDF documents and their converted XML output. After analysing both the outputs, we combined similar properties together to form generalized rules for extraction of figures. Our system perofrms the extraction of fgure captions and afterwards, clean any unnecessary information that is not part of figure caption, generating captions of the figures.

3.3.6 Table Caption Extraction Methodology

Table caption extractor extracts table captions from research articles. Rules created in this unit combines the set of textual features with geometrical features of PDF document to extract table captions. After the extraction of tables, the output of the table captions in cleaned and unnecessary information is removed from the captions, generating captions of tables as output.

3.3.7 Heading Extraction Methodology

The purpose of heading extractor is to extract level 1 headings of the PDF document. Heading extraction also combines geometrical features with textual features to identify headings with high accuracy. Identification of headings is a challenging task due to number of different formats. We have created generalized rules for extracting headings information. We have created regular expressions that identify different formats and we have also targeted geometrical features, like top and bottom padding of the text, to identify headings.

3.3.8 Funding Agency Extraction Methodology

Extraction of funding agency is performed by identifying the “Acknowledgment” section of a research article. After identification of “Acknowledgment” section, a set of heuristics are applied to extract the actual project/ funding agency of the research article. The identification of “Acknowledgement” is also a challenging task, as some paper write it as separate heading and some as footer of the first page. Our funding agency extractor applies generalized rules on the “Acknowledgment” section and extracts the funding agency, which is cleaned by the cleaner to provide the exact name of the agency.

3.4 Extraction Phase

Extraction phase of our approach deals with the extraction of metadata. This phase performs the actual extraction of metadata, using the generalized rules developed in training phase. This phase takes an XML file as in input and extract all the metadata. This phase is responsible to provide out put for the evaluation phase. Each unit in this phase performs as a separate entity, working as a separate unit. The final task of this unit is to combine output of the metadata units and provide user with a single extracted file.

3.5 Evalution Phase

Evaluation phase consist of evaluating our approach CERMINE and GROBID. Evaluation with Riaz *et al.* approach was not possible due to dependency of Riaz *et al.* approach on PDFX and unavailability of PDFX tool. In absence of Riaz *et al.* approach, we calculate and compare our results with CERMINE and GROBID. These tools are considered as very important tools in the field of PDF extraction, and most of the previous work performed has been compared by these approaches. We evaluate our approach using following evaluation parameters: (1) Precision, (2) Recall, and F-measure. These evaluation parameters provide the relevance and accuracy of the results.

Chapter 4

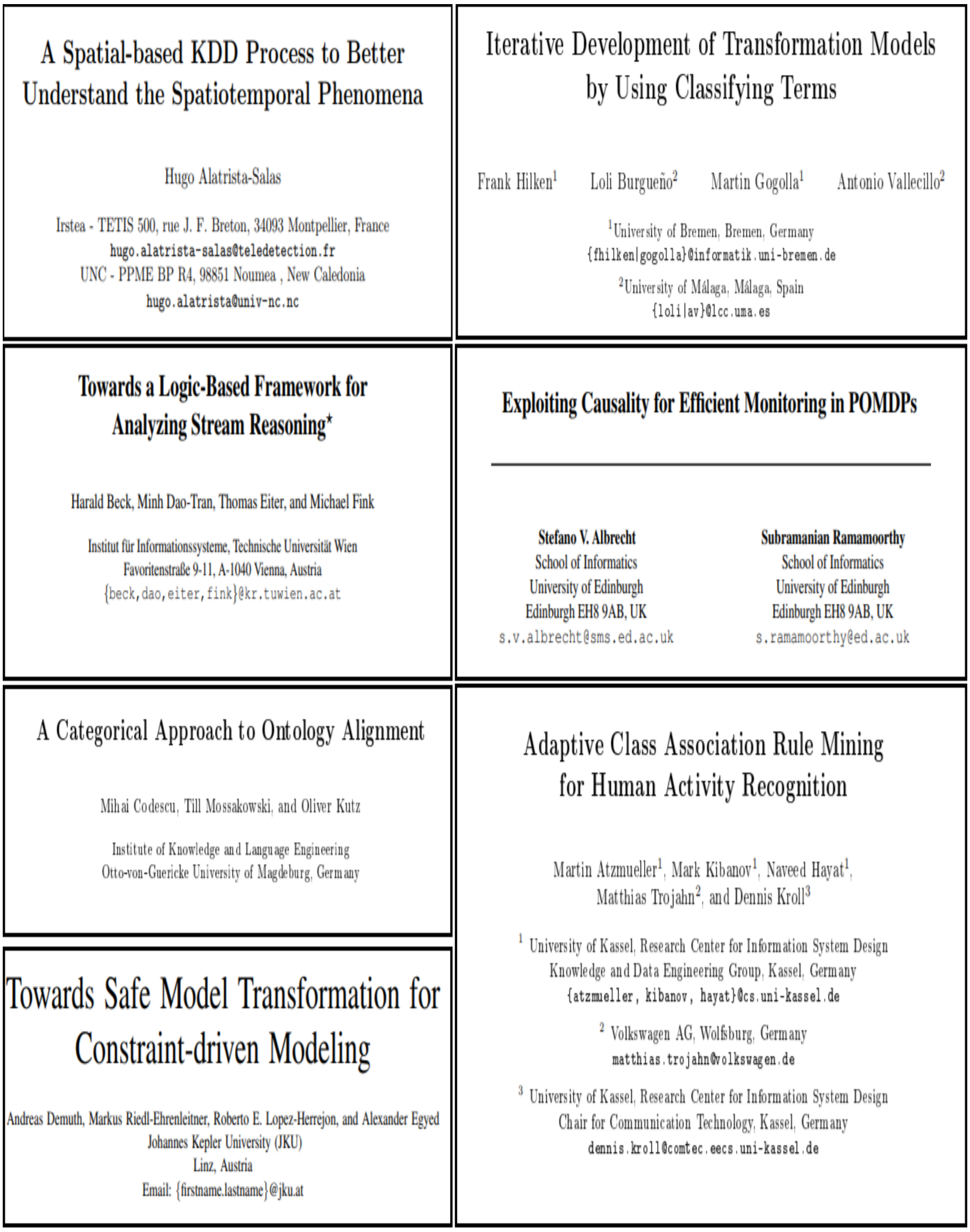
Results

4.1 Statistics of Dataset

Our approach uses multiple datasets for training and evaluation. There are two different datasets used by this approach: (1) CEUR dataset [4], and (2) Dataset prepared in-house. CEUR dataset was provided by ESWC for extracting metadata information from research articles. This dataset is available in two parts: (1) Training Dataset, and (2) Evaluation Dataset. Training dataset consists 45 research articles, whereas evaluation dataset consisting of 40 research articles from different CEURWS proceedings. The second dataset consists 120 research articles from the most recent research articles from the CEURWS proceedings.

4.1.1 Statistics of Training Dataset

These 45 research articles belong to 16 different workshop proceedings. Table 4.1 provides details of volume, proceeding and number of articles selected from that volume for training dataset. This dataset consists of variety of formats, that are used for writing the research articles. Fig. 4.1 shows some of the formats that are used for the header section (Part of research article containing information related to title, author, affiliation, country and email) of the research article. In the same manner, other metadata properties also consists of multiple formats in this dataset. Fig. 4.2 and Fig. 4.3 shows the different formats of tables and figures used in research articles. The tables and figures captions are inserted with different formats, as well as, the positions of captions also vary in research articles. In this research, this dataset is used for training and constructing the generalized rules for metadata extraction.



A Spatial-based KDD Process to Better Understand the Spatiotemporal Phenomena

Hugo Alatrista-Salas

Irstea - TETIS 500, rue J. F. Breton, 34093 Montpellier, France
hugo.alatrista-salas@teledetection.fr
UNC - PPME BP R4, 98851 Noumea , New Caledonia
hugo.alatrista@univ-nc.nc

Iterative Development of Transformation Models by Using Classifying Terms

Frank Hilken[1] Loli Burgueño[2] Martin Gogolla[1] Antonio Vallecillo[2]

[1]University of Bremen, Bremen, Germany
{fhilken|gogolla}@informatik.uni-bremen.de
[2]University of Málaga, Málaga, Spain
{loli|av}@lcc.uma.es

Towards a Logic-Based Framework for Analyzing Stream Reasoning*

Harald Beck, Minh Dao-Tran, Thomas Eiter, and Michael Fink

Institut für Informationssysteme, Technische Universität Wien
Favoritenstraße 9-11, A-1040 Vienna, Austria
{beck,dao,eiter,fink}@kr.tuwien.ac.at

Exploiting Causality for Efficient Monitoring in POMDPs

Stefano V. Albrecht
School of Informatics
University of Edinburgh
Edinburgh EH8 9AB, UK
s.v.albrecht@sms.ed.ac.uk

Subramanian Ramamoorthy
School of Informatics
University of Edinburgh
Edinburgh EH8 9AB, UK
s.ramamoorthy@ed.ac.uk

A Categorical Approach to Ontology Alignment

Mihai Codescu, Till Mossakowski, and Oliver Kutz

Institute of Knowledge and Language Engineering
Otto-von-Guericke University of Magdeburg, Germany

Adaptive Class Association Rule Mining for Human Activity Recognition

Martin Atzmueller[1], Mark Kibanov[1], Naveed Hayat[1], Matthias Trojahn[2], and Dennis Kroll[3]

[1] University of Kassel, Research Center for Information System Design
Knowledge and Data Engineering Group, Kassel, Germany
{atzmueller, kibanov, hayat}@cs.uni-kassel.de

[2] Volkswagen AG, Wolfsburg, Germany
matthias.trojahn@volkswagen.de

[3] University of Kassel, Research Center for Information System Design
Chair for Communication Technology, Kassel, Germany
dennis.kroll@comtec.eecs.uni-kassel.de

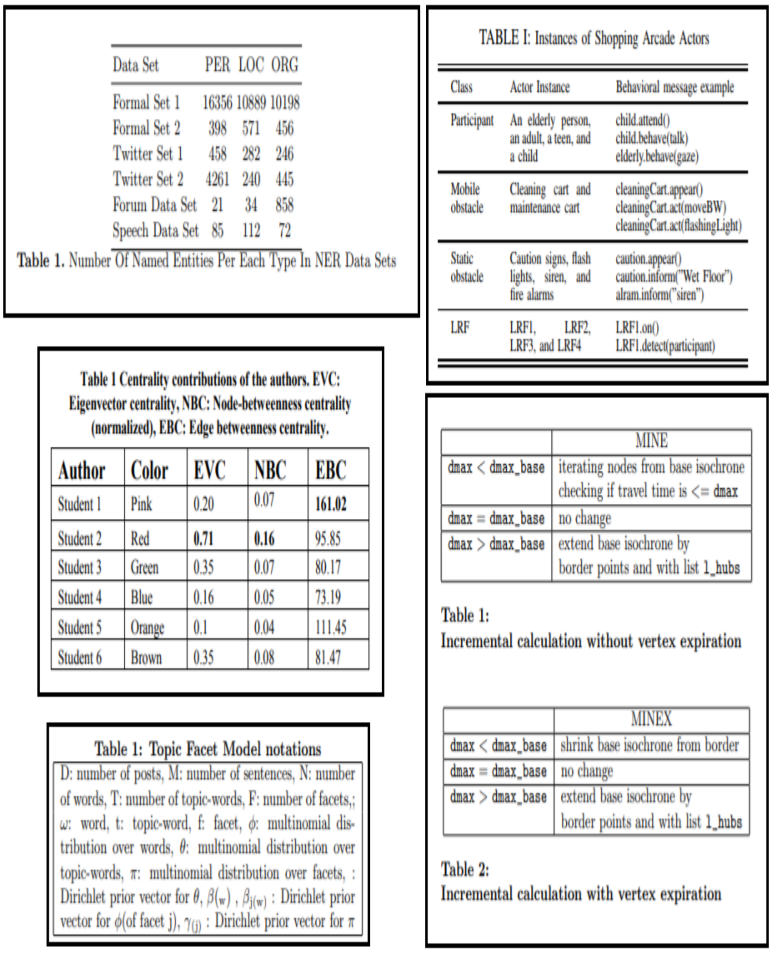
Towards Safe Model Transformation for Constraint-driven Modeling

Andreas Demuth, Markus Riedl-Ehrenleitner, Roberto E. Lopez-Herrejon, and Alexander Egyed
Johannes Kepler University (JKU)
Linz, Austria
Email: {firstname.lastname}@jku.at

Figure 4.1: Formats for header section of research articles from training dataset

Table 4.1: ESWC Training Dataset Stats

Volume	Name	No. of Research Articles
vol-1001	CAiSE 2013 Doctoral Consortium	1
vol-1006	New Generation Enterprise and Business Innovation Systems 2013	1
vol-1303	Ordering and Reasoning 2014	1
vol-1309	Workshops of ICBO 2014: DIKR 2014 / IWOOD 2014 / OBIB 2014	2
vol-1313	Grundlagen von Datenbanken 2014	1
vol-1315	Artificial Intelligence and Cognition 2014	2
vol-1317	Ontology Matching 2014	1
vol-1319	Model-Driven Robot Software Engineering 2014	1
vol-1320	Semantic Web Applications and Tools for Life Sciences 2014	4
vol-1405	Location-Aware Recommendations 2015	1
vol-1500	Analysis of Model Transformations 2015	5
vol-1504	UAI 2015 Workshop on Advances in Causal Inference	2
vol-1514	Model-Driven Engineering, Verification and Validation 2015	6
vol-1518	Visual Aspects of Learning Analytics 2015	9
vol-1521	Mining Ubiquitous and Social Environments 2015	7
vol-1531	Doctoral Symposium at MoDELS 2015	1



Data Set	PER	LOC	ORG
Formal Set 1	16356	10889	10198
Formal Set 2	398	571	456
Twitter Set 1	458	282	246
Twitter Set 2	4261	240	445
Forum Data Set	21	34	858
Speech Data Set	85	112	72

Table 1. Number Of Named Entities Per Each Type In NER Data Sets

Table 1 Centrality contributions of the authors. EVC: Eigenvector centrality, NBC: Node-betweenness centrality (normalized), EBC: Edge betweenness centrality.

Author	Color	EVC	NBC	EBC
Student 1	Pink	0.20	0.07	**161.02**
Student 2	Red	**0.71**	**0.16**	95.85
Student 3	Green	0.35	0.07	80.17
Student 4	Blue	0.16	0.05	73.19
Student 5	Orange	0.1	0.04	111.45
Student 6	Brown	0.35	0.08	81.47

Table 1: Topic Facet Model notations

D: number of posts, M: number of sentences, N: number of words, T: number of topic-words, F: number of facets,; ω: word, t: topic-word, f: facet, ϕ: multinomial distribution over words, θ: multinomial distribution over topic-words, π: multinomial distribution over facets, : Dirichlet prior vector for θ, $\beta(_w)$, $\beta_{j(w)}$: Dirichlet prior vector for ϕ(of facet j), $\gamma_{(j)}$: Dirichlet prior vector for π

TABLE I: Instances of Shopping Arcade Actors

Class	Actor Instance	Behavioral message example

Participant	An elderly person, an adult, a teen, and a child	child.attend()
		child.behave(talk)
		elderly.behave(gaze)

Mobile obstacle	Cleaning cart and maintenance cart	cleaningCart.appear()
		cleaningCart.act(moveBW)
		cleaningCart.act(flashingLight)

Static obstacle	Caution signs, flash lights, siren, and fire alarms	caution.appear()
		caution.inform("Wet Floor")
		alram.inform("siren")

LRF	LRF1, LRF2, LRF3, and LRF4	LRF1.on()
		LRF1.detect(participant)

	MINE
dmax < dmax_base	iterating nodes from base isochrone checking if travel time is <= dmax
dmax = dmax_base	no change
dmax > dmax_base	extend base isochrone by border points and with list 1_hubs

Table 1:
Incremental calculation without vertex expiration

	MINEX
dmax < dmax_base	shrink base isochrone from border
dmax = dmax_base	no change
dmax > dmax_base	extend base isochrone by border points and with list 1_hubs

Table 2:
Incremental calculation with vertex expiration

Figure 4.2: Table caption formats in training dataset

4.1.2 Statistics of Evaluation Dataset

The training dataset consists of 40 research articles belonging from 20 different workshop proceedings of CEURWS. Table 4.2 provides details of volume, proceeding and number of articles selected from that volume for evaluation dataset. Evaluation dataset consists of formats from training dataset with more complexity. In evaluation dataset, body of the paper does not always start with "Abstract" section. In some of the cases the information of authors and email are on multiple lines, instead of single line. Fig. 4.4 shows the formats of inserting authors information in research articles of evaluation dataset. In the same manner, information related to sections, headings, figures and tables also consists of multiple formats from the training dataset that are most difficult to cater. Fig. 4.5 shows the research articles in evaluation dataset, in which the body of research article

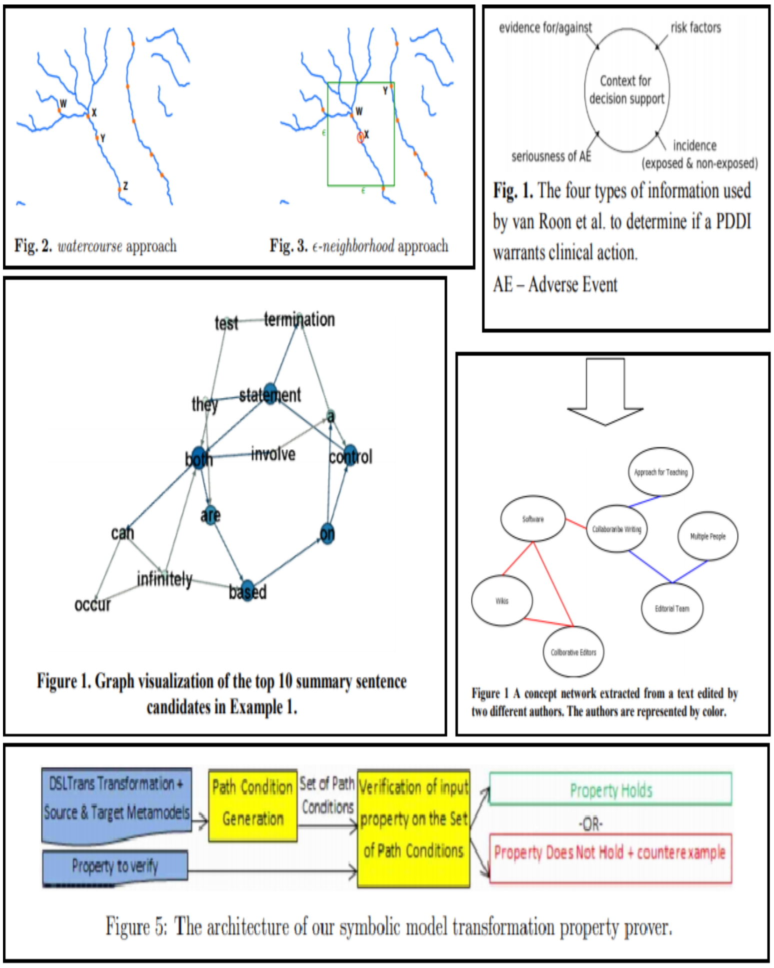


Fig. 2. *watercourse* approach

Fig. 3. *ϵ-neighborhood* approach

Fig. 1. The four types of information used by van Roon et al. to determine if a PDDI warrants clinical action.
AE – Adverse Event

Figure 1. Graph visualization of the top 10 summary sentence candidates in Example 1.

Figure 1 A concept network extracted from a text edited by two different authors. The authors are represented by color.

Figure 5: The architecture of our symbolic model transformation property prover.

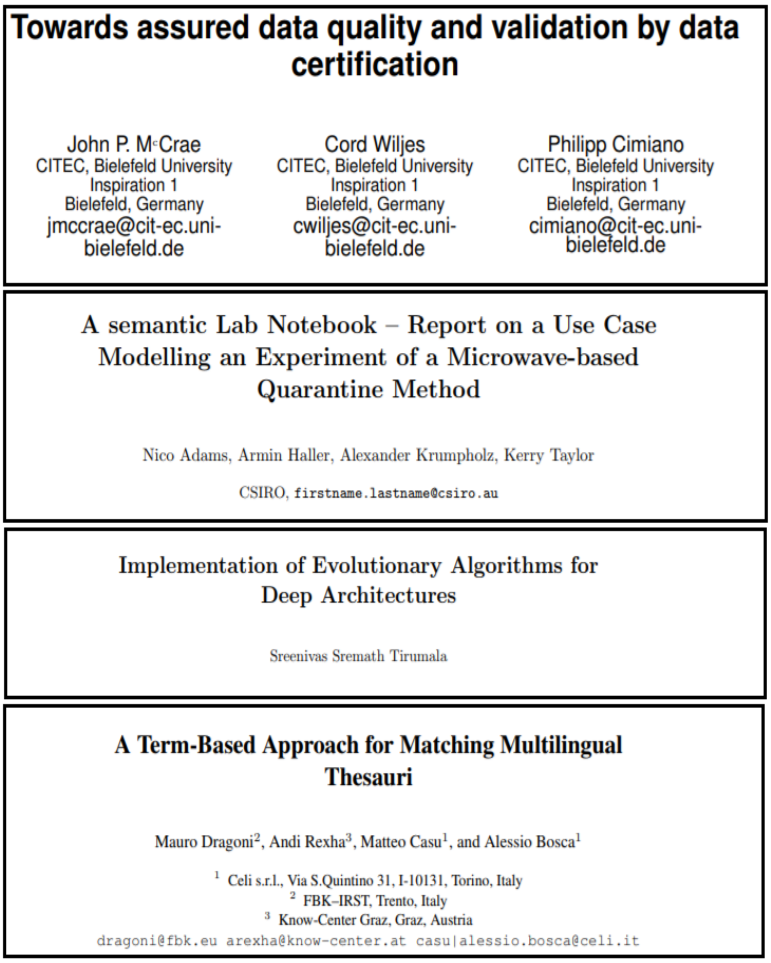
Figure 4.3: Figure caption formats in training dataset

starts with the section other than "Abstract". Some articles only consist of one section (Abstract, Acknowledgement e.tc.). All these formats and features make the evaluation of generalized rules much more accurate, making this dataset much more diverse.

The second dataset used for evaluation purpose consists of 120 research articles, selected from most recent CEURWS proceedings. This dataset contains research articles from 10 different CEUR proceedings. Each volume belongs to different field of study. Table 4.3 provides details of dataset. These papers were selected randomly for evaluation of generalized rules. These articles were selected randomly from the most recent proceedings. Generalized rules were developed from the proceedings before 2015; This dataset consists of research articles from proceedings of 2018, the most recent research articles that we could gather.

Table 4.2: ESWC Evaluation Dataset Stats

Volume	Name	No. of Research Articles
Vol-1006	New Generation Enterprise and Business Innovation Systems 2013	1
Vol-1044	Natural Language Processing and Automated Reasoning 2013	1
Vol-1116	Linked Science 2013	2
Vol-1184	Linked Data on the Web 2014	1
Vol-1215	Linked Data Quality 2014	1
Vol-1303	Ordering and Reasoning 2014	1
Vol-1313	Grundlagen von Datenbanken 2014	3
Vol-1315	Artificial Intelligence and Cognition 2014	3
Vol-1317	Ontology Matching 2014	4
Vol-1319	Model-Driven Robot Software Engineering 2014	1
Vol-1320	Semantic Web Applications and Tools for Life Sciences 2014	1
Vol-1405	Location-Aware Recommendations 2015	4
Vol-1504	UAI 2015 Workshop on Advances in Causal Inference	1
Vol-1531	Doctoral Symposium at MoDELS 2015	3
Vol-1554	MoDELS 2015 - Posters and Demos	3
Vol-1558	EDBT/ICDT Workshops 2016	2
Vol-1559	Software Engineering Workshops 2016	3
Vol-1560	Executable Modeling 2015	1
Vol-1565	Bayesian Modeling Applications Workshop 2015	1
Vol-1567	Bibliometric-enhanced Information Retrieval 2016	1



Towards assured data quality and validation by data certification

John P. McCrae
CITEC, Bielefeld University
Inspiration 1
Bielefeld, Germany
jmccrae@cit-ec.uni-bielefeld.de

Cord Wiljes
CITEC, Bielefeld University
Inspiration 1
Bielefeld, Germany
cwiljes@cit-ec.uni-bielefeld.de

Philipp Cimiano
CITEC, Bielefeld University
Inspiration 1
Bielefeld, Germany
cimiano@cit-ec.uni-bielefeld.de

A semantic Lab Notebook – Report on a Use Case Modelling an Experiment of a Microwave-based Quarantine Method

Nico Adams, Armin Haller, Alexander Krumpholz, Kerry Taylor

CSIRO, `firstname.lastname@csiro.au`

Implementation of Evolutionary Algorithms for Deep Architectures

Sreenivas Sremath Tirumala

A Term-Based Approach for Matching Multilingual Thesauri

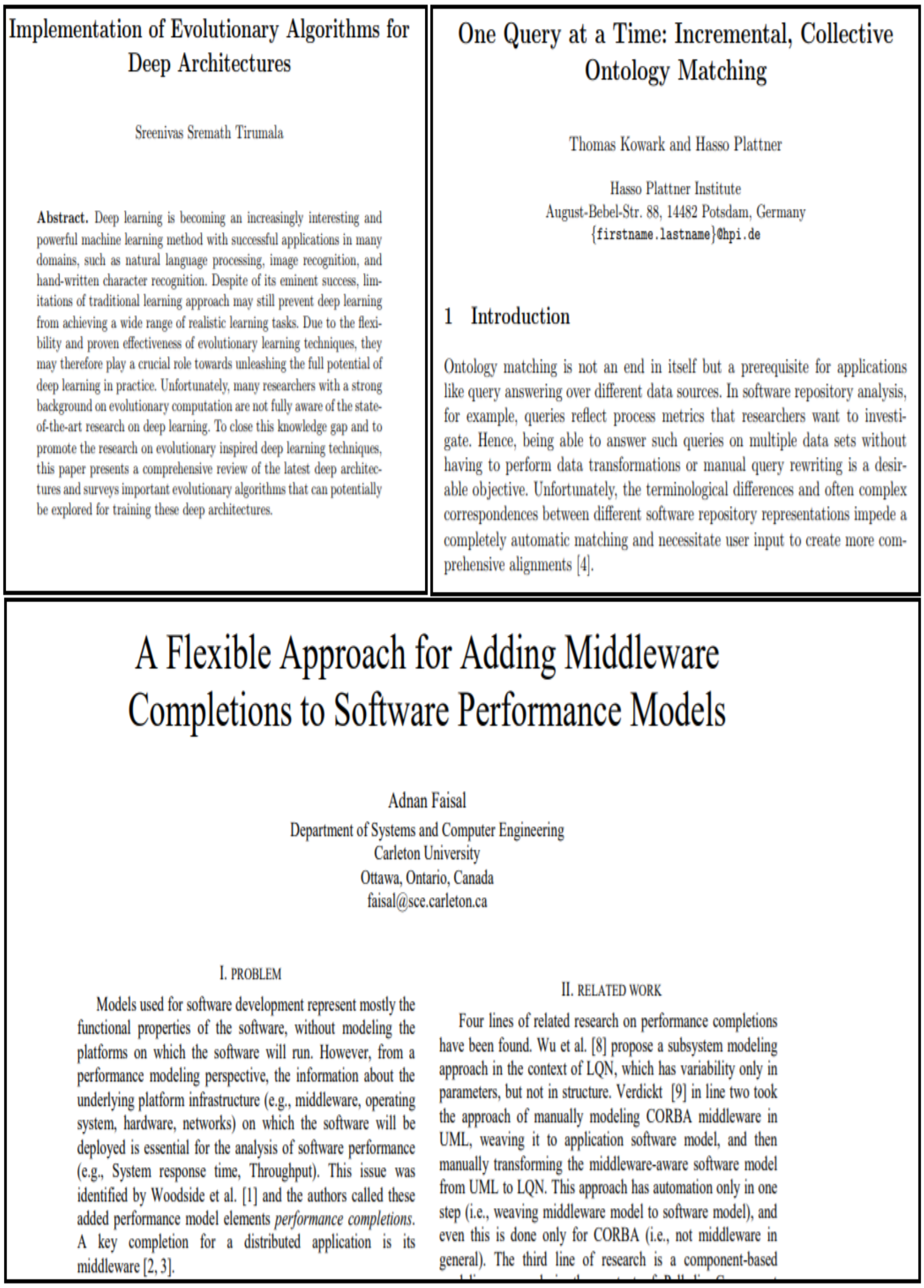
Mauro Dragoni[2], Andi Rexha[3], Matteo Casu[1], and Alessio Bosca[1]

[1] Celi s.r.l., Via S.Quintino 31, I-10131, Torino, Italy
[2] FBK–IRST, Trento, Italy
[3] Know-Center Graz, Graz, Austria
`dragoni@fbk.eu arexha@know-center.at casu|alessio.bosca@celi.it`

Figure 4.4: Author information formats in Evalutaion dataset

4.2 Statistics of PDF To XML Conversion

PDFX [6] converts the PDF document into XML format along with the tagging of PDF document. When this research thesis was started, unfortunately PDFX was unavailable, due to which we were unable to use it in our thesis for converting our PDF documents into XML. We used an online conversion tool [28] to convert the PDF document into XML format. This tool converts the PDF document into XML format and extract its geometrical features. This XML tool converts the PDF document into XML, providing information related to the geomatical features of a research article. Fig. 4.6 shows the converted XML of PDF document. As already discussed, XML tool provides some extra information regarding PDF document. This extra information includes: (1) Page number, (2) Top, (3) Left, (4) Height, and (5) Width.



Implementation of Evolutionary Algorithms for Deep Architectures

Sreenivas Sremath Tirumala

Abstract. Deep learning is becoming an increasingly interesting and powerful machine learning method with successful applications in many domains, such as natural language processing, image recognition, and hand-written character recognition. Despite of its eminent success, limitations of traditional learning approach may still prevent deep learning from achieving a wide range of realistic learning tasks. Due to the flexibility and proven effectiveness of evolutionary learning techniques, they may therefore play a crucial role towards unleashing the full potential of deep learning in practice. Unfortunately, many researchers with a strong background on evolutionary computation are not fully aware of the state-of-the-art research on deep learning. To close this knowledge gap and to promote the research on evolutionary inspired deep learning techniques, this paper presents a comprehensive review of the latest deep architectures and surveys important evolutionary algorithms that can potentially be explored for training these deep architectures.

One Query at a Time: Incremental, Collective Ontology Matching

Thomas Kowark and Hasso Plattner

Hasso Plattner Institute
August-Bebel-Str. 88, 14482 Potsdam, Germany
{firstname.lastname}@hpi.de

1 Introduction

Ontology matching is not an end in itself but a prerequisite for applications like query answering over different data sources. In software repository analysis, for example, queries reflect process metrics that researchers want to investigate. Hence, being able to answer such queries on multiple data sets without having to perform data transformations or manual query rewriting is a desirable objective. Unfortunately, the terminological differences and often complex correspondences between different software repository representations impede a completely automatic matching and necessitate user input to create more comprehensive alignments [4].

A Flexible Approach for Adding Middleware Completions to Software Performance Models

Adnan Faisal
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
faisal@sce.carleton.ca

I. PROBLEM

Models used for software development represent mostly the functional properties of the software, without modeling the platforms on which the software will run. However, from a performance modeling perspective, the information about the underlying platform infrastructure (e.g., middleware, operating system, hardware, networks) on which the software will be deployed is essential for the analysis of software performance (e.g., System response time, Throughput). This issue was identified by Woodside et al. [1] and the authors called these added performance model elements *performance completions*. A key completion for a distributed application is its middleware [2, 3].

II. RELATED WORK

Four lines of related research on performance completions have been found. Wu et al. [8] propose a subsystem modeling approach in the context of LQN, which has variability only in parameters, but not in structure. Verdickt [9] in line two took the approach of manually modeling CORBA middleware in UML, weaving it to application software model, and then manually transforming the middleware-aware software model from UML to LQN. This approach has automation only in one step (i.e., weaving middleware model to software model), and even this is done only for CORBA (i.e., not middleware in general). The third line of research is a component-based

Figure 4.5: Figure caption formats in Evalutaion dataset

A document consists of multiple page/s. Each attribute in document calculated according to the page i.e values of each attribute is calculated at page level, not document level. Top and left attribute refers to the starting point of certain text, as starting position on page can be calculated from top and left distance of text from the page. Height and width refers the height and width of text inside the text. The selected tool calculates the discussed attributes and combines these

Table 4.3: In-house Evaluation Dataset Stats

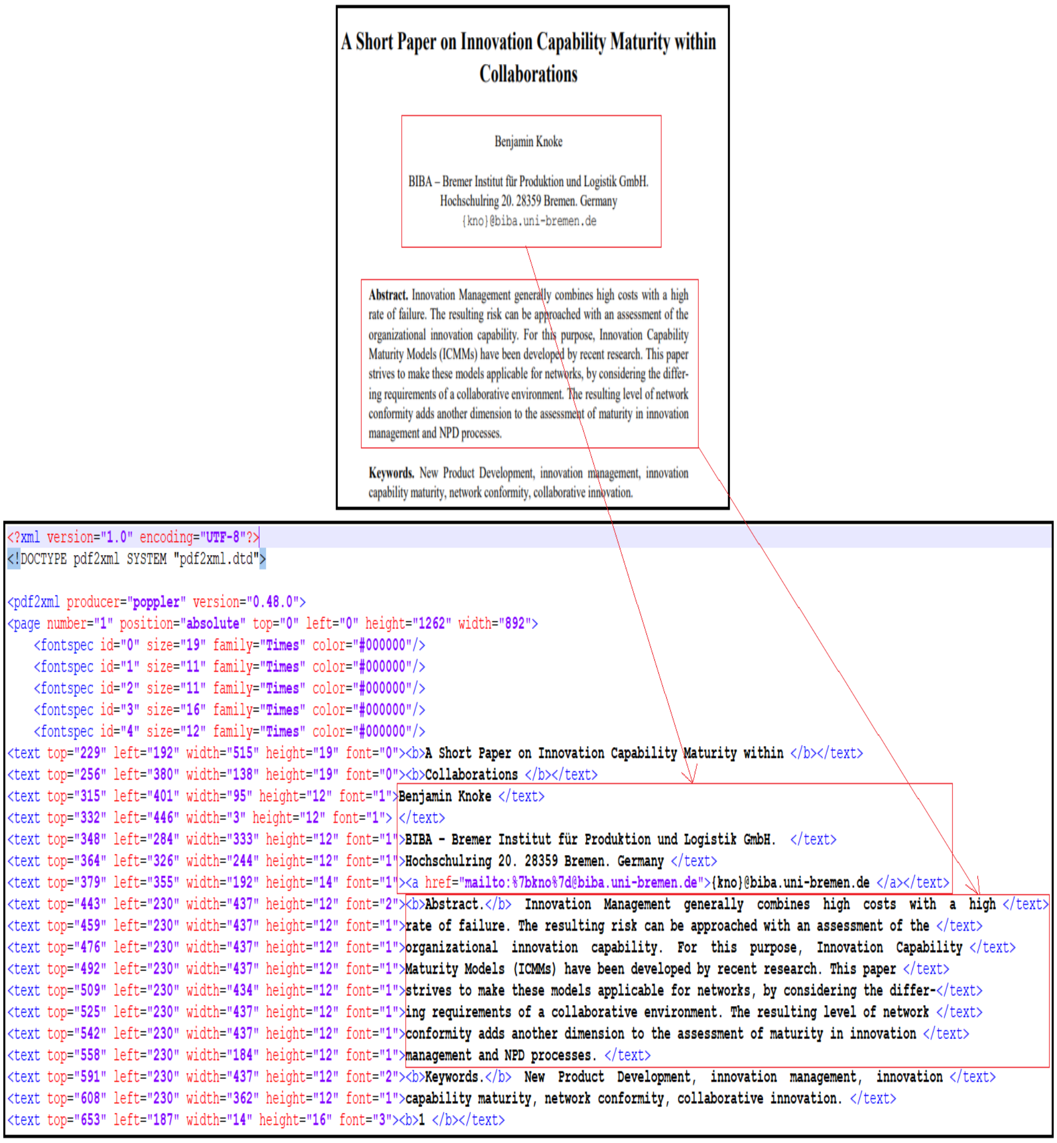
Volume	Name	No. of Research Articles
Vol-2246	Forum Media Technology and All Around Audio Symposium	09
Vol-2254	Multidisciplinary Symposium on Computer Science and ICT	10
Vol-2246	Games-Human Interaction	08
Vol-2246	Practicing Open Enterprise Modelling within OMiLAB	08
Vol-2255	Informatics & Data-Driven Medicine	23
Vol-2244	Natural Language for Artificial Intelligence	14
Vol-2211	Description Logics	18
Vol-2130	Emoji Understanding and Applications in Social Media	08
Vol-2219	Probabilistic Logic Programming	08
Vol-2161	Italian Symposium on Advanced Database Systems	14

attributes with text, providing the complete XML tag.

4.3 Training Phase Results

4.3.1 Results of Title Extraction

Font feature of XML output is used for identification of title. All the XML documents in training dataset had title font information‘ ‘Font=0”. This heuristic was used for identification of title from the PDF document. Fig. 4.7 shows the title in the XML output. As shown in figure, all the titles in the research articles



A Short Paper on Innovation Capability Maturity within Collaborations

Benjamin Knoke

BIBA – Bremer Institut für Produktion und Logistik GmbH.
Hochschulring 20. 28359 Bremen. Germany
{kno}@biba.uni-bremen.de

Abstract. Innovation Management generally combines high costs with a high rate of failure. The resulting risk can be approached with an assessment of the organizational innovation capability. For this purpose, Innovation Capability Maturity Models (ICMMs) have been developed by recent research. This paper strives to make these models applicable for networks, by considering the differing requirements of a collaborative environment. The resulting level of network conformity adds another dimension to the assessment of maturity in innovation management and NPD processes.

Keywords. New Product Development, innovation management, innovation capability maturity, network conformity, collaborative innovation.

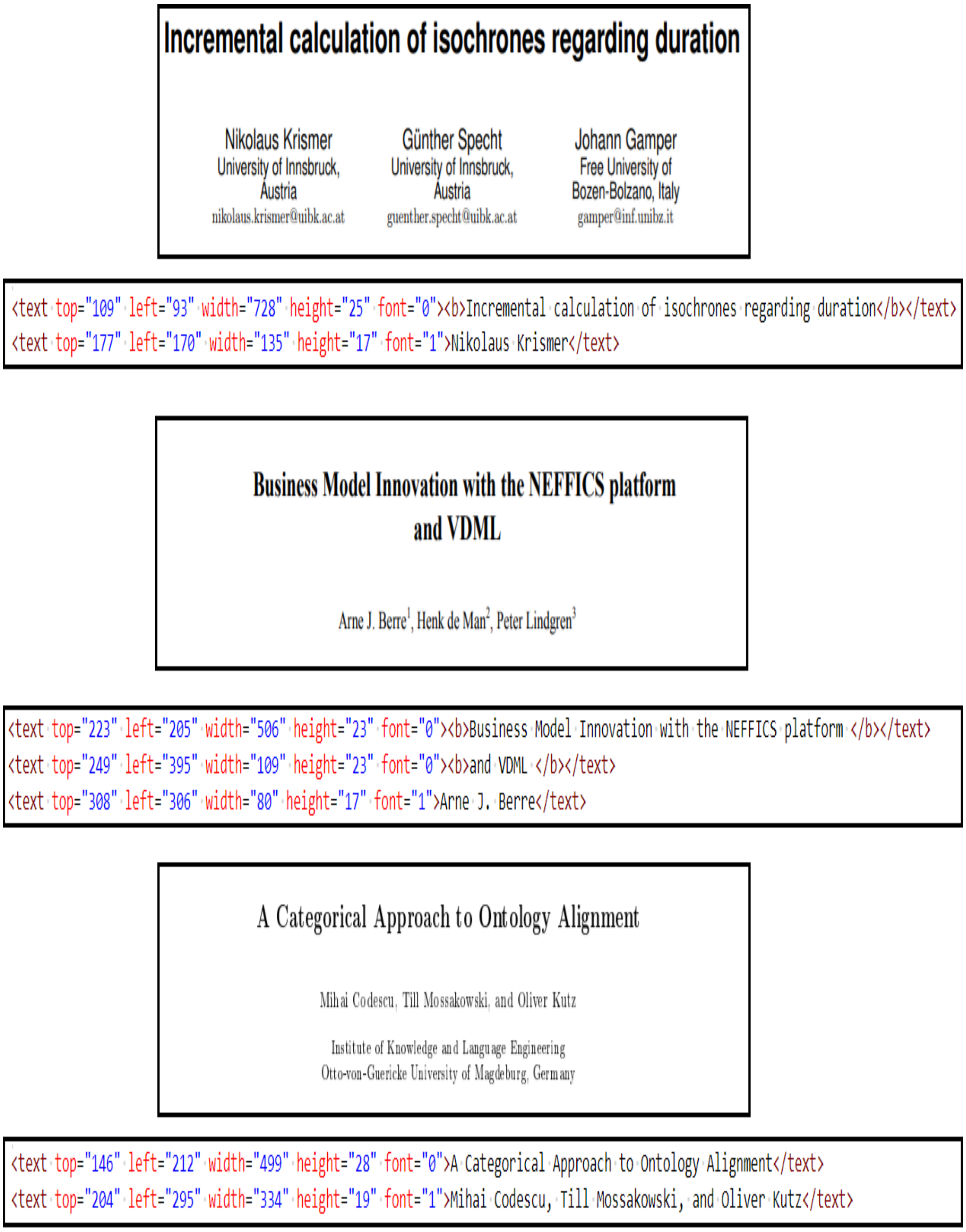
```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE pdf2xml SYSTEM "pdf2xml.dtd">

<pdf2xml producer="poppler" version="0.48.0">
<page number="1" position="absolute" top="0" left="0" height="1262" width="892">
    <fontspec id="0" size="19" family="Times" color="#000000"/>
    <fontspec id="1" size="11" family="Times" color="#000000"/>
    <fontspec id="2" size="11" family="Times" color="#000000"/>
    <fontspec id="3" size="16" family="Times" color="#000000"/>
    <fontspec id="4" size="12" family="Times" color="#000000"/>
<text top="229" left="192" width="515" height="19" font="0"><b>A Short Paper on Innovation Capability Maturity within </b></text>
<text top="256" left="380" width="138" height="19" font="0"><b>Collaborations </b></text>
<text top="315" left="401" width="95" height="12" font="1">Benjamin Knoke </text>
<text top="332" left="446" width="3" height="12" font="1"> </text>
<text top="348" left="284" width="333" height="12" font="1">BIBA - Bremer Institut für Produktion und Logistik GmbH.  </text>
<text top="364" left="326" width="244" height="12" font="1">Hochschulring 20. 28359 Bremen. Germany </text>
<text top="379" left="355" width="192" height="14" font="1"><a href="mailto:%7bkno%7d@biba.uni-bremen.de">{kno}@biba.uni-bremen.de </a></text>
<text top="443" left="230" width="437" height="12" font="2"><b>Abstract.</b>  Innovation  Management  generally  combines  high  costs  with  a  high </text>
<text top="459" left="230" width="437" height="12" font="1">rate of failure. The resulting risk can be approached with an assessment of the </text>
<text top="476" left="230" width="437" height="12" font="1">organizational  innovation  capability.  For  this  purpose,  Innovation  Capability </text>
<text top="492" left="230" width="437" height="12" font="1">Maturity Models (ICMMs) have been developed by recent research. This paper </text>
<text top="509" left="230" width="434" height="12" font="1">strives to make these models applicable for networks, by considering the differ-</text>
<text top="525" left="230" width="437" height="12" font="1">ing requirements of a collaborative environment. The resulting level of network </text>
<text top="542" left="230" width="437" height="12" font="1">conformity adds another dimension to the assessment of maturity in innovation </text>
<text top="558" left="230" width="184" height="12" font="1">management and NPD processes. </text>
<text top="591" left="230" width="437" height="12" font="2"><b>Keywords.</b>  New  Product  Development,  innovation  management,  innovation </text>
<text top="608" left="230" width="362" height="12" font="1">capability maturity, network conformity, collaborative innovation. </text>
<text top="653" left="187" width="14" height="16" font="3"><b>1 </b></text>
```

FIGURE 4.6: PDF to XML conversion output

are represented with font value “0”. Using this heuristic, we were able to get an f-score of 1 on the training dataset.

4.3.2 Results of Email Extraction

Our proposed systems work by identifying the lines from the header section, that are most likely to have email addresses. This identification is performed by a simple heuristic that checks whether a line contains “@” character or not with a set of heuristics. This heuristic is used with some set of rules to identify the email



Incremental calculation of isochrones regarding duration

Nikolaus Krismer
University of Innsbruck,
Austria
nikolaus.krismer@uibk.ac.at

Günther Specht
University of Innsbruck,
Austria
guenther.specht@uibk.ac.at

Johann Gamper
Free University of
Bozen-Bolzano, Italy
gamper@inf.unibz.it

```
<text top="109" left="93" width="728" height="25" font="0"><b>Incremental calculation of isochrones regarding duration</b></text>
<text top="177" left="170" width="135" height="17" font="1">Nikolaus Krismer</text>
```

Business Model Innovation with the NEFFICS platform
and VDML

Arne J. Berre[1], Henk de Man[2], Peter Lindgren[3]

```
<text top="223" left="205" width="506" height="23" font="0"><b>Business Model Innovation with the NEFFICS platform </b></text>
<text top="249" left="395" width="109" height="23" font="0"><b>and VDML </b></text>
<text top="308" left="306" width="80" height="17" font="1">Arne J. Berre</text>
```

A Categorical Approach to Ontology Alignment

Mihai Codescu, Till Mossakowski, and Oliver Kutz

Institute of Knowledge and Language Engineering
Otto-von-Guericke University of Magdeburg, Germany

```
<text top="146" left="212" width="499" height="28" font="0">A Categorical Approach to Ontology Alignment</text>
<text top="204" left="295" width="334" height="19" font="1">Mihai Codescu, Till Mossakowski, and Oliver Kutz</text>
```

FIGURE 4.7: Titles in XML Converted output

address lines correctly. To identify other rules, we were required to know all the possible formats of email addresses that a research article can have. We identified a variety of different formats available in training dataset. Table 4.4 summarizes all the formats that we identified in our training phase.

After the identification of all the possible formats, our system identifies the format used in the research article. All formats provided in Table 4.4 are handled by our proposed approach. Email addresses contain curly braces, comma or pipes are identified by checking brackets, comma or pipes in combination with "email line checking" heuristic. Identified email address are resolved by separating the email in two parts: the recipient part and the domain part. Recipient part is further resolved by removing the curly braces and separating the recipients according to the comma or pipe. Each separated recipient is combined with the domain part providing the actual email addresses.

Some of the research articles provide email addresses, that are dependent on the name of authors, such as "first.middle.lastname@cs.ox.ac.uk". For email addresses like these, our system extracts the authors using Author part extractor and adds the author information into email address to identify the correct email address. In some cases, a research article contains multiple email formats. Our system identifies these formats using the heuristics discussed for email identification. After identification of each email format, these formats are resolved accordingly.

TABLE 4.4: Identified email formats in training dataset

Email
hugo.alatrista@univ-nc.nc
beck,dao,eiter,fink@kr.tuwien.ac.at
mbrochhausen@uams.edu,jschneider@pobox.com,malone@pharmacy.arizona.ed
Email: yonatan.schreiber@cubrc.org
first.middle.lastname@cs.ox.ac.uk
fhilken—gogolla@informatik.uni-bremen.de
Email:xwa,aru,yla@hib.no
atzmueller, kibanov, hayat@cs.uni-kassel.de matthias.trojahn@volkswagen.de

4.3.3 Results of Affiliation and Country Extraction

Our system uses a set of keywords to identify the start of affiliation line, along with some heuristics to find the end of affiliation part. Table 4.5 provides the starting and ending conditions of the affiliation. Start of the affiliation is detected by checking if any of the keywords provided in Table 4.5 is found. All lines after this line are added as one affiliation, until a line containing a country or email is found. The line containing country or email information is considered as ending point of the affiliation. After identification of all the affiliations, extra information that is not part of affiliation is removed by the cleaner and each affiliation is separated and provide as output.

After the extraction of affiliation, country is extracted from the affiliation part. A predefined list of 195 countries in the world, along with abbreviations of some famous countries are used for identification of country. A line containing a country in the header section is considered as a stopping point for the affiliation extractor and identifies country for country extractor. An extra check on country identification is performed because of the use of abbreviations. This extra check ensures that the identified abbreviation is an actual country and not part of a word. Also, identification of country may provide redundant result and output duplicate results. For this purpose, our system extracts the countries and remove the duplicates to ensure a single result for each extracted country.

Table 4.5: Affiliation identification heuristics

Position	Condition
Start	Laboratory, Intelligence, Institut, Division, Faculty, University, College, Universit, Educational, Department, School, Centre, Institute, Group, Universität, Escuela, Engineering, Dept, St., Research
End	Line containing country name or email address

4.3.4 Results of Author Extraction

Author extraction is also performed from the header section of the research article. Author extraction is dependent on three metadata properties: (1) Email, (2) Affiliation, (3) Country, and (4) Title. Output from these respective metadata properties are gathered and removed from the header section, the remaining information left is considered as information of the author. This information of the author is then cleaned and extra characters that are not part of author are removed. Each author of the paper is then separately identified and passed as output of the author extraction unit.

4.3.5 Results of Figure Caption Extraction

Figure caption extraction combines the top and page attributes of XML document with the text features. Critical analysis of the figures revealed that there is always a small amount of gap between a figure caption and body of research article as shown in Fig. 4.8. A manual analysis of all the figures from the XML documents revealed that an average gap between figure caption and body of the document is "26" We used this threshold value with the regular expressions provide in Table 4.6, for creating generalized rule for figure caption extraction.

Top property from the XML document is used to identify the gap between the two sub sequent XML tags. If the top value is greater than threshold, then it is considered that figure caption has ended. On the contrary, if the gap between two subsequent lines is less than 10 then it is considered as the line of figure captions. These properties along with the regular expressions given in Table 4.6 are applied for extraction of figures. Each regular expressions identify the start of the figure section by finding any of the following keywords: (1) Fig, (2) FIG, (3) FIGURE, and (4) Figure. After identification of "fig" keywords, the format of figure is detected, whether it is numeric with dot, numeric with colon, roman number with dot, and roman number with colon. These generalized rules help in identification of figures caption with much more accuracy and high precision.

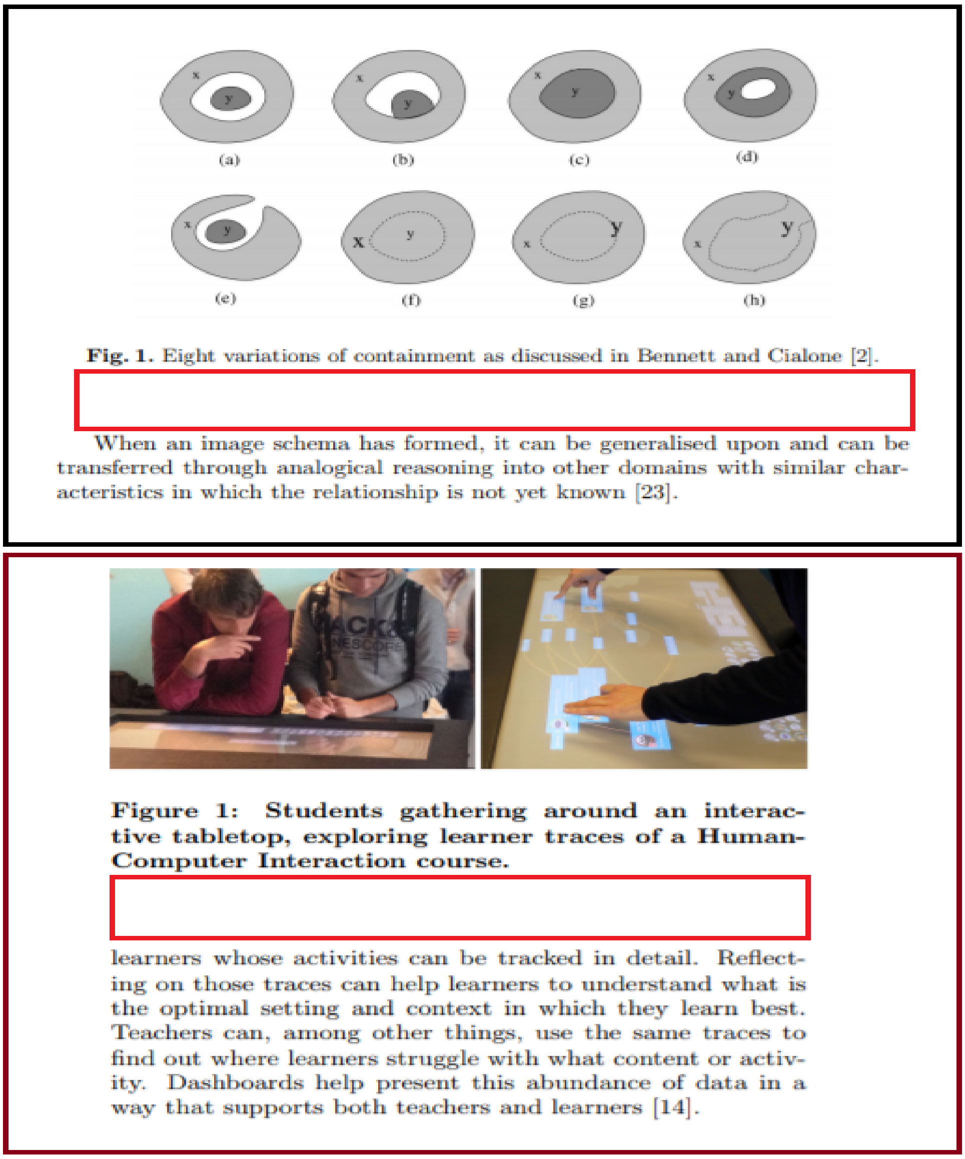


Fig. 1. Eight variations of containment as discussed in Bennett and Cialone [2].

When an image schema has formed, it can be generalised upon and can be transferred through analogical reasoning into other domains with similar characteristics in which the relationship is not yet known [23].

Figure 1: Students gathering around an interactive tabletop, exploring learner traces of a Human-Computer Interaction course.

learners whose activities can be tracked in detail. Reflecting on those traces can help learners to understand what is the optimal setting and context in which they learn best. Teachers can, among other things, use the same traces to find out where learners struggle with what content or activity. Dashboards help present this abundance of data in a way that supports both teachers and learners [14].

Figure 4.8: Gaps between figure captions and document body

4.3.6 Results of Table Caption Extraction

Table extraction utilize top and font property of xml along with some textual heuristics to extract table captions. Analysis of table captions revealed that table captions contain gap at top or bottom of them which can be used to identify generalized rules for extraction of table captions. All the documents output was critically analysed and the gap between table caption and body of document came out to be "25". We used this threshold value along with the regular expressions provided in Table 4.7 to extract table captions. Using "font" property while

Table 4.6: Regular expressions for identification of Figure caption

No.	Regex
1	(Fig|FIG|FIGURE|Figure)\\s[0-9]+[\\s].*
2	(Fig|FIG|FIGURE|Figure)(.\\s|\\s)[0-9]+[.:]+.*
3	(Fig|FIG|FIGURE|Figure)(.\\s|\\s)[IVX]+[.:]+.*
4	(Fig|FIG|FIGURE|Figure)(.\\s|\\s)[IVX]+[\\s]+.*

extracting table captions adds up and extra filter to ensure correctness of the extracted table caption.

Table 4.7: Regular expressions for identification of Table caption

No.	Regex
1	(Table|TABLE)(.\\s|\\s)[0-9]+[\\s]+.*
2	(Table|TABLE)(.\\s|\\s)[0-9]+[.:]+.*
3	(Table|TABLE)(.\\s|\\s)[0-9IVX]+[.:]+.*
4	(Table|TABLE)(.\\s|\\s)[0-9IVX]+[\\s]+.*

4.3.7 Results of Heading Extraction

In this approach, we extracted level 1 headings of the research article. Three geometrical properties: page, top and left, along with regular expressions provided in Table 4.8 are used to extract the headings. The text of PDF document is combined based on page number, top and left. Combining the XML generated output using this format helps in identifying single column and multi column views, as well as identification of the padding becomes easier. After combining the text of XML tags based on page number, top and left properties, we apply the regular expressions to identify the Heading.

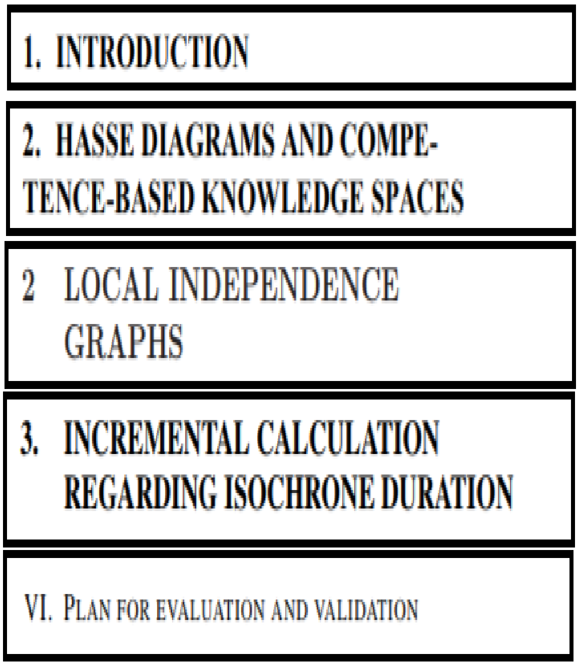
There are variety of formats available in training dataset as sown in Fig. 4.9. The main purpose of combining the text of XML document together was to combine multi-line headings and make them available in single line. The purpose for using the left property along with the page number and top is that it helps in identification of heading in the research articles having two column views. Left property ensures that headings in both columns are identified separately and do not interfere with each other. We came across a case where the top and page number of two headings was some, which resulted in less headings than the actual. Adding the left attribute remove that anomaly and added an extra amount of check to provide high accuracy.

TABLE 4.8: Regular expressions for identification of Table caption

<table>
<tr><th>No.</th><th>Regex</th></tr>
<tr><td>1</td><td><code>[\\d]+(\\.|\\s|)(\\s+|)[A-Z].*</code></td></tr>
<tr><td>2</td><td><code>[IVX]+(\\.|\\s)[\\s]+[A-Z].*</code></td></tr>
</table>

4.3.8 Results of Funding Agency Extraction

The extraction of funding agency is performed by first identifying the "Acknowledgment" section. The problem with this approach was that not all papers had "Acknowledgment" section. We applied multiple heuristics to find the "Acknowledgment" section, from which we extracted the funding agency. Table 4.9 shows the starting and ending keywords used for identification of funding agency. Text between the starting and ending keywords is considered as funding agency. The funding agency cleaner performs cleaning of the extracted funding agency and removes words such as "the", "from" etc.



1. INTRODUCTION

2. HASSE DIAGRAMS AND COMPETENCE-BASED KNOWLEDGE SPACES

2 LOCAL INDEPENDENCE GRAPHS

3. INCREMENTAL CALCULATION REGARDING ISOCHRONE DURATION

VI. PLAN FOR EVALUATION AND VALIDATION

FIGURE 4.9: Formats for heading level 1 of research articles from training dataset

TABLE 4.9: Regular expressions for identification of Funding Agency

No.	Regex
1	supported by, financial sup-port, supported, in part, by, funding of, support from, supported in part, funded by, funding from, fellowship from, financial support of, financial support of, financial support from
2	under grant, fig, within,), under, grant, .

4.4 Extraction Phase Results

Extraction phase consists of the system in which we implemented all the identified generalized rules. This phase starts its working by converting the PDF document in the form of XML using online tool [28]. This converted XML document is passed from series of extractors to extract metadata information. These extractors are build upon the rules and heuristics identified in training phase. Our proposed system extracts all the extracted metadata properties and combines them together, generating a single file in the form of text which includes the extracted information.

4.5 Evalutation Phase Results

Evaluation phase consists of evaluating our approach with CERMINE and GROBID. Most of the past work performed in this field has been compared by these two tools, making them as a benchmark result extractor. Our approach performs well in all the cases. The evaluation parameters chosen for our research includes precision, recall and f-measure. Precision refers to the positive results that are gathered by a system.

(1) Actual results, (2) Retrieved results, and (3) Relevant results for each research article is calculated, which is used to find precision, recall and f-measure. The number of any metadata property in research article refers to as actual results. Retrieved and relevant results are calculated from the output generated according to the output generated by system. Number of results generated by system are referred to as retrieved results, whereas the results that match with the actual results are considered as relevant results. Equation 4.1, 4.2 and 4.3 provides formulas for calculating precision, recall and f-measure. Precision refers to the fraction of relevant results among the total results retrieved by the system. Recall on the other hand calculates fraction of relevant results extracted by the system with retrieved results. F-score considers both precision and recall for calculation of the results and calculate harmonic average of precision and recall, giving an equal amount of importance to both precision and recall. Table 4.10 shows the capabilities of our approach with Riaz *et al.* approach, CERMINE and GROBID.

$$Precision = \frac{Relevantresults}{Retrievedresults}, \tag{4.1}$$

$$Recall = \frac{Relevantresults}{ActualResults}, \tag{4.2}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{4.3}$$

TABLE 4.10: Comparison of metadata properties extraction

Metadata Property	Proposed Architecture	Riaz *et al.* Approach	CERMINE	GROBID
Title	✓	X	✓	✓
Author	✓	✓	✓	✓
Email	✓	✓	✓	✓
Affiliation	✓	✓	✓	✓
Country	✓	✓	✓	✓
Table caption	✓	✓	X	✓
Figure Caption	✓	✓	X	✓
Headings	✓	✓	✓	✓
Funding Agency	✓	✓	X	X

4.5.1 Results of Training Dataset

[!htb] Evaluation of training dataset revelaed that our approach out performs Riaz *et al.* approach, CERMINE, and GROBID. We used 45 papers of training dataset to developed generalized rules for metadata extraction. Table 4.11 shows the results of our approach on training dataset. Our proposed system performed well on training dataset and achieved very high precision, recall and f-measure. We achieved an over all f-score of 0.93 on our training dataset.

We were unable to achieve an f-score of 1, even though we used all the training dataset for development of our generalized rules. Some of the formats that were present in the research articles required rules that were very specific. Adding these rules make our generic rules specific and affected the precision and recall for extraction of other metadata properties. Fig. 4.10 shows an example of table, in which the threshold distance is “16”, which is the common line distance between most of the papers. If we used this threshold, while extraction of PDF documents, all the lines satisfying the pattern would have been extracted resulting in wrong results and low precision and recall. In the same manner, there were some figures and headings which had less threshold value then the average threshold value that we calculated for figures and headings. These were the reasons due to which we were unable to get 100% result accuracy.

Table 4.11: Metadata Extraction results using generalized rules for Training dataset

Metadata Property	Precision	Recall	F-Measure
Title	1.00	1.00	1.00
Author	0.96	0.96	0.95
Email	0.99	0.99	0.99
Affiliation	1.00	0.99	1.00
Country	0.96	0.92	0.93
Table caption	0.98	0.99	0.98
Figure Caption	0.95	0.93	0.94
Headings	0.80	0.82	0.79
Funding Agency	0.90	0.81	0.82
	0.95	**0.93**	**0.93**

Table 4.12 & 4.13 shows the results of the information extracted from training dataset. Our approach out performs CERMINE and GROBID, providing much higher results for the training dataset. Table 4.14 shows the comparison of CERMINE, GROBID and our approach based on f-measures.

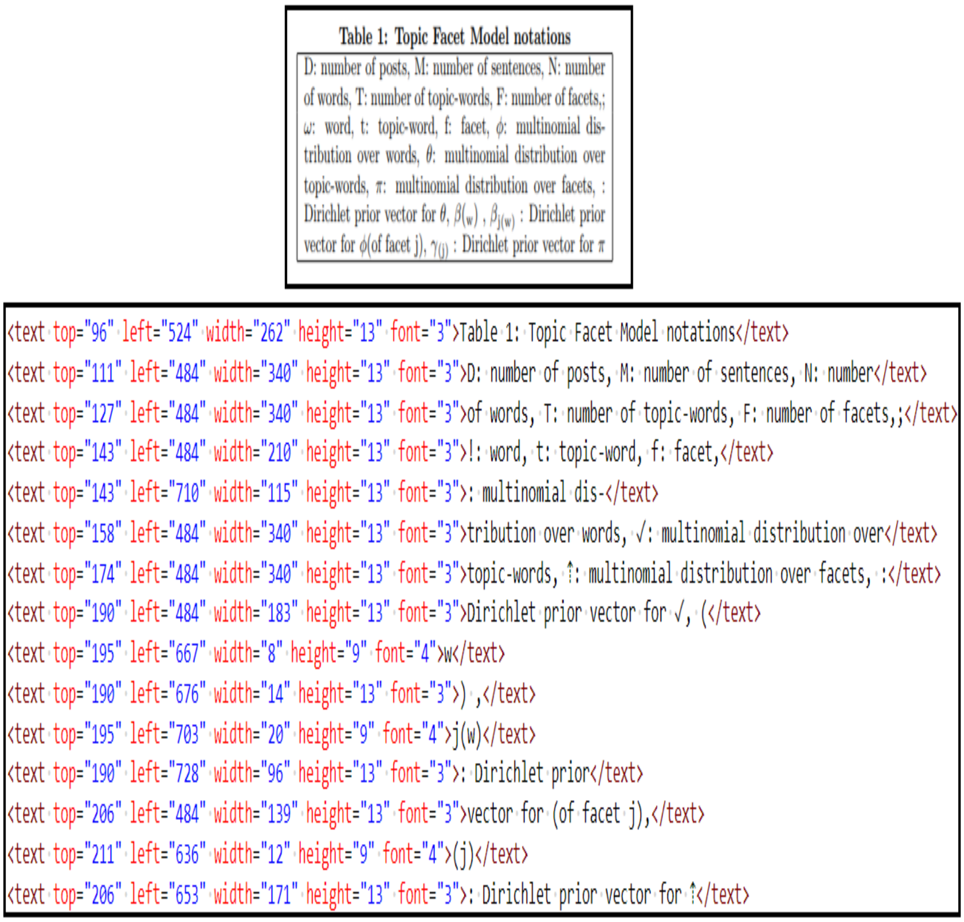


Table 1: Topic Facet Model notations

D: number of posts, M: number of sentences, N: number of words, T: number of topic-words, F: number of facets,; ω: word, t: topic-word, f: facet, ϕ: multinomial distribution over words, θ: multinomial distribution over topic-words, π: multinomial distribution over facets, : Dirichlet prior vector for θ, $\beta(_w)$, $\beta_{j(w)}$: Dirichlet prior vector for ϕ(of facet j), $\gamma_{(j)}$: Dirichlet prior vector for π

```
<text top="96" left="524" width="262" height="13" font="3">Table 1: Topic Facet Model notations</text>
<text top="111" left="484" width="340" height="13" font="3">D: number of posts, M: number of sentences, N: number</text>
<text top="127" left="484" width="340" height="13" font="3">of words, T: number of topic-words, F: number of facets,;</text>
<text top="143" left="484" width="210" height="13" font="3">!: word, t: topic-word, f: facet,</text>
<text top="143" left="710" width="115" height="13" font="3">: multinomial dis-</text>
<text top="158" left="484" width="340" height="13" font="3">tribution over words, √: multinomial distribution over</text>
<text top="174" left="484" width="340" height="13" font="3">topic-words, ⇡: multinomial distribution over facets, :</text>
<text top="190" left="484" width="183" height="13" font="3">Dirichlet prior vector for √, (</text>
<text top="195" left="667" width="8" height="9" font="4">w</text>
<text top="190" left="676" width="14" height="13" font="3">) ,</text>
<text top="195" left="703" width="20" height="9" font="4">j(w)</text>
<text top="190" left="728" width="96" height="13" font="3">: Dirichlet prior</text>
<text top="206" left="484" width="139" height="13" font="3">vector for (of facet j),</text>
<text top="211" left="636" width="12" height="9" font="4">(j)</text>
<text top="206" left="653" width="171" height="13" font="3">: Dirichlet prior vector for ⇡</text>
```

Figure 4.10: Example of research article having very specific properties

4.5.2 Results of Evaluation Datasets

As already discussed, our evaluation dataset is composed of two datasets: (1) ESWC Evaluation dataset, (2) A dataset prepared in-house, that consists of 120 research articles. Our proposed approach performed better than the other approaches, achieving an f-score of 0.85 and 0.87 on evaluation dataset and in-house dataset respectively.

Table 4.15 and Table 4.16 shows the results produced by our proposed approach on evaluation datasets. We were able to achieve high accuracy for extracted parameters. Our heuristics for funding agency, headings, authors, email and affiliation achieved less accuracy then the other heuristics. Heuristics for email and affiliation did not worked as expected because we were unable to find the root element for the start/ end of these metadata. Author extraction performed less because of

Table 4.12: CERMINE results for Metadata extraction on training dataset

Metadata Property	Precision	Recall	F-Measure
Title	1.00	1.00	1.00
Author	0.71	0.70	0.70
Email	0.73	0.62	0.65
Affiliation	0.93	0.89	0.91
Country	0.83	0.80	0.80
Table caption	0.71	0.72	0.70
	0.82	**0.79**	**0.79**

Table 4.13: GROBID results for Metadata extraction on training dataset

Metadata Property	Precision	Recall	F-Measure
Title	1.00	1.00	1.00
Author	0.99	0.98	0.98
Email	0.80	0.70	0.73
Country	0.94	0.91	0.92
Affiliation	0.91	0.93	0.90
Table	0.74	0.73	0.73
Figure	0.79	0.76	0.76
Heading	0.88	0.92	0.88
	0.88	**0.87**	**0.86**

Table 4.14: Comparison of Proposed approach GROBID, and CERMINE

Metadata Property	Generalized Rules	CERMINE	GROBID
Title	1.00	1.00	1.00
Author	0.95	0.70	0.98
Email	0.99	0.65	0.73
Country	1.00	0.91	0.92
Affiliation	0.93	0.80	0.90
Table	0.98	0.00	0.73
Figure	0.94	0.00	0.76
Heading	0.79	0.70	0.88
	0.93	**0.79**	**0.86**

having dependency on affiliation and email extraction. Our heuristics for headings achieved high recall, however less precision because of variety of different formats. These variety of formats are handled by multiple heuristics resulting in some wrong headings. The heuristics for funding agency extracted the funding agency achieved much less results than the other heuristics. On detailed analysis of the funding agency results revealed that our proposed heuristic extracts the funding agency correctly, however our cleaner fails to separate or remove extra information from the extracted text. This inefficiency of funding agency cleaner results in wrong identification of funding agency, providing less precision and recall.

Table 4.17 and Table 4.18 shows results of GROBID for extraction of metadata. GROBID results reveal that it performs better extraction of author, affiliation and headings than our proposed approach, however our approach out performs it in all the other metadata extraction properties.

Table 4.15: Result of proposed approach on ESWC evaluation dataset

Metadata Property	Precision	Recall	F-Measure
Title	0.95	0.95	0.975
Author	0.77	0.84	0.792890304
Email	0.81	0.81	0.806410256
Country	0.94	0.95	0.941666667
Affiliation	0.80	0.84	0.815833333
Table	0.88	0.92	0.89
Figure	0.93	0.95	0.936338384
Headings	0.82	0.86	0.811910194
Funding Agency	0.73	0.66	0.68
	0.85	**0.86**	**0.85**

Table 4.19 and Table 4.20 shows results of CERMINE for extraction of metadata from evaluation dataset. CERMINE was only able to identify the title and country with high f-score. Apart from these two other metadata properties were not extracted with higher f-score. It can be seen in Table 4.18 that results for author, email, affiliation and headings are much less as compared to results computed using our approach and GROBID. The extraction rate on evaluation dataset prepared in-house is comparatively better than the ESWC evaluation dataset. The extraction of metadata properties almost reaches 87% on the evaluation dataset prepared in-house.

Table 4.21 compares the results produced by our proposed approach , GROBID and CERMINE. The over all comparison shows that, our approach outperforms CERMINE and GROBID achieving an f-score of 0.85 on extraction of all the metadata properties. GROBID extracts heading and authors with much higher f-score

TABLE 4.16: Results of proposed approach on dataset prepared in-house

Metadata Property	Precision	Recall	F-Measure
Title	0.94	0.94	0.94
Author	0.76	0.84	0.78
Email	0.84	0.82	0.82
Country	0.96	0.97	0.96
Affiliation	0.84	0.79	0.80
Headings	0.74	0.79	0.74
Figure	0.90	0.91	0.90
Table	0.95	0.95	0.95
Funding Agency	0.93	0.90	0.91
	0.87	**0.88**	**0.87**

than our proposed approach, however our approach performs better extraction of other metadata properties providing results 13% better results than GROBID. CERMINE on the other hand does not achieve high f-scores except for title and country information. Results shows that our approach performs 70% better results than CERMINE for extraction of metadata. If metadata properties that are not extracted by CERMINE and GROBID are excluded, our approach still performs better achieving high f-score. Excluding funding agency from the results shows that our approach achieve an f-score of 0.871, whereas GROBID and CERMINE achieves 0.84 and 0.56 respectively. If funding agency, tables and figures are excluded then our approach achieve an f-score of 0.85, whereas CERMINE and GROBID achieve 0.75 and 0.87 respectively.

Table 4.22 compares the results produced by our proposed approach , GROBID and CERMINE on in-house dataset. Our approach achieves an f-score of 0.87

Table 4.17: Result of GROBID on ESWC evaluation dataset

Metadata Property	Precision	Recall	F-Measure
Title	0.93	0.90	0.90
Author	0.95	0.96	0.95
Email	0.83	0.73	0.75
Country	0.91	0.91	0.91
Affiliation	0.84	0.88	0.85
Table	0.70	0.71	0.70
Figure	0.84	0.85	0.82
Headings	0.90	0.93	0.91
	0.86	**0.86**	**0.85**

followed by GROBID and CERMINE with 0.78 and 0.55 respectively. Results shows that our approach performs 11% and 58% better results than GROBID and CERMINE respectively for extraction of metadata. Excluding funding agency from the results shows that our approach achieves an f-score of 0.897, followed by GROBID and CERMINE with 0.88 and 0.62 respectively. Excluding funding agency, tables and figures from the results shows that our approach secures an f-score of 0.889, whereas CERMINE and GROBID secure 0.83 and 0.91 respectively.

Table 4.23 compares the results of our proposed approach with results of top performing approaches in Semantic Publishing Challenge [4]. Our proposed approach performs 10% better than the best performing tool. The best performing approach for the challenge was Rule/ Heuristic based approach. In this research thesis, we also developed an approach that is also heuristic based, however we try to identify

Table 4.18: Results of GROBID on dataset prepared in-house

Metadata Property	Precision	Recall	F-Measure
Title	0.93	0.93	0.93
Author	0.93	0.95	0.94
Email	0.83	0.64	0.70
Country	0.93	0.92	0.92
Affiliation	0.70	0.71	0.70
Headings	0.79	0.70	0.71
Figure	0.62	0.60	0.57
Table	0.82	0.79	0.80
	0.82	**0.78**	**0.78**

Table 4.19: Result of CERMINE on ESWC evaluation dataset

Metadata Property	Precision	Recall	F-Measure
Title	0.95	0.95	0.98
Author	0.72	0.71	0.71
Email	0.55	0.53	0.53
Country	0.90	0.89	0.89
Affiliation	0.75	0.79	0.76
Headings	0.71	0.65	0.67
	0.76	**0.75**	**0.75**

Table 4.20: Results of CERMINE on dataset prepared in-house

Metadata Property	Precision	Recall	F-Measure
Title	0.80	0.80	0.80
Author	0.86	0.88	0.86
Email	0.60	0.50	0.52
Country	0.89	0.89	0.89
Affiliation	0.74	0.73	0.72
Headings	0.80	0.64	0.70
	0.65	**0.61**	**0.62**

Table 4.21: Results of proposed approach, CERMINE and GROBID on ESWC Evaluation dataset

Metadata Property	Generalized Rules	GROBID	CERMINE
Title	0.98	0.90	0.98
Author	0.79	0.95	0.71
Email	0.81	0.75	0.53
Country	0.94	0.91	0.89
Affiliation	0.82	0.85	0.76
Table	0.89	0.70	0
Figure	0.94	0.82	0
Headings	0.81	0.91	0.67
Funding Agency	0.68	0	0
	0.85	**0.75**	**0.50**

the heuristics in such a way that our heuristics do not get dependent on data, and

Table 4.22: Results of proposed approach with GROBID and CERMINE on the dataset prepared in-house

Metadata Property	Generalized Rules	GROBID	CERMINE
Title	0.94	0.93	0.80
Author	0.78	0.94	0.86
Email	0.82	0.70	0.52
Country	0.96	0.92	0.89
Affiliation	0.80	0.70	0.72
Headings	0.74	0.71	0.70
Figure	0.90	0.57	0
Table	0.95	0.80	0
Funding Agency	0.91	0	0
	0.87	**0.70**	**0.50**

work even if dataset is changed.

Table 4.23: Results comparison of proposed approach with Top 5 approaches of Semantic Publishing Challenge 2016

Approach	Precision	Recall	F-Measure
Generalized Rules - Proposed Approach	0.85	0.86	0.85
Information Extraction from PDF Sources based on Rule-based System using Integrated Formats	0.77	0.78	0.77
An Automatic Workflow for Formalization of Scholarly Articles' Structural and Semantic Elements	0.64	0.63	0.63
Reconstructing the Logical Structure of a Scientific Publication using Machine Learning	0.59	0.60	0.59
ACM: Article Content Miner	0.41	0.43	0.42
Automatically Identify and Label Sections in Scientific Journals using Conditional Random Fields	0.39	0.43	0.39

This section provides details of the results we achieved on multiple datasets and comparison of these results with multiple approaches. We achieved an f-score of 0.93 on training dataset, and 0.85 and 0.87 on evaluation datasets. In this section we compared our approach with CERMINE and GROBID, that are considered as most authentic tools in this field of research. Most of the previous work performed in this field has been compared with the results of these tools. Comparison of our proposed approach results reveal that our approach performs 8% and 39% better than GROBID and CERMINE respectively. Result comparison of evaluation datasets also shows that our approach performs better than CERMINE and GROBID. On ESWC evaluation dataset, our approach performs 13% and 70% better than GROBID and CERMINE respectively. Results on the dataset prepared inhouse shows that our approach perform 11% better than GROBID and 58% better than CERMINE.

Chapter 5

Conclusion and Future Work

The extraction of information from the scientific PDF documents is very challenging task due to variety of formats. Multiple approaches have been proposed to extract information from scientific PDF documents. Base on comprehensive literature review of research articles in this area we identified basic classification of techniques: (1) Rule/ Heuristic based, (2) Machine Learning, (3) Hybrid. This research thesis focuses on making rule-based approach more generalize and extract information from the PDF documents using these generalized rules. Most of the rule-based approaches convert the PDF document into textual or XML format, on which various patterns and rules are applied for extraction of information. Our approach also works by converting the PDF document into XML and utilize the converted XML to extract metadata information.

In this research thesis, we propose a generalized rule-based approach that works by converting PDF document into XML format. Most of the citied techniques, applied heuristics manually, because making these heuristics require an expert attention to craft rules. We also manually identified the common patterns and heuristics from the training dataset and incorporated these heuristics with geometrical and font features of the text. This combination of textual, geometrical and font features helped in constructing generalized rules for the proposed approach. Our system can identify 9 different metadataa properties with high accuracy. These metadata properties include: (1) Title, (2) Author, (3) Affiliation, (4) Author Email, (5) Country, (6) Headings Label, (7) Table Captions, (8) Figure Captions, and (9) Funding Agency. In our proposed system, each metadata property is extracted using its own unit. Some of the metadata properties are dependent on each other, such as Author extraction is dependent on Affiliation and Email metadata. In the same manner Country is extracted from the Affiliation part, making it dependent on the Affiliation. Our system extracts and clean the extracted information by removing extra text that is not part of metadata.

We compared the result of our approach with CERMINE, GROBID and the top 5 approaches proposed in ESWC challenge [4]. Experimental results show that our approach out performs all these approaches achieving an f-score of 0.85 on evaluation dataset. Our approach performs 10& better results than the Riaz *et al.* approach that was the winner of ESWC challenge. Furthermore, our approach performs 12% and 68% better than GROBID and CERMINE respectively on evaluation dataset. We also used a dataset prepared in-house to check if were able to construct generalized rules. Our approach on unseen data achieved an f-score of 0.87 out performing CERMINE and GROBID that achieved 0.78 and 0.55 respectively.

Our future work includes addition of more geometrical and font features then what we have utilized. Our title extraction exploits the font size property of the text only, it can be further extended by incorporating additional features such as geometrical location, page number etc. Our headings, table and figures extraction rely on a threshold value that we identified by critically analysing the PDF documents. Future work includes calculation of this threshold value dynamically for each document by identifying the gaps between the lines. This calculation of threshold value can make this approach more generalized.

Bibliography

[1] A. E. Jinha, “Article 50 million: an estimate of the number of scholarly articles in existence,” *Learned Publishing*, 23(3), pp. 258–263, 2010.

[2] “CEUR Workshop”, http://ceur-ws.org/, accessed on 20 December 2018.

[3] “Extended Semantic Web Confernce”, http://2016.eswc-conferences.org/ , accessed on 30 October 2018.

[4] “Extended Semantic Web Confernce Task 2”, https://github.com/ceurws/lod/wiki/SemPub16_Task2/, accessed on 30 October 2018.

[5] “Semantic Publishing Challenge Rankings in Task 2”, https://github.com/ceurws/lod/wiki/SemPub2016/, accessed on 30 October 2018.

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

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

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

[9] H. Déjean, J-L. Meunier, "A system for converting PDF documents into structured XML format," In *Lecture Notes in Computer Science*, pp. 129–140, 2006.

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

[11] C. Ramakrishnan, A. Patnia, E. Hovy et al. , "Layout-aware text extraction from fulltext PDF of scientific articles. Source Code for Biology and Medicine," In *Source code for biology and medicine*, no. 7(1), pp. 1–7, 2012.

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

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

[14] S. Klink and T. Kieninger, "Rule-based document structure understanding with a fuzzy combination of layout and textual features," In *International Journal On Document Analysis And Recognition*, 4(1), pp. 18–26, 2001.

[15] Kieninger TG, "Table structure recognition based on robust block segmentation," In *Proceedings of SPIE - The International Society for Optical Engineering*, pp. 22–32, 1999.

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

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

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

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

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

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

[22] 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)*, pp. 1081–1085, 2015.

[23] Crystal, A., Land, P, "Metadata and Search: Global Corporate Circle," In *DCMI 2003 Workshop*, 2003.

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

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

[26] Do, H.H.N., Chandrasekaran, M.K., Cho, P.S., Kan, M.Y., “Extracting and matching authors and affiliations in scholarly documents,” In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 219–228, 2013.

[27] Kim, S., Cho, Y., Ahn, K., “Semi-automatic metadata extraction from scientific journal article for full-text XML conversion,” In *Proceedings of the International Conference on Data Mining (DMIN)*, pp. 251–255, 2014

[28] “PDF to XML Conversion tool”, accessed on 10 November 2018, https://www.freefileconvert.com/pdf-xml, accessed on 12 November 2018