

AN ONTOLOGICAL FRAMEWORK FOR INFORMATION EXTRACTION
USING FUZZY RULE-BASED SYSTEM AND WORD SENSE
DISAMBIGUATION

GOHAR ZAMAN

A thesis submitted in
fulfillment of the requirement for the award of the
Doctor of Philosophy in Information Technology

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

DECEMBER 2021

In the name of Allah, Most Gracious, Most Merciful.
I praise and thank Allah.

Special thanks to my beloved father AHMAD ZAMAN.

For dearest,
(Brothers and Sisters)

For their love, support, enthusiasm, encouragement and motivation.

For my Mentor,
Assoc. Prof. Dr. Atta-ur-Rahman

For his incredible help, patience, understanding and support.

For all postgraduate members, fellow friends and housemates.

This thesis is dedicated to all of you.

ACKNOWLEDGEMENT

In the name of ALLAH, believing that there is no God but ALLAH and Hazrat Muhammad (PBUH) is HIS last prophet who made knowledge as the basis for guiding the humanity to the correct path. I am not sufficiently able to find textual representation for thanking ALMIGHTY ALLAH whose blessing the talented teachers have bestowed on me, providing me with fruitful opportunities and enabling me to pursue and carry out this study.

First, my greatest appreciation comes from my good, affectionate, kind-hearted, and most esteemed Supervisor, Associate Prof. Dr Hairulnizam Mahdin, and my mentor, Dr. Atta-ur-Rehman, for their professional and sincere guidance during my study. They have always honestly and truly guided me in all my research work. My target was difficult to attain without their assistance. Their capabilities have allowed me to complete my entire work within the time frame possible.

I will never forget that the Faculty of Computer Science and Information Technology (FSKTM) and Universiti Tun Hussein Onn Malaysia (UTHM) provide educational facilities and research-oriented environment. The UTHM staff and administration's sincere and continuing efforts to make all modern and latest facilities available to impact quality education in all fields are noteworthy. It was their sincere efforts and approach that made us capable in research-oriented field of learning information technology (IT).

Finally, I'm grateful to my parents and family members, their prayers stayed unconditionally with me, and are still with me. I was greatly supported by their heartfelt prayers, guidance, encouragement, affection, and care in every step of my life. In the world. My parents were to me always the greatest source of inspiration and role model. I am very proud to say that my parents are not highly educated, but their position is indescribable in enabling me to achieve the highest academic qualification. I thank them as their selfless and restless encouragement made my entire academic

accomplishments possible. This study, I devote to my parents. Special thanks to my beloved father who inspired me in some way, socially, mentally, spiritually or any other sort. I am also grateful to all UTHM friends for their help, support, motivation, and wonderful time they gave me out of my country during this long stay and never made me feel away from home. Thanks so much.



ABSTRACT

Automatic information extraction (IE) from online published scientific resources (mainly semi-structured and unstructured) like articles, proceedings, editorials etc. is among the hottest areas of research in text mining. This information is essential for various reasons like tagging, searching, indexing the documents and search engine optimization. In this regard, various techniques possessing considerable accuracy besides other merits, have been proposed in the literature. However, their efficiency is limited to domain-specific documents with static and well-defined formats. Whereas the accuracy is significantly challenged with a slight modification in the document format. Hence, it can be safely stated that so far, no scheme is robust enough for broader types, domains, and formats of documents from diverse publishing societies. To address this issue, an Ontological Framework for IE (OFIE) using a fuzzy rule-based system (FRBS) and an efficient word sense disambiguation (WSD) technique is proposed in this research. The FRBS module is responsible for IE in a precise manner by incorporating fuzzy regular expressions with an added tolerance factor conceived experimentally. FRBS is applied to XML and text converted versions of the same input document to extract two streams. Afterwards, the WSD module synthesizes both streams and yields the outcome that is promising semantically as well as syntactically. The domain is significantly wide-ranging and comprises of articles from well-known publishing services like IEEE, ACM, Elsevier, Springer, and few others. It is observed from extensive experiments and contrasting with state-of-the-art techniques that the proposed scheme is robust to changes in format, extracts better information, and exhibits a significant precision, recall and F-score as 89.14%, 89.6% and 89%, respectively in testing phase. As an outcome, the extracted information can be stored in a digital library for the sake of archiving and retrieval by means of extract, transform and load (ETL) process.

ABSTRAK

Pengekstrakan maklumat secara automatik daripada sumber ilmiah yang diterbitkan dalam talian terutamanya dokumen separa berstruktur dan tidak berstruktur seperti artikel, prosiding dan editorial adalah antara bidang penyelidikan yang sedang hangat di dalam perlombongan teks. Maklumat ini penting untuk pelbagai sebab seperti membuat tag, pencarian, pengindeksan dokumen web dan pengoptimuman enjin carian. Dalam hal ini, berbagai teknik dan sistem yang mempunyai ketepatan dan pelbagai merit yang lain telah dicadangkan di dalam literatur. Namun, kecekapannya terbatas hanya pada dokumen dalam domain khusus yang mempunyai format yang statik dan jelas. Tambahan lagi, keupayaannya boleh dicabar walaupun hanya dengan sedikit pengubahsuaian dalam format dokumen. Oleh itu, dapat dinyatakan dengan jelas bahawa tidak ada cara yang tahan lasak untuk pelbagai jenis, domain, dan format dokumen yang lebih terbuka yang ada di dalam dunia penerbitan. Untuk menangani masalah ini, satu kerangka Ontologi untuk Pengekstrakan Maklumat (OFIE) menggunakan sistem berasaskan peraturan kabur (FRBS) dan teknik penyahtaksaan pengesanan perkataan (WSD) yang efisien dicadangkan di dalam penyelidikan ini. Modul FRBS bertanggungjawab untuk mengekstrak maklumat secara tepat melalui penggunaan ekspresi kabus biasa dengan penambahan faktor tolerans yang dihasilkan melalui eksperimen. FRBS diaplikasikan kepada XML dan teks dalam versi yang telah diubah daripada dokumen input yang sama dan kedua-dua aliran ini diekstrak bersama. Selepas itu, modul WSD mensintesis kedua-dua aliran ini dan ia menghasilkan dapatan akhir yang secara semantik dan sintatiknya adalah baik. Domain dokumen adalah sangat besar dan terdiri daripada dokumen yang diterbitkan dari perkhidmatan penerbitan terkenal seperti IEEE, ACM, Elsevier, Springer, dan beberapa yang lain. Pemerhatian yang dijalankan berdasarkan eksperimen-eksperimen secara mendalam dan mendapati ia berbeza dengan teknik-teknik terkini kerana cadangan baru ini adalah lebih tahan lasak terhadap format yang berubah-ubah, mengekstrak lebih banyak

maklumat, menunjukkan ketepatan yang lebih signifikan, penarikan dan skor F yang signifikan di mana ia mencatat 89.14%, 89.6% dan 89%, di dalam fasa ujian. Hasilnya, maklumat yang diekstrak ini dapat disimpan di perpustakaan digital untuk tujuan pengarkiban dan capaian semula melalui proses ekstrak, transformasi dan pemuatan (ETL).



TABLE OF CONTENTS

TITLE	i
DECLARATION	ii
DEDICATION	iii
ACKNOWLEDGEMENT	iv
ABSTRACT	vi
ABSTRAK	vii
LIST OF TABLES	xiii
LIST OF FIGURES	xiv
LIST OF SYMBOLS AND ABBREVIATIONS	xvi
LIST OF APPENDICES	xviii
LIST OF PUBLICATIONS	xix
CHAPTER 1 INTRODUCTION	1
1.1 Research Background	1
1.2 Problem Statement	4
1.3 Research Objectives	5
1.4 Research Scope	6
1.5 Research Significance	6
1.6 Organisation of the Thesis	6
CHAPTER 2 LITERATURE REVIEW	8

2.1	Information Extraction	8
2.1.1	Information extraction from text	10
2.2	Ontology Base Information Extraction	11
2.3	Information Extraction from Scholarly Articles	15
2.4	Systems Structured Information Extraction Research	18
2.4.1	The CiteSeerX system	18
2.4.2	The ParsCit System	19
2.4.3	ArnetMiner	20
2.4.4	Extraction of Bibliographic References	20
2.5	Approaches and Models for Structured-Information Extraction from Research Articles	21
2.5.1	Rule-based Approaches	21
2.5.2	Machine-Learning Based Approaches	26
2.5.3	Fuzzy Rule Based Approaches	31
2.5.4	Other Hybrid Approaches	33
2.6	Word Sense Disambiguation (WSD)	37
2.7	Information Extraction Tools	39
2.8	Discussion (Research Gap)	41
2.9	Chapter Summary	42
CHAPTER 3 RESEARCH METHODOLOGY		43
3.1	Flowchart of Research Process	43
3.2	Proposed Framework Overview	45
3.3	Phase 1: Data Collection and Selection	45
3.3.1	Research Strategy	46
3.3.2	Criteria for Paper Selection and Evaluation	46

3.3.3	Dataset	47
3.4	Phase 2: Ontology Creation	50
3.5	Phase 3: Document Layout Analysis, Conversion and Rule Base Extraction Process	52
3.5.1	Document-Layout Analysis	52
3.5.2	Pattern and Rules Analysis	52
3.5.3	Pdf to XML and TEXT/ Word Conversion	54
3.5.4	Rule Identifier	56
3.5.5	Rule-Based Information Extraction	56
3.5.6	Acknowledgement Extractor	62
3.5.7	Reference Extractor	63
3.6	Phase 4: Fuzzy Regular Expressions	64
3.6.1	Fuzzy Rule-Based System	64
3.7	Phase 5: Word Sense Disambiguation	65
3.8	Digital Library System	67
3.9	Phase 6: Results analysis	67
3.9.1	Performance Evaluation	67
3.10	Comparison Analysis	69
3.11	Chapter summary	69

CHAPTER 4 PROPOSED ONTOLOGICAL FRAMEWORK FOR INFORMATION EXTRACTION

4.1	Proposed OFIE	71
4.2	Phase A: Pdf Document Conversion	73
4.3	Phase B: Structural Information Extraction	74
4.3.1	Fuzzy Rule-Based System	74

4.3.2	Components of Fuzzy Rule-Based System	76
4.4	Phase C: Word Sense Disambiguation	78
4.5	Ontology	81
4.6	Digital Library System	82
4.7	Chapter Summary	83
CHAPTER 5 RESULTS AND DISCUSSION		84
5.1	Implementation	85
5.2	Evaluation	93
5.3	Quantitative Comparison (Empirical)	98
5.3.1	Fuzzy Module Comparison (Partial)	98
5.3.2	Proposed OFIE Comparison	99
5.3.3	Qualitative Comparison	103
5.4	Digital Library System for pdf Sources	104
5.5	Chapter Summary	106
CHAPTER 6 CONCLUSION AND FUTURE WORK		107
6.1	Objectives Accomplished	109
6.2	Contributions	110
6.3	Future Work	111
6.3.1	Machine Learning Evolutionary Approaches	111
6.3.2	Fault Tolerant	111
6.4	Concluding Remarks	112
REFERENCES		113
APPENDIX		120
VITA		134

LIST OF TABLES

2.1	Summary of ontology base IE approaches from research articles	15
2.2	Summary of rule-based approaches, IE from research articles	25
2.3	Summary machine learning approaches, IE from research	30
2.4	Accuracies against datasets of different scientific papers of IE	36
2.5	Summary of hybrid approaches, IE from research articles	37
2.6	Metadata tools for information extraction	40
3.1	Dataset used in the evaluation.	47
3.2	The breakdown of metadata extraction phase into individual	59
3.3	Dissection individual steps logical structural body extraction	62
3.4	Breakdown of ack. and reference extraction into separate steps	63
4.1	Error measurements in fuzzy regular expression	74
4.2	Lookup table	77
4.3	WSD example	81
5.1	Implementation details	85
5.2	Synthesised extracted information.	90
5.3	Comparison of different schemes with information fields	93
5.4	Section based precision testing and training phases	95
5.5	Section based precision in testing and training phases	96
5.6	Comparison fuzzy system module state-of-the-art techniques	98
5.7	Comparison based on overall extraction results	100
5.8	Comparison with (Tkaczyk et al., 2015) for sub-sections	102
5.9	Comparison with state-of-the-art techniques	104

LIST OF FIGURES

2.1	IE transform an unstructured text into sets of facts	9
2.2	Classification of information extraction sub-tasks	10
2.3	General architecture of OBIE system	12
2.4	Number of documents in PubMed files by year of publication.	17
2.5	The number of DBLP documents by release year and form.	18
3.1	Research flow chart	44
3.2	Study source selection	46
3.3	ACM dataset sources	47
3.4	IEEE dataset sources	48
3.5	Springer dataset sources	49
3.6	Elsevier dataset sources	50
3.7	Flow chart of ontology creation.	51
3.8	Document region identification	52
3.9	Logical structural fields of research articles	53
3.10	Pdf-to-text format conversion from example document	54
3.11	Pdf-to-XML format conversion from example document	55
3.12	Pdf-to-WORD format conversion from example document	55
3.13	Proposed rule-based logical structural IE architecture of article	57
3.14	Metadata extraction	58
3.15	Extracted section numbers with headings from example article	60
3.16	Extracted tables numbers with caption from example article	61
3.17	Extracted figure number with caption from example article	61
3.18	Extracted acknowledgement from example article	62
3.19	Extracted references from example article	63
3.20	Word sense disambiguation	66

4.1	The proposed OFIE framework	72
4.2	Schematic of fuzzy system	75
4.3	First input variable society (Soc)	77
4.4	Second input variable structural index (SI)	78
4.5	Output variable tolerance (T)	78
4.6	FRBS briefly	78
4.7	Sentence similarity computation diagram	80
4.8	Proposed ontology of research paper	82
5.1	Main screen	86
5.2	Information extraction from TEXT version	87
5.3	Information extraction from XML version	88
5.4	Synthesised information after WSD module	89
5.5	Prototype of extracted information	91
5.6	Training phase comparison for fuzzy and WSD module	97
5.7	Testing phase comparison for fuzzy and WSD module	97
5.8	Fuzzy module comparison (testing phase)	99
5.9	Comparison in Training phase	101
5.10	Comparison in the testing phase	101
5.11	Section-wise performance comparison	103
5.12	Digital library conceptual model	105



LIST OF SYMBOLS AND ABBREVIATIONS

IE	-	Information Extraction
OFIE	-	Ontological Framework for Information Extraction
CRF	-	Conditional Random Fields
HMM	-	Hidden Markov Model
WSD	-	Word Sense Disambiguation
NLP	-	Natural Language Processing
MUC	-	Message Understanding Conference
ACE	-	Automatic Content Extraction
IR	-	Information retrieval
POS	-	Part-Of-Speech
GATE	-	General Architecture for Text Engineering
SVM	-	Support Vector Machines
ACI	-	Autonomous Citation Indexing
LS	-	Logical Structure
GS	-	Generic Section
OCR	-	Optical character recognition
LNCS	-	Lecture Notes in Computer Science
TE	-	Template Extraction
BW	-	Baum – Welch
KNN	-	K-Nearest Neighbour
DL	-	Digital library
REGEX	-	Regular Expression
FREGEX	-	Fuzzy Regular Expression
SI	-	Structural index
ETL	-	Extract, Transform and Load
FRBS	-	Fuzzy Rule Based System

RTF	-	Rich text format
T	-	Tolerance
Soc	-	Society
E_{avg}	-	Average error
CAD	-	Centre Average Defuzzifier
MIE	-	Mamdani Inference Engine
RDF	-	Resource description framework
RDBMS	-	Relational database management system



PTTHM
PERPUSTAKAAN TUNKU TUN AMINAH

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	List of Journals Selected for Dataset	123



PTTA UTHM
PERPUSTAKAAN TUNKU TUN AMINAH

LIST OF PUBLICATIONS

- i. **Zaman, G.**, Mahdin, H., Hussain, K., Abawajy, J., & Mostafa, S. A. (2021). An Ontological Framework for Information Extraction from Diverse Scientific Sources. *IEEE access*, 9, 42111-42124.
- ii. **Zaman, G.** (2020). Digital Library of Online PDF Sources: An ETL Approach. *IJCSNS*, 20(11), 173.
- iii. **Zaman, G.**, Mahdin, H., Hussain, K., & Rahman, A. (2020). Information extraction from semi and unstructured data sources: A systematic literature review. *ICIC Express Lett.*, 14(6), 593-603.



CHAPTER 1

INTRODUCTION

This chapter provides a detailed introduction to the research study conducted. It mainly provides the background of the studies already accompanied in the area of information extraction. It further describes the statement of the problem being addresses in the underlying research study and its potential objectives are precisely enlisted. Moreover, the research scope is adequately defined based on the provided background and research significance is highlighted.

1.1 Research Background

Scientific repositories maintained by research societies such as IEEE, ACM, Elsevier and Springer have become an increasingly important tool for stakeholders including researchers, businesses, research institutions, government agencies as well as funding agencies (Salloum *et al.*, 2018). These scientific repositories host millions of published documents that provide rich and useful information to the stakeholders. For example, as of October 2019, the IEEE database contains 5 million documents. Similarly, Elsevier publishes more than 430,000 articles annually in 2,500 journals and its archives contain over 13 million documents (Group, 2017) while Wiley Online Library has more than 4 million articles. They all contain pieces of information that are needed by the research community and other interested parties. The published articles are hosted in the form of structured and unstructured portable document format (pdf) of varying sizes. The structured pdf documents have all the necessary metadata information including the table of contents and sections. However, the unstructured

pdf documents contain only the basic metadata fields that include date-time stamp, file size and name or page numbers and used fonts.

The information age challenges us with the effective presentation of large volumes of information that people must absorb to perform their jobs (Nasar & Jaffry, 2018). This issue exists across the board in many occupations. Data must be consolidated and organised so that the user of information can quickly understand and act upon it. Data presentation is perhaps a fairly obvious problem when the data in question involves hundreds or thousands of numeric readings from which someone makes a decision. This increase in scientific content presents significant challenges for researchers who wish to determine, in their particular field of interest, the state of the art. Extracting data manually from such sources is humanly impossible due to the nature and huge volume of data. Thus, automated information extraction is required to serve the purpose. Search engines such as Google on the other hand performs information retrieval based on already extracted and indexed information available over the web (Hofmann *et al.*, 2016). The crawlers continuously keep on indexing the document for the relevant terms. A whole domain, the information extraction (IE), is used to extract potential data nuggets to resolve these and other related problems. The IE focuses mainly on the extraction of structured information from unstructured or semi-structured data. It is widely used in several fields, e.g., the field of medical sciences.

It is apparent that extracting information manually from such a huge number of documents is almost impossible (Azimjonov & Alikhanov, 2018; Lipinski *et al.*, 2013; Ma *et al.*, 2013; Tkaczyk, 2017). Various techniques from different fields have been proposed for efficient information extraction from scientific repositories (Qureshi *et al.*, 2014, 2016; Rizvi *et al.*, 2018b). These techniques include ontology, natural language processing (NLP), machine learning (ML), conditional random fields (CRF) based information extraction and some hybrid techniques (Ahmad *et al.*, 2016; Jayaram & Sangeeta, 2017b; Salloum *et al.*, 2018).

However, extraction of content and metadata from scientific repositories has remained a challenging task. Especially, the huge volume and the varying format of documents pose major technical challenges to efficiently extract the desired information from the repositories. Even the search engines are facing problems in indexing such a massive volume of documents of varying format (Jayaram &

Sangeeta, 2017b). This problem is getting worse as the volume of the generated documents is exponentially increasing (Bodo & Csato, 2017; Feldman & Sanger, n.d.; Groth, Lauruhn, Scerri, & Daniel Jr, 2018). Moreover, the bulk of scientific documents hosted in the publishers' digital libraries (Ahmad & Abawajy, 2014) are mostly unstructured which presents a considerable challenge in reliably and efficiently extracting the required information from such repositories (Do *et al.*, 2013; Kim *et al.*, 2014). Although both information extraction and metadata extraction are generally sensitive to variations in document formats and fields of metadata, existing work does not consider this issue. Also, the extraction of structural information from unstructured and semi-structured published scientific articles has received little attention in the past. However, nowadays, it is among the hottest areas of research due to emerging web technologies like semantic web (Azimjonov & Alikhanov, 2018; Lipinski *et al.*, 2013). Therefore, there is a strong need to overcome these challenges and develop an efficient information extraction mechanism from the published scientific documents.

Ahmad *et al.*, (2016) the authors proposed a rule-based method for information extraction from scientific articles in the form of XML or simple text. The empirical tests were performed on XML files with the help of pdfx online tool. The authors built an ontology and utilised a rule-based approach after crafting the rules by observing the given dataset of documents. The technique possessed an accuracy of 77.5%. However, a major limitation of the technique was that it worked only for one specific conference format. Several approaches have been designed for the extraction process in this area of research such as XML documents and simple text documents (Chen *et al.*, 2018; Clark & Divvala, 2015; Rizvi *et al.*, 2018b) These references never detect the suitable knowledge of document structure. In this field of research, various methods are available to extract data either from XML documents or from Plain-text documents (Casali *et al.*, 2016; Prasad *et al.*, 2018) Not even a single method uses the patterns in XML and text/word formats to identify the needed information in the published research articles. In support of a robust solution, using only one of these formats is not enough. (Nasar & Jaffry, 2018) provide a comprehensive literature review about the information extraction techniques from unstructured and semi-structured scientific resources. It was concluded that there is a significant need for a scheme that can comprehend diverse formats of scientific documents from various societies. Moreover, it is closely observed that a proper hybridisation of more than one techniques as a

framework for information extraction can be promising in terms of accuracy and scope (Dieb *et al.*, 2015; Tkaczyk *et al.*, 2015)

Based on the above discussion and the comprehensive literature review, a novel ontological framework has been proposed for structural information extraction from diverse scientific documents. The documents were taken from various research communities including ACM, IEEE, Springer, Elsevier and others. The framework is equipped with FRBS and WSD (Abd-Rashid *et al.*, 2018) to investigate the accuracy in contrast to the existing techniques based on empirical results. The fuzzy regular expressions (having an in-built Levenshtein-distance measure) (Room, 2019) enables the proposed approach to deal with structural variations and missing information (deleted, inserted, modified). The WSD helps in improving the accuracy of the extracted information using a semantic similarity measure along with auto-correction of words and sentences. A set of comprehensive experiments has been conducted using real datasets from various scientific repositories and validated the proposed approach. The proposed approach has also been compared with various baseline techniques.

1.2 Problem Statement

The problem of obtaining standardised data (structured information) from the scientific documents (from various publishers) for scholars, readers, and data scientists as well as for the indexing services and search engines remains complex and daunting. This is mainly because of the overwhelmingly increasing volume of documents with a variety of formats and layouts resulting in poor indexing and inefficient information retrieval over the web (Chen *et al.*, 2018; Jayaram & Sangeeta, 2017a) Although various techniques, that address this issue, exist (Bodo & Csato, 2017; Groth, Lauruhn, Scerri, & Daniel, 2018; Shah & Jain, 2014), they are developed to address a narrow domain and with specific rules applicable only to certain formats. Nonetheless, when these techniques are investigated over slightly modified formats, their performance is compromised. To overcome this issue of research availability, all relevant information must be precisely extracted from such documents to ensure that all relevant knowledge has been discovered. Therefore, it is paramount to have an efficient approach in extracting that information from published documents. The search engines can only

return general information which is a form of information retrieval; IE helps in providing that information.

Moreover, it is observed that ontology-based dynamic information extraction framework for pdf documents has not been investigated so far in combination with a fuzzy system and word sense disambiguation because:

- information extraction is highly domain-specific that mostly handles a specific document corpus and layout.
- it is hard to comprehend all possible variations among documents in terms of their layout and format even in the same subject area.
- with a little modified document format, the performance of IE systems is greatly compromised in terms of accuracy. This shows that there is a big gap for improvement in terms of accuracy and robustness.
- massive volume & structural differences of scientific documents result in
 - poor indexing (indexing services are unable to adequately index the documents)
 - inefficient searching/discoverability of the deemed data due to lack of unstructured information.

Moreover, it is worthy to investigate the hybrid approaches where each component contributes to overall problem handling. A proper hybridisation of more than one technique, as a framework for information extraction, can be promising in terms of accuracy and scope (Dieb *et al.*, 2015; Tkaczyk *et al.*, 2015).

1.3 Research Objectives

The main objectives of this research are as follows:

- i. To propose an ontology-based framework for information extraction from scientific documents of different scientific communities with diverse formats.
- ii. To develop a Fuzzy Rule-Based System for information extraction based on the framework followed by Word Sense Disambiguation (WSD) to enhance the accuracy.
- iii. To evaluate the proposed framework on the experimental data and compare it with the baseline approaches for information extraction.

1.4 Research Scope

This study focuses on a novel ontology-based approach for extraction of structural information from published scientific articles in pdf format using a Fuzzy Rule-Based System and to investigate its accuracy with existing techniques based on empirical results. The dataset comprised of documents from various research communities including ACM, IEEE, Springer, Elsevier, and others. Each publisher has its own format. For example, title styles, ways to describe the author names and their affiliations, headings and captions of figures and tables etc. Due to this diversity, it is hard to comprehend all these formats using a single information extraction technique. That demands a comprehensive framework to address all the format related issues.

1.5 Research Significance

The research will produce a framework for information extraction from scientific documents with improved accuracy along with considering the diverse nature of documents from ACM, IEEE, Springer, Elsevier, and several others' styles rather than just one type of document. The proposed framework will be robust against varying document formats. Eventually, the scheme can be used for search engines, indexing services and digital libraries for automated information extraction and better utilisation/retrieval. Finally, the approach can be utilised for automated information extraction and better utilisation/retrieval in search engines, indexing services, and digital libraries.

1.6 Organisation of the Thesis

- *Chapter 1* provides a brief overview of the problem statement, priorities and aims, nature and importance of this research analysis with context knowledge of the Ontological System for Information Extraction from various scientific sources.
- *Chapter 2* describes fundamental and theoretical ideas of several information extraction approaches and traditional systems and models that are used to resolve various kinds of information extraction issues. This chapter also

REFERENCES

- Abd-Rashid, A., Abdul-Rahman, S., Yusof, N. N., & Mohamed, A. (2018). Word sense disambiguation using fuzzy semantic-based string similarity model. *Malaysian Journal Of Computing*, 3(2), 154–161.
- Adnan, K., & Akbar, R. (2019a). An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data*, 6(1), 1–38.
- Adnan, K., & Akbar, R. (2019b). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management*, 11, 1847979019890771.
- Ahmad, M., & Abawajy, J. H. (2014). Digital library service quality assessment model. *Procedia-Social and Behavioral Sciences*, 129, 571–580.
- Ahmad, R., Afzal, M. T., & Qadir, M. A. (2016). Information extraction from pdf sources based on rule-based system using integrated formats. *Semantic Web Evaluation Challenge*, 293–308.
- Ahmed, M. W., & Afzal, M. T. (2020). FLAG-pdf: Features oriented metadata extraction framework for scientific publications. *IEEE Access*, 8, 99458–99469.
- Amarnadh, S., Rao, V. S., Prasad, M. A. R., & Rao, V. V. (n.d.). *A Study on Meta Data Extraction Systems and Features of Cloud Monitoring*.
- Anzaroot, S., & McCallum, A. (2014). *A New Dataset for Fine Grained Citation Field Extraction (Author's Manuscript)*. University of Massachusetts, Amherst Amherst United States.
- Azimjonov, J., & Alikhanov, J. (2018). Rule Based Metadata Extraction Framework from Academic Articles. *ArXiv Preprint ArXiv:1807.09009*.
- Biniam, P. (2020). *Ontology-based information extraction from legacy surveillance reports of infectious diseases in animals and humans*.
- Björk, B.-C. (2017). Growth of hybrid open access, 2009–2016. *PeerJ*, 5, e3878.
- Bodó, Z., & Csató, L. (2017). A Hybrid Approach for Scholarly Information Extraction. *Studia Universitatis Babeş-Bolyai, Informatica*, 62(2).
- Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*, 7(2), e012545.
- Burch, M., Pompe, D., & Weiskopf, D. (2015). An analysis and visualization tool for DBLP data. *2015 19th International Conference on Information Visualisation*, 163–170.

- Casali, A., Deco, C., & Beltramone, S. (2016). An assistant to populate repositories: gathering educational digital objects and metadata extraction. *IEEE Revista Iberoamericana de Tecnologías Del Aprendizaje*, 11(2), 87–94.
- Chen, J., Zhang, C., & Niu, Z. (2018). A Two-Step Resume Information Extraction Algorithm. *Mathematical Problems in Engineering*, 2018. <https://doi.org/10.1155/2018/5761287>
- Chenet, M. (2017). *Identify and extract entities from bibliography references in a free text*. University of Twente.
- Chiang, I.-J., Liu, C. C.-H., Tsai, Y.-H., & Kumar, A. (2015). Discovering latent semantics in web documents using fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 23(6), 2122–2134.
- Choudhury, M. H., Wu, J., Ingram, W. A., & Fox, E. A. (2020). A Heuristic Baseline Method for Metadata Extraction from Scanned Electronic Theses and Dissertations. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 515–516.
- Clark, C. A., & Divvala, S. K. (2015). Looking Beyond Text: Extracting Figures, Tables and Captions from Computer Science Papers. *AAAI Workshop: Scholarly Big Data*, 6.
- Constantin, A., Peroni, S., Pettifer, S., Shotton, D., & Vitali, F. (2016). The document components ontology (DoCO). *Semantic Web*, 7(2), 167–181.
- Cuong, N. V., Chandrasekaran, M. K., Kan, M.-Y., & Lee, W. S. (2015). Scholarly document information extraction using extensible features for efficient higher order semi-CRFs. *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 61–64.
- Dash, S., Luhach, A. K., Chilamkurti, N., Baek, S., & Nam, Y. (2019). A Neuro-fuzzy approach for user behaviour classification and prediction. *Journal of Cloud Computing*, 8(1), 17.
- Dieb, T. M., Yoshioka, M., Hara, S., & Newton, M. C. (2015). Framework for automatic information extraction from research papers on nanocrystal devices. *Beilstein Journal of Nanotechnology*, 6, 1872.
- Do, H. H. N., Chandrasekaran, M. K., Cho, P. S., & Kan, M. Y. (2013). Extracting and matching authors and affiliations in scholarly documents. *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 219–228.
- Elizarov, A., Khaydarov, S., & Lipachev, E. (2017). Scientific documents ontologies for semantic representation of digital libraries. *2017 Second Russia and Pacific Conference on Computer Technology and Applications (RPC)*, 1–5.
- Farhat, R., Jebali, B., & Jemni, M. (2015). Ontology based semantic metadata extraction system for learning objects. In *Emerging Issues in Smart Learning* (pp. 247–250). Springer.
- Feldman, R., & Sanger, J. (n.d.). *The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data*.
- Groth, P., Lauruhn, M., Scerri, A., & Daniel Jr, R. (2018). Open information extraction

on scientific text: An evaluation. *ArXiv Preprint ArXiv:1802.05574*.

- Groth, P., Lauruhn, M., Scerri, A., & Daniel, R. (2018). *Open Information Extraction on Scientific Text: An Evaluation*. 3414–3423. <https://doi.org/10.17632/6m5dyx4b58>
- Group, R. (2017). *RELEX. Annual Report*.
- Han, L., Deng, X., & Wu, G. (2017). A knowledge-based word sense disambiguation algorithm utilizing syntactic dependency relation. *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 85–90.
- Hartmann, W. M. (2005). *Modern Acoustics and Signal Processing*.
- Hofmann, K., Li, L., & Radlinski, F. (2016). Online evaluation for information retrieval. *Foundations and Trends in Information Retrieval*, 10(1), 1–117.
- Indrawati, A., Yoganingrum, A., & Yuwono, P. (2019). Evaluating the Quality of the Indonesian Scientific Journal References using ParsCit, CERMINE and GROBID. *Library Philosophy and Practice*, 1–14.
- Jayaram, K., & Sangeeta, K. (2017a). A review: Information extraction techniques from research papers. *IEEE International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2017 - Proceedings, Icimia*, 56–59. <https://doi.org/10.1109/ICIMIA.2017.7975532>
- Jayaram, K., & Sangeeta, K. (2017b). A review: Information extraction techniques from research papers. *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 56–59.
- Jiang, C., Liu, J., Ou, D., Wang, Y., & Yu, L. (2018). Implicit semantics based metadata extraction and matching of scholarly documents. *Journal of Database Management (JDM)*, 29(2), 1–22.
- Junaedi, H., Ikhsan, T. P., & Purwanto, D. D. (2021). Information Extraction From ICMD Documents To Determine The Ratio Factors Function Performance using Fuzzy. *2021 3rd East Indonesia Conference on Computer and Information Technology (EIconCIT)*, 244–247.
- Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. *PloS One*, 9(5), e93949.
- Khusro, S., Latif, A., & Ullah, I. (2015). On methods and tools of table detection, extraction and annotation in pdf documents. *Journal of Information Science*, 41(1), 41–57.
- Kim, S., Cho, Y., & Ahn, K. (2014). Semi-automatic metadata extraction from scientific journal article for full-text XML conversion. *Proceedings of the International Conference on Data Science (ICDATA)*, 1.
- Körner, M. (2017). *Reference String Extraction Using Line-Based Conditional Random Fields*. <http://arxiv.org/abs/1705.08154>
- Kovriguina, L., Shipilo, A., Kozlov, F., Kolchin, M., & Cherny, E. (2015). Metadata extraction from conference proceedings using template-based approach.

Semantic Web Evaluation Challenges, 153–164.

- Kyvik, S., & Aksnes, D. W. (2015). Explaining the increase in publication productivity among academic staff: A generational perspective. *Studies in Higher Education*, 40(8), 1438–1453.
- Lauscher, A., Eckert, K., Galke, L., Scherp, A., Rizvi, S. T. R., Ahmed, S., Dengel, A., Zumstein, P., & Klein, A. (2018). Linked open citation database: Enabling libraries to contribute to an open and interconnected citation graph. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 109–118.
- Lim, C.-G., Jeong, Y.-S., & Choi, H.-J. (2019). Survey of temporal information extraction. *Journal of Information Processing Systems*, 15(4), 931–956.
- Lipinski, M., Yao, K., Breiting, C., Beel, J., & Gipp, B. (2013). Evaluation of header metadata extraction approaches and tools for scientific pdf documents. *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 385–386.
- Liu, R., Gao, L., An, D., Jiang, Z., & Tang, Z. (2017). Automatic document metadata extraction based on deep networks. *National CCF Conference on Natural Language Processing and Chinese Computing*, 305–317.
- Luong, M.-T., Nguyen, T. D., & Kan, M.-Y. (2012). Logical structure recovery in scholarly articles with rich document features. In *Multimedia Storage and Retrieval Innovations for Digital Library Systems* (pp. 270–292). IGI Global.
- Ma, X., Qin, H., Sulaiman, N., Herawan, T., & Abawajy, J. H. (2013). The parameter reduction of the interval-valued fuzzy soft sets and its related algorithms. *IEEE Transactions on Fuzzy Systems*, 22(1), 57–71.
- Magdalena, L. (2015). Fuzzy rule-based systems. In *Springer Handbook of Computational Intelligence* (pp. 203–218). Springer.
- Mahapatra, D., Maharana, C., Panda, S. P., Mohanty, J. P., Talib, A., & Mangaraj, A. (2020). A fuzzy-cluster based semantic information retrieval system. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 675–678.
- Mannai, M., Karâa, W. B. A., & Ghezala, H. H. Ben. (2018). Information Extraction Approaches: A Survey. In *Information and Communication Technology* (pp. 289–297). Springer.
- Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & López-Cózar, E. D. (2020). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 1–36.
- Martínez-Romero, M., O'Connor, M. J., Dorf, M., Vendetti, J., Willrett, D., Egyedi, A. L., Graybeal, J., & Musen, M. A. (2017). Supporting ontology-based standardization of biomedical metadata in the CEDAR Workbench. *Proceedings of the Int Conf Biom Ont (ICBO)*, 1–6.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in*

Neural Information Processing Systems, 3111–3119.

- Mishina, K., Tsuchiya, S., & Watabe, H. (2017). Word sense disambiguation of adjectives using dependency structure and degree of association between sentences. *2017 International Conference on Asian Language Processing (IALP)*, 342–345.
- Mizera-Pietraszko, J., & Machalewski, T. (2016). Extraction of medical terms for word sense disambiguation within multilingual framework. *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, 478–484.
- Moltmann, F. (2017). *Natural language ontology*.
- Nasar, Z., & Jaffry, S. W. (2018). Trust-based situation awareness: Agent-based versus population-based modeling a comparative study. *2018 International Conference on Advancements in Computational Sciences (ICACS)*, 1–7.
- Niklaus, C., Cetto, M., Freitas, A., & Handschuh, S. (2018). A survey on open information extraction. *ArXiv Preprint ArXiv:1806.05599*.
- Ojokoh, B., Zhang, M., & Tang, J. (2011). A trigram hidden Markov model for metadata extraction from heterogeneous references. *Information Sciences*, 181(9), 1538–1551.
- Peng, F., & McCallum, A. (2006). Information extraction from research papers using conditional random fields. *Information Processing & Management*, 42(4), 963–979.
- Popovski, G., Seljak, B. K., & Eftimov, T. (2020). A survey of named-entity recognition methods for food information extraction. *IEEE Access*, 8, 31586–31594.
- Prasad, A., Kaur, M., & Kan, M.-Y. (2018). Neural ParsCit: a deep learning-based reference string parser. *International Journal on Digital Libraries*, 19(4), 323–337.
- Prasath, R. R., & Öztürk, P. (2016). An Approach to Content Extraction from Scientific Articles using Case-Based Reasoning. *Res. Comput. Sci.*, 117, 85–96.
- Qin, W., Elanwar, R., & Betke, M. (2020). Text and metadata extraction from scanned Arabic documents using support vector machines. *Journal of Information Science*, 0165551520961256.
- Qureshi, I. M., Malik, A. N., & Naseem, M. T. (2014). Dynamic resource allocation in OFDM systems using DE and FRBS. *Journal of Intelligent & Fuzzy Systems*, 26(4), 2035–2046.
- Qureshi, I. M., Malik, A. N., & Naseem, M. T. (2016). QoS and rate enhancement in DVB-S2 using fuzzy rule based system. *Journal of Intelligent & Fuzzy Systems*, 30(2), 801–810.
- Rahnama, M., Hasheminejad, S. M. H., & Nasiri, J. A. (2020). Automatic Metadata Extraction From Iranian Theses And Dissertations. *2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 1–5.

- Ramesh, S. H., Dhar, A., Kumar, R. R., Anjaly, V., Sarath, K. S., Pearce, J., & Sundaresan, K. R. (2016). Automatically identify and label sections in scientific journals using conditional random fields. *Semantic Web Evaluation Challenge*, 269–280.
- Rizvi, S. T. R., Mercier, D., Agne, S., Erkel, S., Dengel, A., & Ahmed, S. (2018a). Ontology-based Information Extraction from Technical Documents. *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, 2(Icaart), 493–500. <https://doi.org/10.5220/0006596604930500>
- Rizvi, S. T. R., Mercier, D., Agne, S., Erkel, S., Dengel, A., & Ahmed, S. (2018b). Ontology-based Information Extraction from Technical Documents. *ICAART (2)*, 493–500.
- Romary, L., & Lopez, P. (2015). Grobid-information extraction from scientific publications. *ERCIM News*, 100.
- Room, C. (2019). Levenshtein Distance. *Algorithms*, 12(14), 32.
- Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). Using text mining techniques for extracting information from research articles. In *Intelligent natural language processing: trends and applications* (pp. 373–397). Springer.
- Seol, J.-W., Choi, W.-J., Jeong, H.-S., Hwang, H.-K., & Yoon, H.-M. (2018). Reference Metadata Extraction from Korean Research Papers. *International Conference on Mining Intelligence and Knowledge Exploration*, 42–52.
- Shah, R., & Jain, S. (2014). *Ontology-based Information Extraction: An Overview and a Study of different Approaches*. 87(4), 6–8. <https://doi.org/10.5120/15194-3574>
- Sharma, P., & Tripathi, R. C. (2017). Patent citation: A technique for measuring the knowledge flow of information and innovation. *World Patent Information*, 51, 31–42.
- Sleeman, J., Finin, T., & Joshi, A. (2015). Entity type recognition for heterogeneous semantic graphs. *AI Magazine*, 36(1), 75–86.
- Tang, J. (2016). Aminer: Toward understanding big scholar data. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 467.
- Tansazan, A., & Mahdavi, M. A. (2017). Metadata Extraction from Persian Scientific Papers Using CRF Model. *Library and Information Science Research*, 7(1), 304–321.
- Tkaczyk, D. (2017). New Methods for Metadata Extraction from Scientific Literature. *ArXiv Preprint ArXiv:1710.10201*.
- Tkaczyk, D., Collins, A., Sheridan, P., & Beel, J. (2018). Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 99–108.
- Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJ DAR)*, 18(4),

317–335.

- Tuarob, S., Mitra, P., & Giles, C. L. (2015). A hybrid approach to discover semantic hierarchical sections in scholarly documents. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 1081–1085.
- Vijayarajan, V., Dinakaran, M., Tejaswin, P., & Lohani, M. (2016). A generic framework for ontology-based information retrieval and image retrieval in web data. *Human-Centric Computing and Information Sciences*, 6(1), 1–30.
- Vilnis, L., Belanger, D., Sheldon, D., & McCallum, A. (2015). Bethe projections for non-local inference. *ArXiv Preprint ArXiv:1503.01397*.
- Visser, M., van Eck, N. J., & Waltman, L. (2020). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *ArXiv Preprint ArXiv:2005.10732*.
- Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126, 3–13.
- Ware, M., & Mabe, M. (2015). *The STM report: An overview of scientific and scholarly journal publishing*.
- Wen, L., Li, J., Jin, Y., & Lu, Y. (2016). A method for Word Sense Disambiguation combining contextual semantic features. *2016 International Conference on Asian Language Processing (IALP)*, 283–287.
- Wu, J., & Giles, C. L. (2020). Scholarly Very Large Data: Challenges For Digital Libraries. *Challenges For Large Scale Networking (LSN) Workshop on Huge Data: A Computing, Networking and Distributed Systems Perspective*.
- Yadav, P., Remala, N., & Pervin, N. (2019). RecCite: A Hybrid Approach to Recommend Potential Papers. *2019 IEEE International Conference on Big Data (Big Data)*, 2956–2964.
- Yang, Q., & Yuan, X. (2021). Interrelated information pair extraction algorithm of visual attention for form documents. *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, 560–567.
- Zhou, Q., Jiang, Z., & Yang, F. (2020). Sentences Similarity Based on Deep Structured Semantic Model and Semantic Role Labeling. *2020 International Conference on Asian Language Processing (IALP)*, 40–44.