IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 21, 2021, accepted February 18, 2021, date of publication March 2, 2021, date of current version March 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063181

An Ontological Framework for Information Extraction From Diverse Scientific Sources

GOHAR ZAMAN[1], HAIRULNIZAM MAHDIN[1], (Member, IEEE), KHALID HUSSAIN[2], ATTA-UR-RAHMAN[3], JEMAL ABAWAJY[4], (Senior Member, IEEE), AND SALAMA A. MOSTAFA[1]

[1]Center of Intelligent and Autonomous System, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat 86400, Malaysia
[2]Barani Institute of Sciences (Sahiwal), PMAS Arid Agriculture University, Rawalpindi 46000, Pakistan
[3]Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia
[4]School of Information Technology, Deakin University, Geelong, VIC 3217, Australia

Corresponding authors: Hairulnizam Mahdin (hairuln@uthm.edu.my) and Jemal Abawajy (jemal.abawajy@deakin.edu.au)

This work was supported by the Ministry of Education Malaysia (MOE) through the Fundamental Research Grant Scheme for Research Acculturation of Early Career Researchers (FRGS-Racer) under Grant RACER/1/2019/ICT04/UTHM/1 Vote: K154.

ABSTRACT Automatic information extraction from online published scientific documents is useful in various applications such as tagging, web indexing and search engine optimization. As a result, automatic information extraction has become among the hottest areas of research in text mining. Although various information extraction techniques have been proposed in the literature, their efficiency demands domain specific documents with static and well-defined format. Furthermore, their accuracy is challenged with a slight modification in the format. To overcome these issues, a novel ontological framework for information extraction (OFIE) using fuzzy rule-base (FRB) and word sense disambiguation (WSD) is proposed. The proposed approach is validated with a significantly wider document domains sourced from well-known publishing services such as IEEE, ACM, Elsevier, and Springer. We have also compared the proposed information extraction approach against state-of-the-art techniques. The results of the experiment show that the proposed approach is less sensitive to changes in the document format and has a significantly better average accuracy of 89.14% and F-score as 89%.

INDEX TERMS Information extraction, semi structure scientific documents, fuzzy rule base, word sense disambiguation, ontological framework.

I. INTRODUCTION

Scientific repositories maintained by research societies such as IEEE, ACM, Elsevier and Springer have become an increasingly important tool for diverse stakeholders that include researchers, businesses, research institutions, government agencies as well as funding agencies [1]. These scientific repositories host millions of published documents that provide rich and useful information to the stakeholders [2]. For example, as of October 2019, the IEEE database contains 5 million documents [3]. Similarly, Elsevier publishes more than 430,000 articles annually in 2,500 journals and its archives contain over 13 million documents [4]. Also, Wiley Online Library have more than 4 million articles. They all contain some piece of information that is 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 information. However, the unstructured documents contain only the basic metadata fields that include date-time stamp, file size and name or page numbers and used fonts.

It is apparent that manually retrieving information from such huge documents is near to impossible [5]–[9]. Various techniques from different fields have been proposed to automate information extraction from scientific repositories [10]–[12]. These techniques include ontology-based, natural language processing (NLP), machine learning (ML), conditional random fields (CRF) based information extraction and some hybrid techniques [1], [2], [13], [14].

The associate editor coordinating the review of this manuscript and approving it for publication was Davide Aloini.

However, automated content and metadata extraction from the scientific repositories has remained challenging. Specially, the huge volume and the varying format of the documents pose major technical challenges to efficiently extract the desired information from the repositories. Even the search engines are facing problems in indexing such massive volume and varying format documents [14]. This problem is getting worse as the volumes of the generated documents are exponentially increasing rapidly [15]–[17]. Moreover, the bulk of scientific documents hosted in the publishers digital libraries [18], [19] are mostly unstructured documents, which presents a considerable challenge to reliably and efficiently extract required information from such repositories [20]–[22]. Although, in general both information extraction and metadata extraction are sensitive to variations in the document formats and fields of metadata, exiting work do not consider this issue. Also, the extraction of structural information from unstructured/semi-structured published scientific articles has received little attention [7]–[9]. Therefore, there is a strong need to overcome these challenges and develop an efficient information extraction mechanism from the published scientific documents.

In this paper, we propose an efficient ontology-based approach for structural information extraction from the scientific documents. The proposed approach uses fuzzy rule-base (FRB) and word sense disambiguation (WSD). The Fuzzy regular expressions (having in-built Levenshtein-distance measure) enables the proposed approach to deal with structural variations and missing information (deleted, inserted, modified). The WSD helps in fixing the extracted information using semantic similarity measure along with auto-correction of words and generates the final stream. The proposed approach is effective to the type and amount of information being extracted as well as able to take various types of the targeted document into consideration like text, docx and XML. We have conducted comprehensive experiments using real-data sets from various scientific repositories and validated the proposed approach. We have also compared the proposed approach with various baseline techniques. The contribution of this paper can be summarized as follows:

- An adaptive and robust ontological framework for information extraction (OFIE) using fuzzy rule-base system (FRBS), plain & fuzzy regular expressions and word sense disambiguation (WSD) [23] using word2vec approach [24] has been proposed.
- An extensive experimental analysis of the proposed approach with real data from various repositories mainly from IEEE, ACM, Springer and Elsevier and some others, with significant variations.
- We compare the proposed approach using quantitative (in terms of numbers and experimental outputs) and qualitative (in terms of the features and dimensions of extraction fields) methods and show that the proposed approach outperforms in terms of several performance metrics.

The rest of the paper is organized as follows. Section 2 discusses literature review and Section 3 contains the proposed approach. Section 4 contains discussion of the results. Section 5 concludes the paper.

II. PROBLEM OVERVIEW AND LITERATURE REVIEW

In this section, we present an overview of the problem and a comprehensive literature review with emphases on the work related to the problem addressed in this paper.

A. PROBLEM OVERVIEW

Scientific repositories such as IEEE, Springer and ACM play an increasingly important role in modern research. These repositories maintain a large number of documents on research outcomes. Researchers produce massive amounts of information from their research outcomes as the number of published scholarly articles has increased between 8% and 9% annually [25]. This also implies that researchers consume massive amounts of information from many different scientific sources. As modern research relies heavily on already published research results, effective access to the huge scientific papers duly published by different publishers is crucial [2]. However, the process of Information Extraction of these documents on specific subject is inefficient due to the massive volumes of the documents and their structural differences which results in poor indexing and inefficient information retrieval over the web [14], [26]. It is not easy to extract information from these documents effectively and thus automatically extracting content and metadata from scientific repositories remains a challenge. It is not easy to extract information from these documents in an efficient manner and therefore automatic retrieval of content and metadata from scientific repositories remains a challenge.

Although various techniques have been proposed to address these challenges [15], [16], [27] there are several gaps in the exiting work. Exiting solutions are mainly developed to address only a narrow domain and with very specific rules that are only applicable to limited formats. Moreover, when the exiting techniques are investigated over slightly modified document formats, their performance is considerably compromised [17]. Moreover, ontology based dynamic Information Extraction framework for PDF and MS Word documents that recognizes a wide variety of document resources published in scientific community and extracts the complete structural information from them has not been investigated so far. Although some of the concepts are partially investigated, a proper hybridization of more than one technique as a framework for information extraction can be promising in terms of accuracy and scope [28], [29]. A recent comprehensive literature review about the information extraction techniques from unstructured and semi structured scientific resources have been discussed in [9]. The authors concluded that there is a dire need of a scheme that can comprehend diverse formats of scientific documents from various society repositories.

These unsolved challenges have served as a prime motivation behind the current study. To fill this research gaps,

a new approach that can comprehensively and efficiently extract all pertinent information from the entire spectrum of publications with precision is paramount.

B. RELATED WORK

A variety of techniques were proposed to automate the extraction of information from documents in scientific repositories. In this section, we present some of the most relevant and state-of-the-art approaches.

A rule-based method for information extraction from scientific documents in the form of XML or simple text is discussed in [13]. The authors built an ontology and utilized a rule-based approach after crafting the rules by observing the given dataset in the documents. The empirical tests were performed on XML files with the help of PDFx online tool and an accuracy of 77.5% was observed. The major limitation of the technique was that it was designed for only a specific conference format paper. Various methods for extracting information from XML documents [30]–[34] and from plain-text document [35]–[38] have also been developed. However, none of these approaches utilize the patterns in both XML and Text formats to identify the desired information of the published research articles.

In [39], the authors used PDFBox tool to get pure text and formatting values of the target document. The authors further developed their own tool called PAXAT which works on rich text features (RTF) of the document taken from published article from ACM, IEEE, SPRINGER and ArVix. For example, formatting values like the line height, font type, font size and alignment of different metadata items that include the title, author and affiliation (take help from given template). It also works on redundant text such as dates that appear on both header and footer. The technique, however, was confined to extracting the paper title, authors, and their affiliations only.

An approach called CERMINE (Content ExtRactor and MINEr) is proposed to retrieve different parts of an article in [28]. CERMINE implements different algorithms for extracting different part of an article. For example, the K-Mean algorithm is used for clustering of lines, Docstrum Algorithm is used for page segmentation, and the Support Vector Machine (SVM) algorithm is used for classification purpose. The CERMINE approach was a PDF to Word conversion rather than just information extraction.

In [40], the authors investigated PDFBox, TET (Text Extraction Toolkit), PDF2TET and Table Seer Algorithm techniques. Authors presented Tabular Ontology for extraction based on Semantic Relationship consist of Columns and Rows include Cell, header, Body and Associated Text Regions. Authors use Pre-defined layout approaches, Border Lines, statistical approach, Heuristic Approach. Rule Based Approach play important role in Table Data Extraction. The technique was designed for tabular data extraction only.

In [41], the authors investigated PDFBox and PARSCIT algorithm. Basically, The PARSCIT algorithm performs reference string parsing. Sometimes it is also called Citation Parsing or Citation Extraction from the given set of references. Parsing means to resolve a sentence into its components and describe their role according to the context. The given technique is mainly focused on the extraction of citations.

In [42], the authors investigated conditional random fields (CRF) and hidden Markov Model (HMM) techniques, and their estimation is improved and trained by using Particle Swarm Optimization (PSO). PSO searches the optimal value between CRF and HMM and finds the optimize answer. Purpose of this technique was to generate the citations for the given paper including its digital object identifier (DoI), website, journal name, pages, date of publication and volume etc., mainly available on first page.

In [43], the authors use CRF Chunker, because sometime author and organization names make ambiguities. So, they make a chunk of patterns for extraction. Same in case of number of pages, days, volume number make ambiguity and location names may mix with organization name. The scheme used to extract references' metadata from the last pages of research paper that include author, title, date, pages, location, organization, journal, book title, publisher, website.

In [44], the authors used Data Mining tool called Rapid Miner and SVM for Classification. The technique extracts metadata from research articles include a Title, Abstract, Keywords, an Author Name, Affiliation, Email, and Address. The information was extracted from the first page of a research paper only. Also, the scheme is limited to two department papers dataset only.

In [45], a CRF based method for extraction of citations from Bio-Medical papers that includes Title, Author, Source, Year and Volume is discussed. However, the technique focused on limited fields. In [46], a structural SVM classification technique is used to extract references metadata including citation number, authors, title, journal, volume, year, and page numbers from MEDLINE society. Experimental results showed that the approach performed better than the normal SVM and CRF based techniques in terms of extraction accuracy. In [47], authors investigated Hidden Markov Model (HMM) with the Viterbi algorithm. But the scheme is limited to references metadata extraction. In [48], authors proposed long-short term memory (LSTM) based deep learning approach for references extraction from the articles.

Table 1 summarizes the above state-of-the-art techniques for information extraction including their limitations, approach used and dataset information. The observed technical limitations to the existing schemes can be summarized as follows:

a. Their effectiveness to the type (simple like title or complex like authors affiliation etc.)
b. Amount of information being extracted (less information less issues and vice versa. Mainly schemes only focus on limited information like available on first or last page of the paper)

TABLE 1. Summary of related work review.

No.	Technique	Information Type	Technical Limitations	Dataset
[39]	PDFBox and PAXAT	Metadata of Scientific Articles that includes Title, Author Name & its Affiliation.	Only extract title, authors & its affiliation	ACM, IEEE, SPRINGER & ArVix. Total Articles 10,177.
[28]	K-Mean, Docstrum Algorithm and SVM. CERMINE	Metadata of Scientific Articles that includes Title, Author, and its Affiliation, Abstract, Keywords, Journal Name, Volume, Issue, Page, year, DOI, References and Complete Body of Article.	It is a PDF to Word Conversion tool that translates line by line. Not meant for extracting the structure of the article	CORA Datasets, PMC, GROTOAP, GROTOAP2, Cite Seer and PubMed. Total 1160 journals.
[40]	PDFBox, TET, PDF2TET and Table Seer Algorithm	Tabular data of Articles.	Only tabular data extraction	26 Wall Street journal's articles, Financial statements, UW-III. Total Articles 22 consist of 110 pages with 215 tables etc.
[41]	PDFBox and PARSCIT Algorithm	Reference Metadata that includes date, author, title, volume, pages, source, editor, issue, publisher, and location.	Confined to reference metadata extraction	PubMed Articles. Total Articles 10,000.
[42]	CRF, HMM and PSO	Metadata from the header of the paper that include title, author, address, summary, unit, email, date, abstract, telephone number, keyword, URL, degree, ISSN and extract the metadata from the last page i.e., from references.	Extract mainly information from first and last pages of the paper only	CORA Datasets. Total Articles 1500.
[43]	CRF Chunker	Reference metadata that include author, title, date, pages, location, organization, journal, book title, publisher, and website.	Mainly focused to reference metadata extraction	CORA Dataset, FLUX-CIM and CS-SW. Total Reference Strings 1176
[44]	Rapid Miner and SVM	Metadata of Scientific Articles that includes title, Author, Author Affiliation, Email, Address, Keywords and Abstract.	Extract information from only first page of research paper and are limited to only two Departmental papers as Datasets	Publications of two departments i.e., Department of Mathematics and Faculty of Sciences. Total Articles 100.
[45]	CRF	Citations from Bio-Medical Papers includes Title, Author, Source, Year and Volume.	Author is limited to just Bio-Medical Domain and extract only Citations	PubMed Central (PMC). Total Articles 672. Total numbers of Citation are 27606.
[46]	SVM and CRF	Reference Metadata that include Citation Number, Author, title, Journal, Volume, Year, and Pagination.	Single style reference parsing of MEDLINE Articles	MEDLINE Datasets. Total Articles 7000.
[47]	HMM and Viterbi Algorithm	Reference Metadata having 13 fields including Author Name, Title, Date, Editor, Book Title, Journal, Volume etc.	Extract the Reference Metadata only	CORA Datasets. Total References 500.
[48]	LSTM, Deep Learning, ParsCit, CRF, Word2Vec	Reference Parsing that includes, Author Name, Pages, Title, Date, Editor, Book Title, Journal, Volume etc.	Extract Citation of the document only	CORA Datasets. Total References 500.

c. Targeted corpus/structure (schemes perform good for coherent structure of the document and fail on the diverse structures and/or with missing/modified information)
d. Variations in the input document format (text, word, XML etc.). Mainly schemes target XML as input document type.
e. Post correction of extracted information (mainly schemes are confined to the extracted information and do not post-process or correct it).

Based on the discussion above and the comprehensive literature review, this study focuses on a novel ontology-based framework for information extraction (OFIE) from scientific documents available in PDF and DOC/DOCX formats, equipped with FRBS and WSD to investigate the accuracy with existing techniques based on empirical results.

Table 2 contrasts the proposed scheme and state-of-the-art schemes (selected for comparison in the results section) in terms of:

- Information type (structural components) being extracted.
- Type of data set
- Type of input documents (online published articles)
- Format of input documents (XML/DOC converted PDF documents)

TABLE 2. Schemes selected for comparison with information fields.

	Proposed OFIE	CERMINE[28]	PDF-Extract[13]	PAXAT[39]	ParsCit[41]
Title	✓	✓	×	✓	✓
Authors' names	✓	✓	✓	✓	✓
Authors' email	✓	✓	✓	✓	✓
Authors' affiliation	✓	✓	✓	✓	✓
Abstract	✓	✓	✓	×	×
Keywords	✓	✓	✓	×	×
Section headings	✓	×	✓	×	×
Sections sub-headings	✓	×	✓	×	×
Figure number with caption	✓	×	✓	×	×
Table number with caption	✓	×	✓	×	×
Acknowledgement	✓	×	×	×	×
References	✓	✓	×	✓	×

III. PROPOSED ONTOLOGICAL FRAMEWORK FOR INFORMATION EXTRACTION(OFIE)

In this section, we present the proposed ontological framework for information extraction (OFIE) using fuzzy rule-base system (FRBS) and word sense disambiguation (WSD). In the proposed approach, a backend ontology, based on which the entire IE framework is defined, will work as structural criteria for information extraction. Also, fuzzy regular expressions are used to address the variations in the extracted tokens as an extension to simple regular expressions. This makes the proposed scheme robust in terms of information extraction from the documents with the formats different from those considered in the examples during training phase. Furthermore, instead of working on just one type of document conversion, multiple document conversions namely, XML, Text and Word are processed under a separate rule for each conversion. Without loss of generality, the rules carved for XML are not applicable to text and doc conversion and vice versa. That is why we use Word2Vec to achieve word sense disambiguation to derive more accuracy in the conversion process.

Figure 1 depicts the proposed the schematic of the system architecture for information extraction from the scientific documents. For clarity, the architecture is divided into four phases. Detail of each phase is given subsequently. In the following subsections, we will describe each component of the proposed approach in detail.

A. PDF DOCUMENT CONVERSION

The scientific repositories currently maintain research articles in a PDF format. The first step is to convert the PDF format to XML, plain text and DOC/DOCX formats. The main purpose for converting the source documents (articles) into more than one formats is to facilitate fine-grained information extraction from the documents. The main purpose for converting the source documents (articles) into more than one formats is to facilitate fine-grained information extraction from the documents. Our analysis of many documents that include PDF, XML, plain text and DOC/DOCX formats regarding ease of information extraction led us to conclude that some part of information may be better extracted from XML, plain text, and/or from DOC/DOCX. For example, structural information such as title and author details can be adequately extracted from XML document. Similarly, figures and tables details can be extracted from text/docx documents which are better than XML.

Further, in this phase, unnecessary details, omission, data cleansing and other type of pre-processing is performed. In this phase different parts of the research papers like title, authors, funding agency/ acknowledgement are identified using different rules (mainly in the form of regular expressions) and the rest of the tokens are discarded. For example, in this case, we are not interested in publication year, journal ISSN and paper's main text etc. so such information can be filtered out. First, the given PDF document is converted

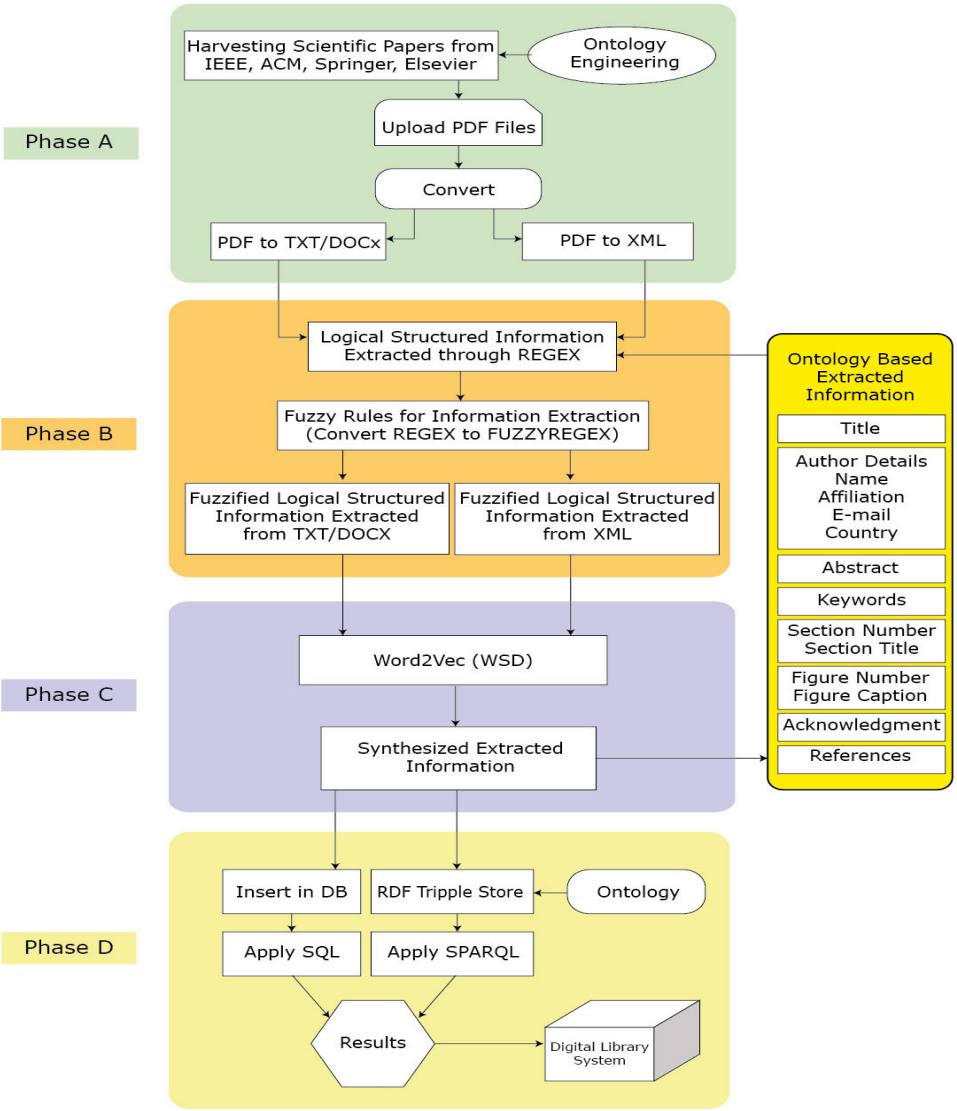

FIGURE 1. The proposed OFIE architecture.

to XML, plain text and DOC/DOCX formats. Purpose for converting the source documents (articles) into more than one formats is that each conversion has its own pros and cons and from the experiments, it is observed that some part of information may be better extracted either from XML or plain text, or DOCX. To compliment, all the formats are used. Mainly, XML provides a better extraction support due to its tag nature, DOC/DOCX provides rich text format (RTF) features like fonts, headings and numbering etc. for a better understanding and text provide plain text in a better way like the abstract, acknowledgement part etc.

B. STRUCTURAL INFORMATION EXTRACTION

In this phase, the structural information is extracted from the documents converted in Phase A. This task is mainly carried out by carving the rules (patterns) in the form of regular expressions (REGEX) to detect the desired token from the document text and extract the relevant information. However, in case of REGEX, it looks for an exact match between the pattern and the text token. In case of mismatch, document crawler is unable to extract the desired information. To overcome this issue, the REGEX rules are converted into Fuzzy regular expressions (FREGEX). In case of FREGEX, it looks for an approximate match between pattern and text token. That approximation error is calculated by Levenshtein distance formula that equates two strings even in case of deletion, insertion, and substitution of characters in the text token [49]. The distance formula calculates the error which is roughly equal to count of mistakes divided by average between length of pattern and length of text token. Table 3 shows an example of fuzzy match

TABLE 3. Error measure in FREJ.

Error	Text Token	Type of Error
0	abstract	None
2	absrac	Deleted
1	aabstract	Inserted
1	absrpact	Modified

between pattern ''abstract'' and some erroneous text tokens containing deleted, inserted, and modified characters or substrings [50].

1) FUZZY RULE BASED SYSTEM

Generally, the documents in the dataset repository have varying formats as described in the dataset subsection. By closely observing the structure of these documents (belong to various societies), REGEX are carefully carved and converted to equivalent FREGEX. In this regard, it is very important to figure out the appropriate FREGEX (index) upon detecting the format. Similarly, several types of information are being extracted from title to bibliography. This is referred to as structural index (SI) that specifies whether the given text token is title, abstract, keyword etc. The third parameter in this regard is tolerance (T) that specifies the extent to which the distance between pattern and text token can be tolerated. It is worth mentioning here, that high tolerance does not always mean error is avoided. In some cases, more tolerance can result in poor detection and/or accuracy.

Based on these experiments, a bank of FREGEX for XML, text/docx pattern against various formats is created. To automate the process and to get the appropriate tolerance against given society, and SI, the FRBS is designed in MATLAB. There are two input variables namely, society (Soc), and Structural Index (SI) while there is one output variable and Tolerance (T). This relationship is shown in Figure 2. This is mainly because information extraction from different societies (IEEE, ACM etc.) exhibits different number of errors due to variations in structural components (title, name etc.), as observed by experimentation.

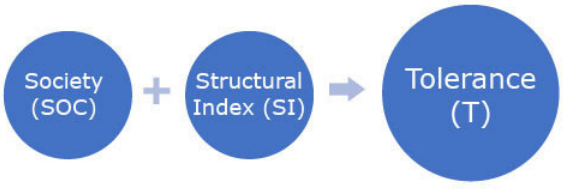

FIGURE 2. Schematic of fuzzy system.

After several experiments on extracting information from several societies with several formats, following observations were made (heuristics).

1. In one society information against a certain component extracted without any error or single error.
2. In second society information against same component extracted with a greater number of errors.
3. After several experiments, these errors were obtained again each society and its structural components and consequently averaged. The average error can be written mathematically as,

$$E_{avg} = \frac{1}{S.I} \sum_{s} \sum_{i} e_{s.i} \tag{1}$$

where S is total number of societies, I is total number of structural indices in a paper and $e_{s.i}e_{s.i}$ is error against a particular society and its structural index.

1) Heuristically, during investigating plain regular expressions, it was observed that mainly the average error ($E_{avg}E_{avg}$) falls in the range 0 to 4.
2) That error is used as the tolerance variable determined by FRBS for fuzzy regular expression used against the given society and structural index.

To fine tune the performance by tolerating the average error, a FRBS is designed to estimate the exact tolerance being used by the fuzzy regular expression.

The sample rule can be expressed as:

IF (SOC = 'index' AND SI = 'index') THEN (AND T = V. High)

There are four main components of FRBS. Namely fuzzifier, defuzzifier, inference engine and the rule base. Here, we have used Triangular Fuzzifier, Center Average Defuzzifier (CAD) and Mamdani Inference Engine (MIE) [51].

C. WORD SENSE DISAMBIGUATION

As shown in Figure 1, this module is responsible for synthesizing and fine tuning the extracted information from XML and text/docx converted versions in Phase B. Among the primary challenges in structural information extraction is information loss that occurs while converting the input document from one format to another. Many PDF to text, xml and word conversion libraries result into errors during the phase of conversion. These errors tend to affect the performance of extraction task. To mitigate such type of errors, WSD is a necessity and a value addition in the proposed technique to improve the accuracy in the overall information extraction process. In the proposed scheme, WSD is performed using word to vector (word2vec) approach that is a Neural Network based similarity model where the words or the concepts are represented in terms of n-dimensional vectors in a huge vector space [24]. The model performs autocorrection of misspelled words, word sense correction and sentence sense/sequence correction by augmenting the two streams duly extracted from same document converted in XML and text/docx formats. The process is shown in Figure 3. Upon receiving both streams, it creates the order/semantic vector with the help of lexical database and corpus. Consequently, it performs word order/ semantic similarity and after applying the sentence similarity it generates the final synthesized output stream.

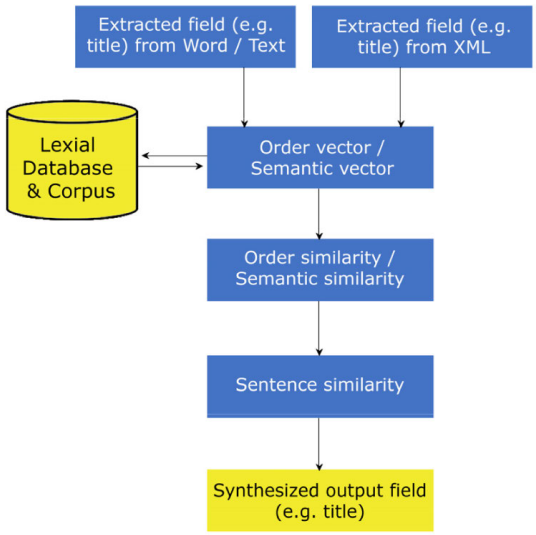

FIGURE 3. Word Sense disambiguation.

D. ONTOLOGY AND DIGITAL LIBRARY

As depicted in Figure 1, this module stores, manages and retrieve all the extracted structural information from the scientific documents. Ontologies are meant to comprehend and synthesized the information in a particular field [25]. For example, authors, title, sub-title, sections, and other required information. The ontology is designed/engineered in Protégé software. Figure 4 shows the proposed ontology of the scientific documents. After successful extraction, all the information is sent to the RDF (Resource Description Framework) generator block [25]. The block generates the subject, the predicate (property), and the object. The relationship of the three is called a triple. All the triples are stored in a triple store, where the SPARQL queries can be applied to search the results. As an alternate approach, the extracted information can be inserted into a Relational Database and searched/retrieved by SQL queries instead. Moreover, it can be exported as comma separated values (CSV) and accessed in MS Excel.

Consequently, the ontology is mapped to the digital library for utilization of the extracted information [18], [19].

IV. PERFORMANCE ANALYSIS

This section presents the implementation of the proposed scheme and analysis of the results. Performance metrics along with the analysis of the results are discussed. We also compared the proposed work against the approaches proposed in [10], [13], [28]. The reason behind selection of these techniques is stated in the ''Related Work'' and Table 2.

A. IMPLEMENTATION

We have implemented the proposed approach and table 4 shows the software packages and libraries used in the implementation. Figure 5 shows the main screen of the prototype.

FIGURE 4. Proposed Ontology of research Paper.

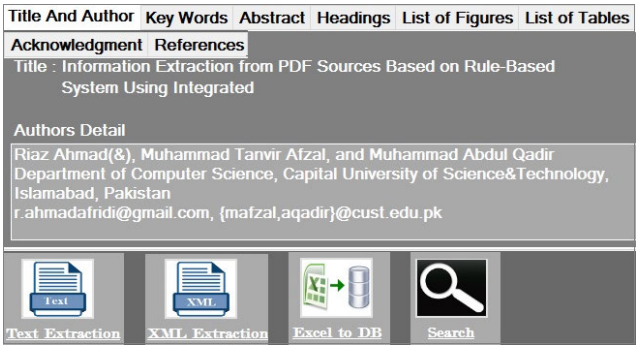

FIGURE 5. Prototype main screen.

TABLE 4. Implementation detail.

Sr	Item	Value
1	Programming Languages	C#, Java, Python
2	Packages/Libraries used	• REGEX • FUZZY REGEX in Java (FREJ) • Fuzzywuzzy
3	Fuzzy Error Types	The 3 types of error are: • Insertion (i) • Deletion (d) • Substitution (s)
4	Error tolerance range	[1-4]
5	Error measure	Levenshtein distance aka edit-distance measure[49]
6	Information type	Specified by the ontology given in Fig. 4
7	WSD method	Word2Vec

The files are uploaded and all necessary information such as author name, email, affiliation, and other contents are extracted. Generation of text and docx from PDF writer is also automated process but for clarification and differentiation purposes, separate buttons are provided in the prototype. To reveal the information extracted from the tabs, the extracted information is categorized (in tabs) as Title and Author, Keywords, Headings, List of Figures (captions), List of Tables (captions), References and Acknowledgement.

In the current view, the extracted Title and Author names are displayed. Here Fuzzy REGEX (regular expression) augmented with word sense disambiguation, helps in precisely identifying the author names. As different societies have different ways to write the author names like first and last name, first second and last name and so on, there was a dire need to have an approach to disembogue this issue. Further, it shows the set of extracted keywords for the same published document. The Fuzzy REGEX rules obtain all variations of keywords styles associated with various publishing societies. Different societies use different characters to separate the keywords like comma, semi-colon, line separation etc. In the shown example, comma separation is detected. Similarly, it shows the extracted abstract from the sample paper.

Next, all the extracted Headings of the research paper including main and sub-sections are depicted. This information can be used for generating table of contents against the document for better indexing, search, and retrieval. Here Fuzzy regular expression was helpful in identifying the main and sub-sections of any depth like section number 2.1.2.5.6. Further, it shows the extracted list of figures' captions in the input scientific document. The documents may contain arbitrary number of figures and the proposed scheme can extract all the figures' captions precisely. Again, various societies have different ways to present the figure captions, the proposed scheme not only extract the figures captions for the underlying four societies (ACM, IEEE, Elsevier, and Springer) but other societies' papers can also be treated with the significant accuracy. Here Fuzzy REGEX can precisely unify the tokens like ''Figure'', ''Fig'' etc. followed by a '.', ':' and/or space.

Similar, approach is used for Table captions (next tab). Similarly, it shows the extracted captions of the tables in the paper. The tables' caption formats vary from society to society like word 'table' or 'tab', single line or multi-line caption text, and various types of table numbering styles etc. Likewise, it shows the extracted acknowledgement section of the paper. It is normally an optional part in many societies. It comes under different heading like 'special thanks', 'funding agency' and 'acknowledgement' while sometimes both 'funding agency and acknowledgement' etc. This section carries the authors' acknowledgements to a person, an organization, and/or the funding agency etc. In case the paper does not contain information, a null string will be returned. The length of acknowledgement section can be arbitrary in this experiment. Most of the societies provide this section and then it is up to the authors whether they utilize it or not. Finally, it shows the extracted references from the. It is well-known fact there are several ways for references. We have addressed all the well-known as well as the customized reference styles in the FREJ, like APA, IEEE etc. The scheme is robust against several standard reference styles used in the target societies as well as many others in the literature. Figure 6 shows the extracted information from a sample document in a summarized form, duly collected from the tabs shown in Figure 5.

B. EVALUATION METRICS

The performance is measured in terms of three metrics: precision (P), recall (R) and F-measure (F). Same performance metrices have been used in previous works [10], [13], [28]. These performance metrics are defined as follows [13], [28]:

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$F - measure = \frac{2PR}{P + R} \tag{4}$$

where TP refers to the true positives or the number of rows to which the scheme correctly assigned the category it recognizes (Title, Authors etc.). In the case when that category is incorrectly assigned a false positive (FP) is generated, while FN (false negatives) represents the number of rows for which the scheme was not able to recognize the category it was constructed for.

C. DATASET

For the evaluation of the proposed scheme, we used 500 published papers composed of journals, conferences, etc. sourced from several scientific repositories. Table 5 shows the dataset along the various sources and distribution. For instance, the variations of 12 ACM data sources are given in Figure 7. The data set was divided for training and testing phases as 70% and 30% division, respectively. Also, the CORA dataset is used.

TABLE 5. Dataset used in the evaluation.

Sr	Repositories	Variations	Papers
1	IEEE	31	100
2	ACM	12	100
3	Elsevier	37	100
4	Springer	50	100
5	Other (Hindawi, MDPI, BMC, PubMed, PLOS etc.)	70	100
	Total	200	500

D. COMPARISON

In this section the proposed approach is compared with similar approaches in the literature. Two types of comparisons are made, that are quantitative (in terms of numbers and experimental outputs) and qualitative (in terms of the features and dimensions of extraction fields).

To make the comparison fair, we evaluated the proposed scheme and the other schemes using the datasets used in [10], [13], [28]. Table 6 shows the precision (in percentage) in training and testing phases for various sections of the paper separately. At the end aggregate averages of all the sections is calculated which are 89.14% and 91.21%, respectively for testing and training phases, respectively.

1) QUANTITATIVE COMPARISON (EMPIRICAL)

The comparison is made in terms of precision, F-score, and recall, and it is given in Table 7. The proposed scheme is better

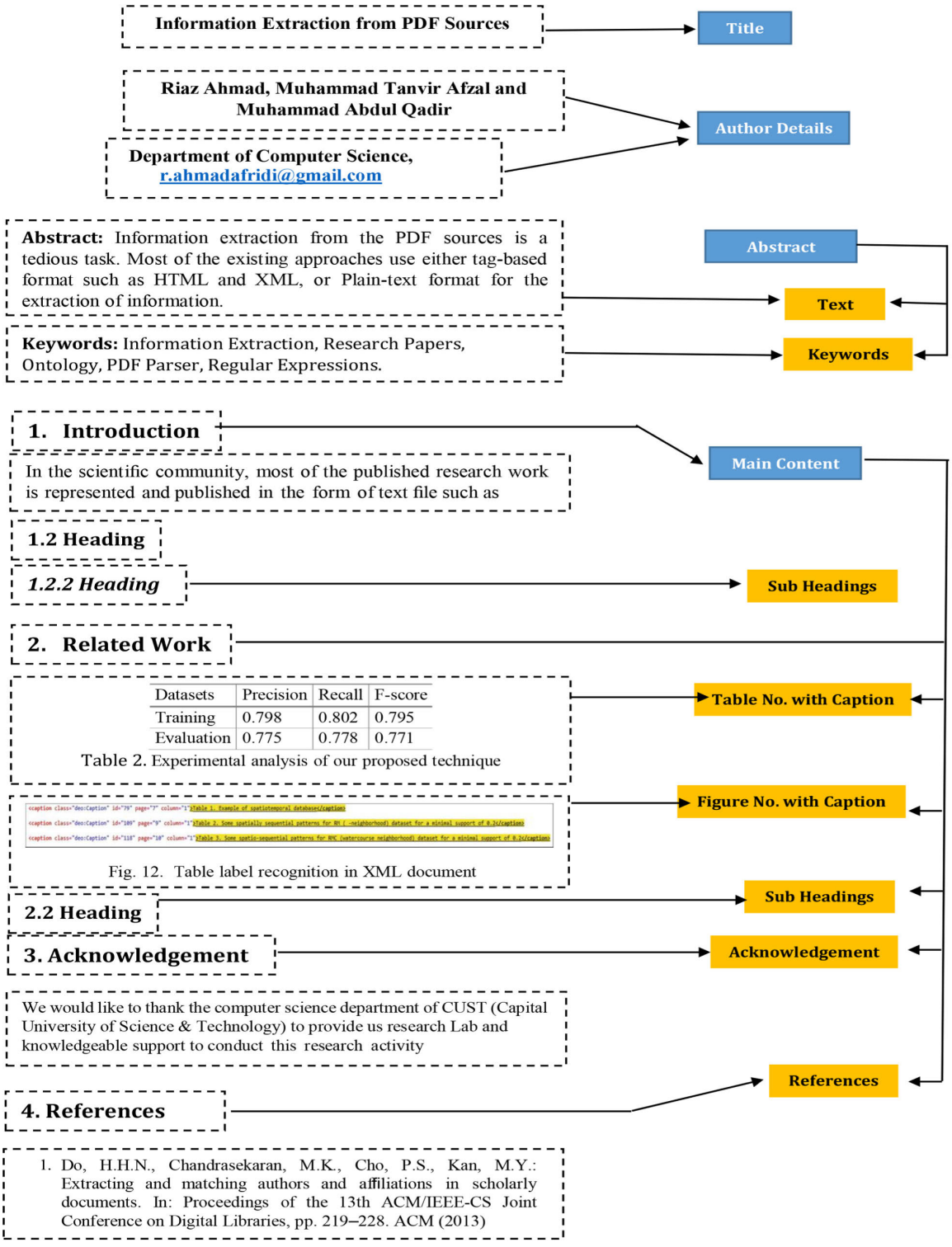

FIGURE 6. Structural information extracted from an example article.

in terms of precision, recall and F-score. However, the scheme is [10] has a performance closer to the proposed scheme but it is worth mentioning here, that it only focuses on tabular information extraction rather that whole paper's information. Nonetheless, the schemes in [13] and CERMINE [28] are like the proposed one in terms of type of information to be

TABLE 6. Section based precision in testing and training phases.

Sr.	Section	Precision (Testing)	Precision (Training)
1	Title and Author details (name, affiliation, email)	83%	84.5%
2	Keywords	95%	97%
3	Headings	88%	90%
4	References	92%	93.5%
5	Acknowledgement	92%	93.5%
6	List of Figures	87%	90%
7	List of Tables	87%	90%
	Overall (average)	**89.14%**	**91.21%**

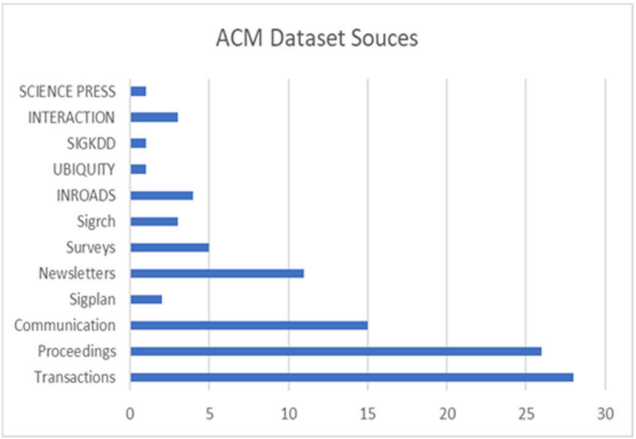

FIGURE 7. ACM dataset sources.

TABLE 7. Comparison based on overall extraction results.

Dataset	Precision	Recall	F-Score	Scheme
Training	0.798	0.802	0.795	[13]
Training	0.89	0.88	0.86	[10]
Training	**0.912**	**0.916**	**0.901**	**OFIE**
Testing	0.775	0.778	0.771	[13]
Testing	0.76	0.9	0.86	[10]
Testing	0.81	0.747	0.775	CERMINE[28]
Testing	**0.8914**	**0.896**	**0.890**	**OFIE**

extracted. It is apparent that the proposed scheme performs significantly better in terms of precision, recall and F-score, compared to all the schemes.

Since the scheme in [28] provides granular level precision values for the components like title, name etc., a detailed comparison is given in Table 8. The proposed scheme outperforms for several sections like title, authors' names, authors' affiliation, authors' email, and keywords. However, for references/bibliography section the scheme in [28] has 4.1% better accuracy. Moreover, the sections like headings, acknowledgement, figures, and tables are not covered in [28]. Moreover,

TABLE 8. Comparison with [28] for sub-sections.

Sr.	Section	OFIE	CERMINE[28]
1	Title	**96%**	95.5%
2	Author name	**91%**	90.2%
3	Affiliation	**90%**	88.2%
4	Email	**55%**	51.7%
5	Keywords	**95%**	89.9%
6	Headings	88%	X
7	References	92%	**96.1%**
8	Acknowledgement	92%	X
9	List of Figures	87%	X
10	List of Tables	87%	X

FIGURE 8. Comparison in training phase.

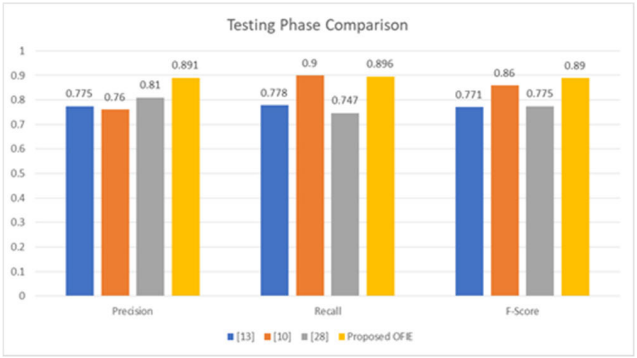

FIGURE 9. Comparison in testing phase.

in [28], authors are intended to extract the entire article text but for the comparison purpose we only consider the common parts. Similarly, the entries marked as 'X' are not covered in [28] that are headings, acknowledgement, list of figures and list of tables. Moreover, the [28] was compared on the same dataset.

Figure 8 and Figure 9 show the comparison of the schemes in training and testing phases, respectively. It is comprised of the values given in Table 7. In Figure 9, the recall value of [10] is slightly better than the proposed scheme in testing phase. However, in terms of precision and F-score, proposed scheme outperforms in both testing and training phases in all schemes.

TABLE 9. Comparison.

Parameters	[10]	[13]	[28]	OFIE
Target Document Type	Scientific Publications	Scientific Publications	Scientific Publications	Scientific Publications
Scope of document/ Data set	Industrial Reports	ESWC Conference papers	Published articles in various domains like PubMed, CORA etc.	Published articles of IEEE, ACM, Elsevier, Springer and others
Dataset size	48 papers	95 papers	1160 papers	500 papers
Information to be extracted	Ontology based Tabular Data Extraction only	Ontology based paper logical structure	Metadata, Full-text, Bibliography and XML equivalent document	Ontology based paper logical structure + keywords and bibliography
Proposed Technique	Ontology based Extraction	Heuristic based	Supervised and Unsupervised Machine Learning	Ontology, Fuzzy System and WSD

2) QUALITATIVE COMPARISON

This part contains the qualitative comparison of the proposed scheme with the previous schemes. The comparison is given in Table 9. From the comparison, it is apparent that the proposed scheme is superior to [10] in terms of obtaining the structural information from the document. Similarly, the proposed scheme is superior to the approaches proposed in [10] and in [13] in terms of the target range and diversity of the document formats and volume of information extracted. As far as scheme in [28] is concerned, it is mainly focusing on metadata, full text of the document, complete breakdown of the bibliography part (title, name, volume, publishing venue, pages, and years etc.). In contrast to the proposed scheme where we are mainly interested in structural information (title, author, details, table of contents, references and list of figures and tables) which is more than just metadata. However, full-text extraction is not part of the scope of the proposed scheme. That is why, Table 6 contains a separate comparison with [28] section-wise details, table of contents, references and list of figures and tables) which is more than just metadata. However, full-text extraction is not part of the scope of the paper.

E. SUMMARY

Based on the comprehensive empirical analysis, it is evident that the proposed framework outperforms several techniques in terms of average accuracy, amount of information extraction and diversity of the dataset. In the framework, however, there are many components like fuzzy regular expressions and WSD etc. each playing their role toward the aggregate success. It is somewhat hard to describe their individual role towards the overall accuracy of the system. However, since the approaches are working in a sequential way, so it is apparent that the former component's efficacy will strengthen the next component as so on. For instance, if the fuzzy regular expression does not provide the due input to the WSD, it will not be able to fine tune the accuracy. Nonetheless, on a nutshell, intuitively, it can be safely stated that the system's accuracy is mainly enriched by the fuzzy regular expressions followed by the WSD semantic and syntactical fine tuning.

V. CONCLUSION AND FUTURE WORK

This paper proposed an ontological framework for information extraction, repository and retrieval using Fuzzy rule base and word sense disambiguation. The Fuzzy Rule Base transforms the plain regular expression into fuzzy regular expressions with a tunable tolerance in terms of insertion, deletion, and substitution errors in the pattern. Four research societies are targeted for the information extraction, though the proposed scheme works significantly well for the other societies as well. Once the information is extracted, it is transformed into an RDF object and stored in the compatible RDF triple store for efficient retrieval. The proposed approach can be a great help in building the digital libraries supported with an automatic ETL (extraction, transformation, and loading) process. The proposed scheme is promising in terms of computational complexity as well as accuracy. In future, the scheme may be extended to involve machine learning for automated information extraction that can encompass wider number of societies and corpuses.

It is worth investigating machine learning and evolutionary computing techniques especially their hybrid counterparts for sake of information extraction especially in the domain of scientific (Wiley, BMC etc.) research under the proposed framework. Further experiments that include scientific publication from other societies is part of our future work. The information extraction process, in practice as well as in the proposed approach is assumed to be a backend/offline process where there are several documents are downloaded and information is extracted. However, it is important to also consider the difficulty component of the solution for a digital library and information system in real time. In future, this factor may also be assessed for the existing as well as the upcoming schemes.

ACKNOWLEDGMENT

The help of Maliha Omar, Mohib Ullah Khan and Umar Farooq is greatly appreciated.

REFERENCES

[1] K. Shaalan *et al.*, Eds., *Intelligent Natural Language Processing: Trends and Applications* (Studies in Computational Intelligence), vol. 740. Springer, 2018, doi: 10.1007/978-3-319-67056-0_18.
[2] F. Peng and A. McCallum, "Information extraction from research papers using conditional random fields," *Inf. Process. Manage.*, vol. 42, no. 4, pp. 963–979, Jul. 2006.
[3] *IEEE Xplore*. Accessed: Jan. 2, 2020. [Online]. Available: https://ieeexplore.ieee.org/Xplore/home.jsp
[4] *RELEX. Annual Report*, R. Group, New York, NY, USA, 2017.
[5] D. Tkaczyk, "New methods for metadata extraction from scientific literature," 2017, *arXiv:1710.10201*. [Online]. Available: http://arxiv.org/abs/1710.10201

[6] X. Ma, H. Qin, N. Sulaiman, T. Herawan, and J. H. Abawajy, "The parameter reduction of the interval-valued fuzzy soft sets and its related algorithms," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 1, pp. 57–71, Feb. 2014.
[7] J. Azimjonov and J. Alikhanov, "Rule based metadata extraction framework from academic articles," 2018, *arXiv:1807.09009*. [Online]. Available: http://arxiv.org/abs/1807.09009
[8] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp, "Evaluation of header metadata extraction approaches and tools for scientific PDF documents," in *Proc. 13th ACM/IEEE-CS Joint Conf. Digit. Libraries (JCDL)*, 2013, pp. 385–386.
[9] G. Zaman, H. Mahdin, and K. Hussain, "Information extraction from semi and unstructured data sources: A systematic literature review," *ICIC Express Lett.*, vol. 14, no. 6, pp. 593–603, Jun. 2020.
[10] S. T. R. Rizvi, D. Mercier, S. Agne, S. Erkel, A. Dengel, and S. Ahmed, "Ontology-based information extraction from technical documents," in *Proc. 10th Int. Conf. Agents Artif. Intell.*, 2018, pp. 493–500.
[11] Atta-ur-Rahman, I. M. Qureshi, A. N. Malik, and M. T. Naseem, "QoS and rate enhancement in DVB-S2 using fuzzy rule based system," *J. Intell. Fuzzy Syst.*, vol. 30, no. 2, pp. 801–810, Feb. 2016.
[12] Atta-ur-Rahman, I. M. Qureshi, A. N. Malik, and M. T. Naseem, "Dynamic resource allocation in OFDM systems using DE and FRBS," *J. Intell. Fuzzy Syst.*, vol. 26, no. 4, pp. 2035–2046, 2014.
[13] R. Ahmad, M. T. Afzal, and M. A. Qadir, "Information extraction from PDF sources based on rule-based system using integrated formats," in *Semantic Web Challenges. SemWebEval* (Communications in Computer and Information Science), vol. 641, H. Sack, S. Dietze, A. Tordai, and C. Lange, Eds. Cham, Switzerland: Springer, 2016. [Online]. Available: link: https://link.springer.com/chapter/10.1007/978-3-319-46565-4_23, doi: 10.1007/978-3-319-46565-4_23.
[14] K. Jayaram and K. Sangeeta, "A review: Information extraction techniques from research papers," in *Proc. Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, Feb. 2017, pp. 56–59.
[15] Z. Bodó and L. Csató, "A hybrid approach for scholarly information extraction," *Studia Univ. Babes-Bolyai, Inform.*, vol. 62, no. 2, pp. 5–16, 2017.
[16] P. Groth, M. Lauruhn, A. Scerri, and R. Daniel, "Open information extraction on scientific text: An evaluation," 2018, *arXiv:1802.05574*. [Online]. Available: http://arxiv.org/abs/1802.05574
[17] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data*, vol. 34, no. 1. Cambridge, U.K.: Cambridge Univ. Press, 2007, pp. xii and 410.
[18] M. Ahmad and J. H. Abawajy, "Digital library service quality assessment model," *Procedia-Social Behav. Sci.*, vol. 129, pp. 571–580, May 2014.
[19] M. Safar, "Digital Library of Online PDF Sources: An ETL Approach," *IJCSNS*, vol. 20, no. 11, p. 173, 2020.
[20] M.-T. Luong, T. D. Nguyen, and M.-Y. Kan, "Logical structure recovery in scholarly articles with rich document features," in *Multimedia Storage and Retrieval Innovations for Digital Library Systems*. Hershey, PA, USA: IGI Global, 2012, pp. 270–292.
[21] H. H. N. Do, M. K. Chandrasekaran, P. S. Cho, and M. Y. Kan, "Extracting and matching authors and affiliations in scholarly documents," in *Proc. 13th ACM/IEEE-CS Joint Conf. Digit. Libraries (JCDL)*, 2013, pp. 219–228.
[22] S. Kim, Y. Cho, and K. Ahn, "Semi-automatic metadata extraction from scientific journal article for full-text XML conversion," in *Proc. Int. Conf. Data Sci. (ICDATA)*, 2014, p. 1.
[23] A. Abd-Rashid, S. Abdul-Rahman, N. N. Yusof, and A. Mohamed, "Word sense disambiguation using fuzzy semantic-based string similarity model," *Malaysian J. Comput.*, vol. 3, no. 2, pp. 154–161, 2018.
[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[25] D. Allemang and J. Hendler, *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Amsterdam, The Netherlands: Elsevier, 2011.
[26] J. Chen, C. Zhang, and Z. Niu, "A two-step resume information extraction algorithm," *Math. Problems Eng.*, vol. 2018, pp. 1–8, May 2018, doi: 10.1155/2018/5761287.
[27] R. Shah and S. Jain, "Ontology-based information extraction: An overview and a study of different approaches," *Int. J. Comput. Appl.*, vol. 87, no. 4, pp. 6–8, Feb. 2014, doi: 10.5120/15194-3574.
[28] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski, "CERMINE: Automatic extraction of structured metadata from scientific literature," *Int. J. Document Anal. Recognit.*, vol. 18, no. 4, pp. 317–335, Dec. 2015.
[29] T. M. Dieb, M. Yoshioka, S. Hara, and M. C. Newton, "Framework for automatic information extraction from research papers on nanocrystal devices," *Beilstein J. Nanotechnol.*, vol. 6, no. 1, pp. 1872–1882, 2015.
[30] X. Li, "The comparison of QlikView and tableau: A theoretical approach combined with practical experiencies," M.S. thesis, Dept. Bus. Econ., Masters Manage., Univ. Hasselt, Hasselt, Belgium, 2014.
[31] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: An overview from a database perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 6, pp. 866–883, Dec. 1996.
[32] S. Jebbara and P. Cimiano, "Aspect-based sentiment analysis using a two-step neural network architecture," in *Semantic Web Challenges. SemWebEval* (Communications in Computer and Information Science), vol. 641, H. Sack, S. Dietze, A. Tordai, and C. Lange, Eds. Cham, Switzerland: Springer, 2016. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46565-4_12, doi: 10.1007/978-3-319-46565-4_12.
[33] S.-T. Kousta, D. P. Vinson, and G. Vigliocco, "Emotion words, regardless of polarity, have a processing advantage over neutral words," *Cognition*, vol. 112, no. 3, pp. 473–481, Sep. 2009.
[34] S. Sun, G. Kong, and C. Zhao, "Polarity words distance-weight count for opinion analysis of online news comments," *Procedia Eng.*, vol. 15, pp. 1916–1920, Dec. 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877705811018583?via%3Dihub
[35] A. Agarwal, F. Biadsy, and K. R. Mckeown, "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams," in *Proc. 12th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2009, pp. 24–32.
[36] A. C. E. S. Lima, L. N. D. Castro, and J. M. Corchado, "A polarity analysis framework for Twitter messages," *Appl. Math. Comput.*, vol. 270, pp. 756–767, Nov. 2015.
[37] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2004, pp. 168–177.
[38] A. Krouska, C. Troussas, and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," in *Proc. 7th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, Jul. 2016, pp. 1–5.
[39] C. Jiang, J. Liu, D. Ou, Y. Wang, and L. Yu, "Implicit semantics based metadata extraction and matching of scholarly documents," *J. Database Manage.*, vol. 29, no. 2, pp. 1–22, Apr. 2018.
[40] S. Khusro, A. Latif, and I. Ullah, "On methods and tools of table detection, extraction and annotation in PDF documents," *J. Inf. Sci.*, vol. 41, no. 1, pp. 41–57, Feb. 2015.
[41] R. Kern and S. Klampfl, "Extraction of references using layout and formatting information from scientific articles," *D-Lib Mag.*, vol. 19, no. 9/10, Sep./Oct. 2013. [Online]. Available: http://www.dlib.org/dlib/september13/kern/09kern.html, doi: 10.1045/september2013-kern.
[42] Z. Shuxin, X. Zhonghong, and C. Yuehong, "Information extraction from research papers based on conditional random field model," *Telkomnika Indonesian J. Electr. Eng.*, vol. 11, no. 3, pp. 1213–1220, Mar. 2013.
[43] T. Groza, A. Grimnes, and S. Handschuh, "Reference information extraction and processing using random conditional fields," *Inf. Technol. Library*, vol. 31, no. 2, pp. 6–20, 2012.
[44] A. Kovačević, D. Ivanović, B. Milosavljević, Z. Konjović, and D. Surla, "Automatic extraction of metadata from scientific publications for CRIS systems," *Program*, vol. 45, no. 4, pp. 376–396, Sep. 2011.
[45] Q. Zhang, Y.-G. Cao, and H. Yu, "Parsing citations in biomedical articles using conditional random fields," *Comput. Biol. Med.*, vol. 41, no. 4, pp. 190–194, Apr. 2011.
[46] X. Zhang, J. Zou, D. X. Le, and G. R. Thoma, "A structural SVM approach for reference parsing," in *Proc. 9th Int. Conf. Mach. Learn. Appl.*, Dec. 2010, pp. 479–484.
[47] B. Ojokoh, M. Zhang, and J. Tang, "A trigram hidden Markov model for metadata extraction from heterogeneous references," *Inf. Sci.*, vol. 181, no. 9, pp. 1538–1551, May 2011.
[48] A. Prasad, M. Kaur, and M.-Y. Kan, "Neural ParsCit: A deep learning-based reference string parser," *Int. J. Digit. Libraries*, vol. 19, no. 4, pp. 323–337, Nov. 2018.
[49] K. U. Schulz and S. Mihov, "Fast string correction with Levenshtein automata," *Int. J. Document Anal. Recognit.*, vol. 5, no. 1, pp. 67–85, Nov. 2002.
[50] *Fuzzy Regular Expressions for Java—FREJ*. Accessed: Dec. 25, 2019. [Online]. Available: http://frej.sourceforge.net/javadocs/index.html
[51] Atta-ur-Rahman, S. Dash, A. K. Luhach, N. Chilamkurti, S. Baek, and Y. Nam, "A neuro-fuzzy approach for user behaviour classification and prediction," *J. Cloud Comput.*, vol. 8, no. 1, p. 17, Dec. 2019.

GOHAR ZAMAN is currently a Postgraduate Research Student with the Faculty of Computer Science and Information Technology (FSKTM), Universiti Tun Hussein Onn Malaysia. His research interests include information extraction, data mining, ontologies, NLP, and automatic text categorization.

HAIRULNIZAM MAHDIN (Member, IEEE) is currently an Associate Professor with the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. His current research interests include IoT and blockchain. He is a member of the Malaysia Board of Technologist (MBOT). He has been actively involved in many conferences internationally serving as conferences in various capacity, including chair, general co-chair, vice-chair, best paper award chair, publication chair, session chair, and program committee. He has also guest edited many special issue journals.

KHALID HUSSAIN joined Academia, in 2008, as a full-time Faculty Member. He is currently working as a Professor and the Dean Faculty of Computing, Barani Institute of Sciences Sahiwal. He is also working as a Campus Director in Burewla Campus. He has vast university/industry experience. During his tenure in the industry, he served in the defense related projects and in recognition of his services, he has been awarded commendation certificates by multiple government agencies. He has been involved in numerous research projects. He helped to setup a pioneer setup for information/network security certification in Pakistan. He also introduced EC Council certification under the first academia industry partnership. He did his Ph.D. from Malaysia, under a fully funded UTM/HEC scholarship. He published 63 articles. In which 27 are ISI Indexed Impact Factor, 13 are in HEC approved journal and 23 are in IEEE and ACM conferences. He also has a book chapter and three books with the title *Information Security Handbook* is going to publish in couple of months. He has successfully completed six applied research project in the domain of information security funded by NESCOM. Up till now 37 M.S. students completed his research thesis under his supervision. He is also supervising 13 M.S. and five Ph.D. student in which two Ph.D. students completed their Ph.D. He received the Gold Medal for his contribution towards Information Security SATHA, in 2015.

ATTA-UR-RAHMAN received the B.S. degree in computer science from the University of the Punjab, Lahore, Pakistan, in 2004, the M.S. degree in EE from International Islamic University, Islamabad, Pakistan, in 2008, and the Ph.D. degree in EE from ISRA University, Islamabad, in 2012. He is currently working as an Assistant Professor with the College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University (IAU), Dammam, Saudi Arabia. Since 2003, he has been involved in teaching and research. He has authored/coauthored more than 100 publications in conferences, books, and journals of good reputation. His research interests include digital communication, DSP, information and coding theory, AI, and applied soft computing.

JEMAL ABAWAJY (Senior Member, IEEE) is currently a Full Professor with the Faculty of Science, Engineering, and Built Environment, Deakin University, Australia. His leadership is extensive spanning industrial, academic, and professional areas. He is also the Director of the Distributing System Security (DSS). He is actively involved in funded research supervising a large number of Ph.D. students, postdoctoral researchers, research assistants, and visiting scholars in the area of cloud computing, big data, network and system security, decision support systems, and e-health. He is the author/coauthor of five books and ten conference volumes, more than 250 refereed articles in conferences, book chapters, and journals. He is a Senior Member of the IEEE Technical Committee on Scalable Computing (TCSC), the IEEE Technical Committee on Dependable Computing and Fault Tolerance, and the IEEE Communication Society. He has been actively involved in the organization of more than 200 national and international conferences in various capacity, including the Chair, the General Co-Chair, the Vice-Chair, the Best Paper Award Chair, the Publication Chair, the Session Chair, and a Program Committee Member. He has served on the Editorial Board of numerous international journals.

SALAMA A. MOSTAFA received the B.Sc. degree in computer science from the University of Mosul, Iraq, in 2003, and the M.Sc. and Ph.D. degrees in information and communication technology from the Universiti Tenaga Nasional (UNITEN), Malaysia, in 2011 and 2016, respectively. He is currently a Lecturer with the Department of Software Engineering, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM). His research interests include soft computing, data mining, software agents, and intelligent autonomous systems.