IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 9, No. 3, June, 2022, pp. 683-696
Journal homepage: http://iieta.org/journals/mmep

Assessment of Information Extraction Techniques, Models and Systems

Atta-ur Rahman[1*], Dhiaa Musleh[1], Majed Nabil[1], Haya Alubaidan[1], Mohammed Gollapalli[2], Gomathi Krishnasamy[2], Dakheel Almoqbil[3], Mohammad Aftab Alam Khan[4], Mehwash Farooqui[4], Mohammed Imran Basheer Ahmed[4], Mohammed Salih Ahmed[4], Maqsood Mahmud[5]

[1] Department of Computer Science (CS), College of Computer Science and Information Technology (CCSIT), Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
[2] Department of Computer Information Systems (CIS), College of Computer Science and Information Technology (CCSIT), Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
[3] Department of Networks and Communications (NC), College of Computer Science and Information Technology (CCSIT), Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
[4] Department of Computer Engineering (CE), College of Computer Science and Information Technology (CCSIT), Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
[5] Department of Management Information System (MIS), College of Business Administration (CBA), Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia

Corresponding Author Email: aaurrahman@iau.edu.sa

https://doi.org/10.18280/mmep.090315

Received: 31 January 2022
Accepted: 13 June 2022

Keywords:
information extraction (IE), digital libraries, ontologies, NLP, ML, HMM, CRF, WSD

ABSTRACT

The present article aims to review and evaluate the practiced and classical techniques, tools, models, and systems concerning automatic information extraction (IE) from published scientific documents like research articles, patents, theses, technical reports, and case studies etc. IE is performed for various reasons such as better indexing, archiving, searching, and retrieving. That is mainly used by the search engines and the indexing services as well the digital libraries and semantic web. In this regard, several studies have been conducted targeting various nature of documents. The study pays special consideration to the successful IE models, algorithms and approaches applied to structural IE from published documents. To grasp this, the paper is classified into several segments and each segment covers a significant aspect of IE. Furthermore, to validate their benefits and drawbacks, a comparative study of all the approaches have been conducted in terms of various performance factors like precision, accuracy, recall and F-score. Potential areas of improvement have been emphasized as research gap for the scholars in the closely related areas. Ultimately, a comprehensive summary of the evaluation is presented in tabular form and review is concluded. It was observed that the hybrid methods outperform the other methods due to their versatile nature to address various document formats.

1. INTRODUCTION

With the rapid growth in the technological aspects in the web, a huge volume of documents has been witnessed to be created on the daily basis. These documents are mainly unstructured or semi-structured, the search engines and the web crawlers are unable to index and consequently search them effectively. Hence, it is necessary to develop techniques for extracting useful information (like metadata etc.) from such documents that are generally unstructured and/or semi-structured. Since most of these documents are in portable document format (PDF), the techniques can be seen as converters which take unstructured texts to be more computer-friendly input and output information in a particular format. As documents on the web and from other sources is increasing, it becomes increasingly important to develop such techniques [1]. Due to increasing demand to obtain structured information from documents, information extraction (IE) has gained a significant importance in research. IE refers to the automated extractions of information from unstructured sources [2]. The role of IE is to locate in a particular domain a predefined collection of definitions, ignoring other unrelated information. The domain consists of a corpus of texts together with a specified need for information. In other words, from unstructured text, IE is about deriving organized factual knowledge. The volume and variety of data creates additional difficulties in identifying useful information [3]. Since the growth rates of unstructured information are very high and growing in recent times, the key challenges in IE, mining and analysis must be understood. The main challenges to collect useful information are the scalability, dimensionality, and heterogeneity of unstructured data [4]. The big questions are transforming unstructured data into a structured format to improve representation. Analytics need for efficient and accurate transformation by identifying new methods to extract semantics and contextual information through analysis, management, and query. Technological advances have encouraged rapid data volume growth in recent years [5]. The volume, diversity and the rapidity of big data also have altered the paradigm of system computing capacity. In addition,

unstructured data from different sources was predicted to grow to 90% in few years. It is estimated that 95% of global data for the year 2020 will have unstructured data with an annual rate of growth estimated at 65% [6]. Unstructured data exists in various formats like text, images, audio, video, blogs, and websites [7], non-standard and schema-less [8], further it comes from various sources like, social media, sensors [7]. Due to size and complexity of such data, the IE is a tedious task because it involves the format sensitive approaches whose effectiveness fluctuates severely with the slight change in the format of the documents. That is why, no single win-win scheme has been introduced that can handle all formats at the same time. The IE process is used to extract structured information from the data pipeline for analysis in the form of entities, relationships, facts, terms, and other information. Efficient and accurate data transformation leads to improved data analysis and IE performance. Various IE methods were proposed to extract structured and useful information from unstructured data to ultimately assist in the management, processing, and analysis of non-structured data [8]. Various documents (emails, web pages, newsgroups, news articles, business reports, research papers, blogs, abstracts, proposals) and a report on output from the source document is provided in accordance with certain specific criteria [8]. In the form of non-structured data, natural language texts carry textual information that does not have a predefined data model. The information and relationships represented by the data are therefore not explicitly indicated or formatted. The task of working with textual information is a difficult due to its variability [9]. The IE process is one of the important data analytical tasks in which structured information is extracted from unstructured data [10]. The IE is defined as "extracting instances from unstructured data from predefined categories, establishing a structured and unequivocal representation and relationship between entities." Documents are collected as inputs and different representations of relevant data that meet different criteria. IE techniques efficiently analyze the text in free form using structured format, extracting the valuable and relevant information [11]. The ultimate objective of IE techniques is to identify the key facts from text to enrich knowledgebases [12].

Rest of the paper is categorized as follows: section 2 contains ontology-based IE techniques, section 3 introduces the IE from articles. IE systems are elaborated in section 4 while section 5 contains approaches and methods to IE from research articles. Section 6 convers IE tools while section 7 concludes the paper.

2. LITERATURE REVIEW ON IE

2.1 Ontology based IE

Ontology based IE (OBIE) emerged as recent field in IE. The ontologies are used as basis of IE and the data is usually perceived through the ontology. This is to be noted that ontology is characterized as a mutual conceptualization, formal and explicit specification of a field [13]. Since ontologies are domain specific, IE also supports this idea and results in a powerful combination [14]. An OBIE system that, through a method driven by ontologies, processes unstructured or semi-structured NL text to extract certain information and presents the output using ontologies. It should be noted that this concept involves systems that construct ontology by processing natural language text [15] in addition to systems that recognize (and present) knowledge relevant to ontology. Although ontology construction is not commonly correlated with the IE, this can be an essential step in this process. Furthermore, ontology design itself extracts some knowledge as it defines the related domain concepts and relationships. Figure 1 shows the general architecture of OBIE system that comprehends various fields in IE.

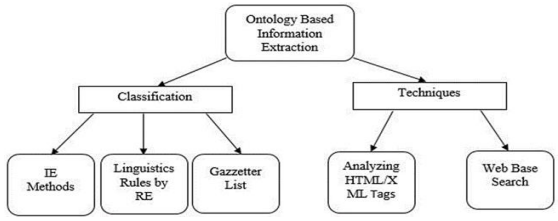

Figure 1. General architecture of OBIE system

Constantin et al. [16] introduces Document Component Ontology (DoCO), which was designed to describe various aspects related to scientific and other academic content as a general ontological unification framework. Its primary objective was to enhance the interoperability and sharing of academic documents and related services when multiple storage formats are used. The structural and rhetorical basis of DoCO together with hybrid structures describing components in terms of complementary structural and rhetorical performance are presented in the following sections. The usefulness of ontology in practice is then illustrated by presenting diverse applications that use DoCO to annotate and retrieve document components from academic articles as well as other activities of the semantic publishing community that directly use or promote DoCO as a complete ontology for document components formulation in RDF. In this article, the most common components of DoCO, such as paragraphs, figures, tables, chapters, references, body-to-body issues and the like were officially described. Moreover, author described tools and approaches that use DoCO for different purposes, for example, annotation of pdf documents or the semiology of the components of intended academic articles. The author plans to add additional mapping elements, such as JATS, metadata elements soon. The authors are also working on the extension of the current implementation of pdfX to identify other purely rhetorical document components (e.g., methods, materials, experiment, data, result, evaluation, discussion). In the XML conversion outputs, all these components are adequately doco-noted. The automatic converse of all structures retrieved and declared in the XML outputs to RDF according to DoCO and other relevant models will form another planned development for pdfX. Martínez-Romero et al. [17] developed tools that enable scientists to identify and dynamically develop new terms and sets of values in ontologies for recording their data. This work was incorporated into the CEDAR Workbench web-based platform. The resulting integrated environment offers a range of highly interactive interfaces for the production and release of ontologically rich metadata. In this article, the author outlines main characteristics developed by CEDAR that make it very simple to build web-based metadata acquisition forms and then enrich the forms with ontology concepts. Users may also set reusable field groups, known as elements. For instance, the fields describing a published piece may be grouped (e.g., author, title, year, publishing type, etc.) to form a publishing element that can then be reused in multiple templates. The Metadata Editor can be used to create an acquisition interface based on the form to type metadata for

that template after a template has already been created. Scientists who enter metadata using the Metadata Editor are prompted with drop-down lists, automatic suggestions and check tips in real-time which significantly decrease their error rate when entering and repairing metadata. The values specified in the templates are driven by these indicators.

Farhat et al. [18] proposed an OBIE system approach allowing the automatic semantically extraction of metadata from a particular sub-set of IEEE LOM standards metadata. A LOM metadata set and a domain ontology is the input for system. System's output is a set of semantic metadata in the form of the RDF. In the field of e-learning, seminal metadata research projects have little practical influence. This can partly be explained by the fact that it is hard and complicated for authors to add semi-metadata to a learning object. Moreover, there are already many learning objects, and it will be a tremendous task to enrich them with semantic metadata. Also, semantic metadata are not universal as they depend on domain ontologies that in many cases vary from community to community. Therefore, every time the learning object is used in a new context, the task must be repeated. The experiment results have shown that automatic generation of semantic metadata with existing technologies can be discussed.

The system of services for automatic processing of scientific collections, which are part of digital library collections, is presented in Ref. [19]. These services are based on ontologies for the representation of scientific documents and methods of semantic analysis of mathematical documents. The tools developed automatically check the validity of document compliance, convert them to required formats and generate metadata. This method is based on an analysis of the document structure and its stylistic characteristics. This is why, two technologies have been tried: the processing and conversion of unstructured data in a reading form by the machine. The author specifies the rules for selecting blocks for an article to extract metadata based on characteristics. In particular, such features include the style of the articles (font, font size, selection, etc.). Some additional features allow improving the quality of the metadata extraction, e.g., text pattern (for example, the location of an "Annotation" word in front of an annotation block or an email-address-template-type record) and the block location in the text (for example, the document starts with the title of the article). The author shows many features that are used for the structural analysis of the collection of scientific papers that were published in the 11th All-Russian Congress' material on basic problems in theoretical and mechanical engineering. Table 1 explains the features, metrics, datasets, approaches and performance of the research articles.

2.2 Classification of current OBIE systems

Methods employed by the OBIE are:

Linguistic Rules and Regular Expressions (RE)

The basic approach behind the strategy is to define regular expressions (RE) capturing such information types. For instance, the < NP > expression (watched), where < NP > denotes a noun phrase, might catch movie names (represented by the noun phrase) in a collection of documents. By defining a set of rules like this, a large amount of knowledge can be extracted [19]. Often the set of RE is implemented using a sequence of finite-state automatons (FSAs).

Classification Techniques

Various classification methods were used in IE, for example Support Vector Machines (SVM) [20], decision trees and maximum entropy models. [20] Offers a thorough overview of these methods and divides them into categories as techniques for "supervised classification". Sequence tagging methods such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF) fall in this group.

Analyzing HTML/XML Tags

Use the tags of documents, OBIE and IE systems that use html or xml pages as input will extract those information types. For example, a system that is aware of tables 'html tags that extract information from tables that are present in html pages [21]. The table's first row represents attributes, and the remaining rows represent the values of the attributes for individual records or instances. XML documents will provide more opportunities to extract information in this way as they allow users to define their tags.

Web-Based Search

To use queries for knowledge extraction on web-based search engines seems like a new technique [22]. The basic concept behind this strategy is to use the web as a broad corpus.

Only the ontology to be utilized for the extraction of information can be classified by OBIE systems. The ontology as an input into the system is one approach. With this approach, ontology can be manually built, or an ontology built off-shelf by others can be used. Approaches [16, 23] seem to have been adopted by most OBIE systems. Building an ontology from scratch or using an existing ontology as the basis can lead to ontology building. Some OBIE systems build an ontology and extract no instances. The ontology is updated through IE process by adding new classes and properties.

An ontology consists of various components, including classes, properties of data type, and properties of objects (including taxonomic relationships), instances (objects), instant property values and limitation. Based on the Ontology Components, OBIE systems may be classified [24]. Building systems in ontology generally extract class information only. Many elements of ontology are taken from systems which construct an ontology and find information concerning instances. During the ontology construction process, it extracts class names, taxonomies, and type of data.

Whilst all OBIE systems extract information from text in the natural language, they can have very different sources [25]. Some systems can handle any type of text in a natural language, whereas others have specific document structure requirements or target websites.

Table 1. Summary of ontology base IE approaches

Ref.	Features	Datasets	Metric	Performance
[16]	Extraction of metadata, body and references	117 Scientific papers from Balisage Conference	Precision Recall	88%
[17]	Extraction of metadata	Scientific Articles	NA	NA
[18]	Extraction of metadata	IEEE LOM files	Precision Recall	85%
[19]	Extraction of title, authors and block of literature	Scientific Publications	NA	NA

3. IE FROM SCHOLARLY ARTICLES

Initially, scientific literature was distributed in print form. But from the past 30 years, due to media transition, it is unknown how much literature is being published online. However, for some common online sources, statistics are known. For example, DBLP database, which offers bibliographic information, currently contains around 3 million records [26]. There are 57 million documents in Scopus database [27], which contains publications from a much broader variety of disciplines than DBLP or PubMed have. Approximately 2.2 million new scientific articles were published in 2016, according to the resources [28]. The increased number of publications, online digital libraries and accessibility in scientific literature were a major cause of the rapid growth of numerous scientific papers. International association of scientific, technical, and medical editors' (IASTME) report shows that publishers are increasing by 4-5% annually. Moreover, there are around 28,100 journals in English since 2014 [29]. This increase in scientific content presents crucial challenges for researchers interested in determining state of the art in their field. To conduct systemic literature reviews (SLR), several relevant research repositories require first literature. Subsequently, manual analysis filters the acquired results. The findings from these science articles are consolidated after acquiring the relevant literature to determine the state of the art of the field. This entire process of SLR is essential to the scientists, as it helps to analyze the research gaps and establish room for innovation, though it is time taking. Under one of the guidelines for the SLR, it may take up to 1 year to conduct a quality assessment [28]. In addition, an SLR with single/multiple human resource/s may take up to 186 weeks per [30]. Many research institutions and scientific publishers, such as ACM, IEEE, and Springer, have made digital repositories available to researchers. These libraries tend to offer user-friendly search filters that query millions of research documents using metadata from scientific articles. Thus, the extraction of metadata from scientific papers will eventually help save researchers time while SLR. The next is to read, assess and consolidate results of acquired literature. From the point of view of the researcher, this entire process is extremely important, but time consuming, laborious, and complicated. A variety of attempts have also been made to estimate the total variety of scientific posts [31], or a certain subset of them [32]. For example, Björk [33] reported that articles published by 2006 were approximately 1,350,000 using ISI database. Wu and Lee [34] used this finding and a variety of theories relating to a steady rise in the number of scholars, journals, and papers, and resulted at an estimate of more than 50 million articles ever written as of 2009. Finally, by examining the scope of two prominent academic search engines, Khabsa and Giles [35] analyzed the number of scholarly documents in English and accessible on the web: Google Scholar and Microsoft Academic Search. According to their figures, at least 114 million documents are available on the internet, with as low as 27 million without subscription or charge. According to the study [36], the statistical data derived from DBLP, and PubMed indicate similar patterns. Sadly, keeping the track of the latest articles is a big challenge for the because of the large and rising volume of scientific literature. The above discussion shows the importance of IE from research papers in terms of better searching, indexing and disposal of right information for the intended users.

4. SYSTEMS FOR STRUCTURED INFORMATION EXTRACTION FROM RESEARCH ARTICLES

This section describes the most well-known systems for extracting metadata from academic papers.

4.1 CiteSeerX

CiteSeerX is one of the very first scientific literature search engines. This integrates Autonomous Citation Indexing (ACI), a program that indexes electronic-format scholarly literature, such as Postscript and PDF files. CiteSeerX relies on data extraction approaches to construct quotation indexes [34]. More specifically, CiteSeer was designed to provide a high degree of versatility, enabling the use of different metadata extraction approaches. CiteSeer implements a blackboard architecture [37] consisting of three key components, namely (i) Information Sources, (ii) a Blackboard, and (iii) a Control component (CC), to provide this scalable solution. The component Information Sources is composed of experts, which corresponds to modules specializing in some part of the problem. Such specialists are the different approaches to solving the knowledge extraction problem. As for the Blackboard, it refers to a global database that contains the input data, partial solutions, and other information provided by the experts to facilitate problem resolution. Finally, the CC refers to the workflow manager that creates runtime decisions about the direction of problem consisting more specifically of the experts responsible for scheduling the information sources selected by the CC, depending on the problem of extraction. Using this method, CiteSeer can be integrated with other frameworks specializing in the metadata extraction.

4.2 ParsCit

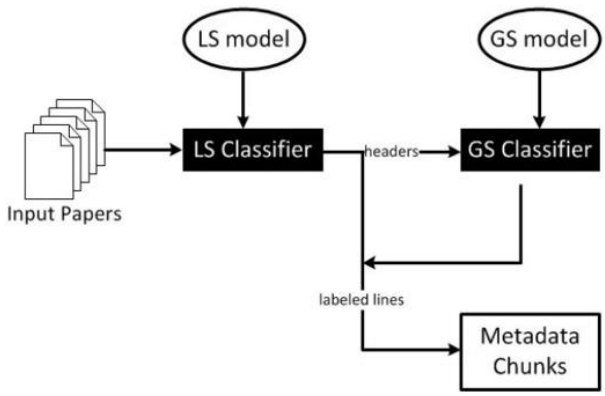

Figure 2. ParsCit architecture

The ParsCit focuses on extraction and parsing of reference strings, i.e., the text strings in the bibliography or reference portion of a publication related to a previous document [38]. All tasks are organized as supervised procedures for ML, using CRF as their learning mechanism. ParsCit uses Conditional Random Fields to mark the reference string token sequences, and it is combined with heuristic processing to classify reference strings from an unstructured text file, and to retrieve the citation contexts [39]. ParsCit starts by locating the references in the text using a collection of heuristics that consists of searching by a designated reference segment (e.g., "References," "Bibliography,") To do so the text is iteratively separated into strings that are likely to be labels in the

reference segment. After extracting the reference portion, ParsCit segments it into single strings. It uses RE for segmentation to define the different types of reference start markers. Three cases were considered for the start of a guide, namely (i) beginning with square brackets or parenthetical markers (e.g., [1], (1), (ii) beginning with naked numbers (e.g., 1), (iii) starting with naked numbers (e.g., APA style1). In the case that no reference markers are identified, a set heuristic is implemented to obtain the individual references, such as where individual strings start and end based on the length of previous lines, where strings tend to be lists of author names, and end punctuation as the final line of a quote usually ends with a time span. The CRF model is implemented after the obtainment of the list of individual references. It is depicted in Figure 2.

4.3 ArnetMiner

ArnetMiner is a program whose purpose is the extraction and mining of virtual academic networks [40]. The one more important to this work is the automated web-based extraction of researcher profiles, which resorts to Conditional Random Fields models to solve the labeling problem. This problem involves extracting the value of property (e.g., name, affiliation, homepage, telephone, research interest) from a personal profile (researcher), which is a non-trivial problem to solve as the researchers' format differs widely between various websites [41]. A three-step approach, namely, pertinent page recognition, pre-processing and extraction is proposed to solve this problem [42]. The program acquires a list of web pages using a search engine (e.g., Google API) for the appropriate page identification and assigned a researcher name, and then identifies the homepage using a binary classifier such as an SVM. When this list is collected in a pre-processing step, the text of the pages is split into tokens to which labels are allocated. Five types of tokens were identified, such as standard-word, special-word, phrase, for example, RE used to identify the special words. The token labeling process begins after the token recognition, using a CRF model. For example, the labeling function in the case of special word tokens consists of assigning each token to the classes like affiliation, name, address, phone.

4.4 Extracting bibliographic references from patents

Patrice Lopez developed a framework for the automated retrieval of bibliographic references in multilingual texts for patent documents [43]. The method of identification is achieved by using the Conditional Random Fields. There are several differences for the CRF used, and certain modifications, depending on whether the function is relevant to patent documents or academic papers, as references may appear in the text in patent documents, as opposed to what occurs in academic articles with references in a separate section. Those modifications refer to the set of characteristics used [44]. The text body of the target document (e.g., PDF) is extracted in a first step. The reference blocks for patents and non-patents are then listed in the text recurring to two separate CRF versions. A CRF then annotate non-patent references for the identification of a set of 12 classes (e.g., author, title, book title). Then the use of regular expression normalizes both patent and non-patent resulting from the annotation via CRF. Later, different online bibliographic databases are accessed in a last step to verify the resulting reference.

5. APPROACHES AND MODELS FOR STRUCTURED IE FROM RESEARCH ARTICLES

Several scholars have contributed to IE domain over the decades [2]. Scientific literature uses various machine learning (ML) and NLP approaches to retrieve metadata. ML and Rule-based provide some of the commonly used methods. Among the techniques used in ML are HMM, SVM and CRF. Following section outlines working method of each technique.

5.1 Rule-based approaches

It refers to a system which relies on predefined instructions to retrieve desired data. Several researchers have employed rule-based methods focused on text structure & formats in the sense of metadata extraction. Such as title, abstract, keywords, authors name and affiliation, acknowledgement, funding information and reference section. Research published in Ref. [45] uses rules which are based on textual and geometrical characteristics. It highlighted the extraction of the entities from an article's metadata: abstract-body, abstract-heading, affiliation, profile, title, highlight keyword-body, keyword-heading, page number, membership, pseudo-coding, publish-info, reader-service, synopsis, and text-body. They used a valid rule-base on multiple realms. Research suggests that they have reasonable effects when rules are used along with fuzzy matches. All pdf documents and metadata extracted were indexed and stored at the end of the process. In the framework, six different metadata (title, abstract, keywords, body text, conclusion, and references) from one pdf file are extracted in 3-5 seconds. These metadata are the most common in the IE systems from scientific articles and useful for the indexing, classifying, tagging, and searching the document. In total, 15-20 pdf files are extracted and stored per minute by the framework. The extraction speed is 9-10 times faster than the current packages for metadata extraction. The overall accuracy for scientific paper is 97.71% whereas the accuracy of the metadata extraction for the title is 91.21%. For abstract the accuracy of the results is 98.13%, for the keyword or index terms 92.53%, for body-text is 99.37%. Tuarob et al. [46] Use a set of RE to extract metadata, along with section headings (Title, Author Details, and Abstract). There are three subtasks to the problem: The 1) section headers, 2) standard section recognition (i.e., abstract, initiation) and 3) section hierarchy. Authors used the RE in the standard sections which record lexical patterns. It was demonstrated that a set of simple heuristics to capture sub-section relationship can be used effectively for your final task. This paper uses two datasets. The first collection of data contains 100 academic documents manually chosen for various types of publications from the CiteseerX repository. A second dataset contains 117 randomly selected PDF documents from the repository having diverse academic documents, specifically conference papers, newspapers, theses, and academic articles. In the *Section Boundary Detection*, algorithms achieve an F-score of 92.38%, standard Section Recognition accuracy of 96%, and accuracy of 95.51% for section positioning. Rizvi et al. [47] presents a new system for the IE from documents of user-related table information. The system presented uses rules which are interspersed with regular, generic expressions and can be applied to any documents irrespective of their nature. It is robust and can deal with different layouts of documents. The system has two main modules: table detection and ontology. All tables from a given document are extracted by the

detection module while only relevant tables from all detected tables are extracted from the second module. The generalized use of ontologies allows the system to adapt itself according to a new group of documents from any other field. The system was assessed on 80 actual technical documents containing 2033 tablets from 20 different industrial brands. Evaluation shows a precision of 0.88, 1 and 0.93 in accuracy, retrieval, and F score as given in Table 2.

Casali et al. [48] proposes a system architecture that assists the institutional repositories manager within a limited website in the collection of Spanish and English text documents. Acceptable documents can therefore be detected to be uploaded to the repository. The metadata for which the author publishes the document are also automatically extracted, in accordance with certain regulations, such as titles, classes, writers, languages, keywords and the relevant contact data. Email and authors' affiliation of the document to be uploaded will be contact data. The task of collecting information is becoming increasingly complex because of the large increase in the Web and the heterogeneity of its pages. The information collection system is responsible for the collection of data in well-defined collections. The collection should be restricted to specific domains to extract relevant information. Clark and Divvala [49] suggested a rule-based approach to analyze a page structure by detecting body text chunks and identifying the areas that figures or tables could lie within the text by reasoning about the empty regions. This method can extract a wide range of figures if they are different from the main article text, as they do not make strong assumptions regarding the format of the figures in the document. The algorithm also shows a corresponding sub-part that works even if individual subtitles are adjacent to multiple figures. It also contains procedures to leverage specific assumptions of consistency and format for identifying the titles, text, and subtitles of each article. It presents a new dataset of 150 informatics papers with basic labels for the location of the figures, tables, and subtitles. When tested against this data set, algorithm achieves 96% accuracy, which exceeds the prior art. To allow future research, we publish our dataset, code, and scripts for evaluation on our project site.

According to Ahmad et al. [50], paper introduces a comprehensive approach for obtaining components of information in PDF format from CEUR Workshop research papers. The proposed framework uses strong technologies translated to XML and designed rules from plain text formats. The in-depth study has described different circumstances in which the XML document or plain text database can help to retrieve the structured information required more accurately. The extracted information includes authors, affiliation & address, first level text headings, table and figure captions, funding statement and project information. The method cannot work in certain traditional cases because of limitations of the designed rules developed for the CEUR dataset. Two datasets from the ESWC challenge are chosen for experimental analysis. The first preparation dataset consists of 45 research articles, while the second evaluation dataset contains 40 research papers. The RDF contains 2,488 triples and 1,815, respectively, for each preparation and evaluation dataset. In both datasets, the execution of eight intended queries against each study paper produces 360 and 320 (CSV extension) files, respectively. The generated CSV files are analyzed with the gold standard dataset using evaluation tool. In the ESWC challenge website, the gold standard dataset and validation tool are available. The overall F-score for both datasets is 78.3%.

The study [51] focuses on developing ontology describing paper metadata, resources cited, metadata extraction procedures by named entities. Developing tool crawling PDF papers using methods of metadata extraction, publishing results as Linked Open Data. Finally, developing a metadata library and PDF full text of papers for the extraction of context information. That uses RE based on the style attributes of the HTML page, NLP, acronym resolution heuristics and the extraction of named entities. Further work involves improving the extraction procedures for named entities.

The study in Ref. [52] presents a preliminary basic model work to extract metadata from the cover pages of the scanned theses and dissertations (ETDs). The process started by converting scanned pages into images and then using OCR tools to create text files. A series of carefully crafted RE is used for each field, capturing patterns in seven fields of metadata: titles, authors, years, graduates, academic programs, institutions, and consultants. The method is evaluated with a dataset of basic truth consisting of corrected metadata from Virginia Tech and MIT libraries. In the area of ETD text files our heuristic procedure achieves up to 97% accuracy. The approach offers a strong foundation of ML. This is the first work to extract metadata from non-born digital ETDs to our best knowledge. Summary of rule-based approaches is provided in Table 3.

Table 2. Comparison of various approaches

Datasets	Precision	Recall	F-score
[45]	0.97	0.85	0.90
[46]	0.95	0.96	0.92
[47]	0.88	0.95	0.93
[49]	0.96	0.89	0.90
[50]	0.77	0.76	0.77
[52]	0.97	0.91	0.92

Table 3. Summary of rule-based (RE) approaches

Ref	Nature of text	Datasets	Metric
[45]	Title, abstract, keywords, page# and body text	Different open access computer science journals	P, R, F
[46]	Title, author details, abstract, headers	CiteSeer X repository	P, R, F
[47]	Tables with caption	Industrial reports	P, R, F
[48]	Title, author details and affiliations	Educational digital libraries	P, R, F
[49]	Tables and figures with caption	150 computer science papers	P, R, F
[50]	Title, abstract, caption of figures and tables, headings, references	ESWC datasets	P, R, F
[51]	Title, author details and full text body	ESWC datasets	P, R, F
[52]	Title, authors, years, graduates, institutions.	Virginia Tech Thesis and MIT libraries	P, R, F

5.2 Machine-learning based approaches

This section contains the review and evaluation of ML based approaches in the IE from published articles.

5.2.1 Hidden Markov model (HMM)

HMM has good analytical grounds that are basically robust and successful to build. The main downside is their dependency on the training details. It is widely used in many areas like NLP, pattern matching etc. HMM can be described in such a way.

λ = (A, B, π), is simplified notation for an HMM. Other notation is used in Hidden Markov Models:

A = state transition probabilities (a_{ij})

B = observation probability matrix ($b_j(k)$)

N = number of states in the model {1, 2, …, N} or the state at time $t \rightarrow s_t$

M = number of distinct observation symbols per state

Q = $\{q_0, q_1, \ldots, q_{N-1}\}$ = distinct states of the Markov process

T = length of the observation sequence

V = {0, 1, ..., M-1} = set of possible observations

O = (O_0, O_1, ..., O_{T-1}) = observation sequence

π = initial state distribution (π_i)

s = state or state sequence (s_1, s_2, ..., s_n)

x_k = hidden state

z_k = observation

Three types of problems can be solved by the HMM, namely, evaluation problems, decoding problems and learning/ optimization problems.

Prasad et al. [53] used the HMM for parsing of reference strings together with profound learning. The architecture makes it possible to induct features that are tuned to the parsing of the reference strings. In comparison with constructed and dictionary features, the author demonstrates the superiority of abstract numerical representations of the word learned from unlabelled data. They have been exploring a unified way in multilingual datasets to translate the reference string into English, to scan it using the technique proposed and to propagate it into the original string. The report emphasizes the importance of using HMM layer to increase model robustness and achieve the same performance with state-of-the-art hand-crafted systems. In all, the model proposed, which has already been accepted as a strong foundation of this task, has significant F-Score gains over the existing systems.

In Ref. [54], Trigram HMMs are used to derive metadata from citations. To develop the model, a minimum of twenty features are used as vocabulary. These features include period, comma, letter of capital, all numbers. The research describing bigrams for network training used a self-created test dataset composed of 713 citation strings from 250 scientific papers.

5.2.2 Conditional random fields (CRF)

CRF is a mathematical model having the capacity to incorporate neighborhood effect. A study [55] presents a metadata extraction framework that includes the identification of the header along with the references in English and Persian. Model CRF was used in the extraction of header and reference metadata. By defining various features, this model may be modified. In a set of 100 science documents from various Iranian journals, the method proposed has been tested. This model has greater precision than other models than Markov in text tagging. The model is based on statistics, on the other hand. Extracting metadata from papers of different layouts and styles, while using statistics, produces better results than rules. The use of this model is therefore a good solution. The proposed method was assessed by the measure F. For each token, the measure F is calculated. For metadata, Persian references, and English references respectively, an Average F-score is 96.89%, 93.87% and 94.75% was observed.

Ramesh et al. [56] proposed an automated approach for the identification, using the sequence of multiple CRF models that build on the prediction of each model, of sections and their labels including the metadata (title, name, and association of authors), referred metadata and citation. An XML was then generated and transformed into an RDF document to build a knowledge base. They have shown that assessment values after the challenge comparable to those of the SemPub-2016 model matching approaches, thus highlighting the value of integrating ML and NLP techniques. Author trained on 146 CEUR-WS workshop documents. System yielded an F-measure of 0.612 on average with precision and recall as 0.629 and 0.62, respectively).

Study in Ref. [57] proposed a new method to extract bibliographic information from heterogeneous references using the CRF, such as author name, title, year of publication, volume, edition, and journal name. The model of the CRF was selected because the fields in the reference list are often sequentially listed and have patterns. CRF is a statistical model for the prediction of sequential and structured labels by considering the next-to-neighbour samples. Their structure was initially normalized with its different patterned field relations to extract bibliographical data from a reference list. In addition, several features, including punctuation, number, capital letter, word length, person/procedure dictionaries, and indication words were used to boost the CRF classifier. Finally, the accuracy of the proposed model was measured as 97.10% by analysis of 1415 heterogeneous references cited in academic papers published in Korea. In the citation extraction pipeline the extraction of single reference strings from the scientific publications reference section is an important step.

Körner [58] divided this task into two steps using the CRF technique by first detecting the reference areas and then grouping the lines into reference strings within such a field. They propose an EXCITE classification model which considers every line as a potentially part of a reference string in the article. When using random lines instead of constructing the graphic model based on the individual words, dependences, and patterns typical for referrals, the overall complexity of the model is reduced. Anzaroot and McCallum [59] and Vilnis et al. [60] demonstrate many articles that focus on developing the underlying CRF models to expand the scope of global contexts. These studies discuss citation extraction as a UMASS dataset method for enhanced CRF models.

Rahnama et al. [61] introduced two layers of the CRF model. The first page of the review document is considered in the study as it includes potential details about metadata headers. The first layer describes wider components which can include information regarding metadata from the document content. The header, author, description, body, and footnote details are the components. Since the body class does not contain useful data for extracting metadata features, it is not processed any further, but it processes header and author. On the other side, as footnotes usually contain information about publishers, conference information, and additional information about authors which may contain email and writers' relationship. For header, author, and footnote content, therefore, a second layer of CRF was created. This extra layer permits extracting the real metadata and defining roles in the segment. Results are tested on 100 articles, while dataset and respective corpus are freely available on GitHub.

Another study [62] focused on the enhancement of standard CRF efficiency through the implementation of semi-CRF higher-order principles. The transformation between chains of

variable length sequences can be modelled by these models, thereby giving them greater control than standard linear chain CRFs. The measurements are conducted using ParsCit dataset semi-Markov CRFs with linear-chain CRFs as baseline and first order, second order, and third order. CRF already offers state-of-the-art outcomes in metadata processing activities. These models often struggle with HMM constraints.

If we contrast the HMM and CRF models, HMM is based on Bayes rule, while CRF is based on maximum entropy rule. Further here are advantages and disadvantages.

Advantage compared to HMM: Since CRF does not have as strict independence assumptions as HMM does, it can accommodate any context information. Its feature design is flexible.

Disadvantage: CRF is highly computationally complex at the training stage of the algorithm. It makes it very difficult to re-train the model when newer data becomes available.

5.2.3 Support vector machines (SVM)

SVM is another method to extract metadata. This is widely used for regression and classification. Cost function can be expressed as Equation:

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

The cost is 0 if the predicted value and the actual value are of the same sign. If they are not, we then calculate the loss value.

In Ref. [63], SVM technique was used to develop and implement an accurate automatic algorithm to extract rich metadata directly from a PDF article. The algorithm takes a single publication, analyses, and releases a structured machine-readable metadata file including title, abstract, keywords, author's full names and e-mail addresses, journal name, volume, issue, pages of the publication and year. The algorithm designed as a universal solution to handle diverse nature of documents. This was achieved by using supervised and unsupervised ML algorithms trained on large numbers of data sets and resulting in an increased system performance and adaptability to new documents. Heuristics were also accompanied by ML solutions. This approach has proven to be good in practice even in contrast to simple ML based approaches (Table 4).

Table 4. Accuracies against datasets of different papers of IE

Datasets	Precision	Recall	F-score
[53]	0.96	0.89	0.90
[54]	0.77	0.76	0.77
[55]	0.97	0.91	0.92

It requires checking on a variety of classifiers including Decision Tree (DT), K-Nearest Neighbor (kNN), Naïve Bayes (NB) and SVM. This means that the best results are obtained where five different SVM classifiers are used, each classifier is used to classify one category. This resulted in an F-score calculation of over 85% for all categories except keywords. An overview of the literature reviewed on the extraction of metadata from scientific publications is given in Table 5. Moreover, it enlists summary of ML approaches with the precision, recall and F-measure (P, R, F) accordingly.

Table 5. Summary of ML based approaches

Ref.	Approach	Features	Datasets	Metric
[53]	HMMs	References and its metadata	Multilingual Scientific Papers	P, R, F
[54]	Trigram HMM	Metadata from citation	713 citation strings from 250 papers	P, R, F
[55]	CRF	Header and references metadata	Scientific documents from Iranian journals	P, R, F
[56]	CRF	Title, author details & its affiliations and citation	CEUR-WS scientific Documents	P, R, F
[57]	CRF	Metadata from references	Academic papers published in Korea	P, R, F
[58]	CRF	Citation	EXCITE dataset	P, R, F
[60]	CRF	Citation	UMASS Dataset	P, R, F
[61]	CRF	Header and footer metadata	Github Corpus	P, R, F
[62]	CRF, HMM	Metadata from references	ParsCit Dataset	P, R, F
[63]	SVM, CRF	Title, author details, Abstract, Keywords and metadata of references	Citeseer-X and Mendely dataset	P, R, F
[20]	SVM	Figures with captions	Arabic book page	P, R, F

Study in Ref. [20] described the LABA system for logically analyzing scanned Arabic book pages, based on multiple SVM. The function of a text area can be detected by analyzing different image characteristics and by extracting the figures, title, and quotation from the documents. SVM has been used this research in many ways. If all fields are defined by a single classifier, and several classifiers are used to classify each type, then the results are compared.

5.3 Other hybrid approaches

There are several studies that either use hybrid approach to retrieve metadata, or use other approaches not mentioned in the sections above. In Ref. [64] for automatic metadata extraction, authors suggest the use of formatting templates. The pure texts and their related formatting information including line height, font type and size are recognized in parallel for direct metadata recognition. In contrast to the optical character recognition (OCR)-based methods, authors use the open source PDFBox and compare the proposed PAXAT system for well-known metadata extraction approaches, namely, arXiv, ACM, ACL, and other publicly accessible and institutionally subscribed outlets. Title, author names, affiliations and matching accuracies were 0.9798, 0.9425, 0.9298, and 0.9109, respectively. The extraction results for the 9,992 articles (excluding the 185 journal papers with partial experiment results) are shown in Table 6. From title to author-affiliation matching, the precision decreases

because this method provides optimum specificity for title, which has less formatting variety in most academic documents.

Table 6. Accuracies against datasets of different papers

Datasets	Precision	Recall	F-score
[64]	0.96	0.89	0.90
[65]	0.77	0.76	0.77
[66]	0.97	0.91	0.92

Tkaczyk et al. [65] introduces CERMINE, a born-digital system for extracting both metadata and bibliography from scientific papers. Automated extraction tools from CERMINE support several operations, such as smart scanning, recognition of related papers, development of citation and authoring networks. The modular architecture and the application of supervised and unsupervised techniques of ML make it flexible and easy to adapt to modern document formats. The evaluation against a large and complicated dataset shows good results for the key measures and the whole extraction workflow and outperforms similar IE schemes. In Ref. [66], the approach employs an AI method, case-based reasoning (CBR), based on the notion that similar problems have similar solutions. CBR is the paradigm for problem solving that resolves new problems based on solutions to similar past issues. Each case is contextualized and has a problem description part which can be represented in either vector, structured or text representations and the solution part. In this case, the previous cases are saved in a case base. The key steps consist of CBR: checking, reuse and retaining. Method learns and stores tag sequence in a case base from each test. If a new tag sequence needs to be classified, the system checks its case base to see if a similar tag was experienced before, otherwise, a new case was added using the proposed algorithm. Study in Ref. [63] illustrate the process of extracting entities from Elsevier database quotations from the texts including unstructured text files, sentence classification and extraction. The process extracts authors name(s). Chenet [67] Illustrates excerpts from non-structured texts. To obtain the correct documents mentioned in the unstructured text an end-to-end process carried out, starting from raw data, and finishing off with correct extraction of entities. Metadata extraction from academic papers is useful for several applications. Liu et al. [68] introduced the deep learning networks in many fields such as computer vision (CV), and NLP using deep learning. First, a deep learning network was used to shape the image information and the text information of headings, which allow to extract metadata with little loss of information. Two typical tasks are done: object detection in the CV field, and NLP sequence marking. Finally, the two networks created by the two tasks are combined to give extraction results. Ahmed and Afzal [69] Proposed an IE scheme for publishers with diverse styles of logical structures for articles aka FLAG-PDFe to extract unique metadata from a research paper placed. The approach builds on separate and generic features based on textual and geometrical information from the raw materials of research papers. Different physical layout components of an article are identified by the separate features in the first step. Since journals follow unique formats of publishing and layout, they develop generic features to deal with these diversified patterns. SVM was used in the third phase, to learn and extract Logical Layout Structure (LLS) from article. The study results are achieved through the gold standard data set. The results show 0.877 recall, 0.928 precision and 0.897 F-score. Compared to the best approach in ESWC challenge, the approach achieved 16% f-score gain. In Ref. [70], it is stated that websites, articles, and other documents cannot neglect the tables while searching. There is a vast and rich literature that outlines different aspects of the identification, extraction, discovery, classification, and annotation of the tables. Author outlined the shortcomings of existing techniques and address them appropriately. Zhou et al. [71] used data sets named as SEYMORE which contain 935 headers and OURS with 75000 headers for sentence similarity. Results show metadata removal system performs far better than other systems. Metadata and content must be annotated to extract the content of a document automatically in a structured manner. Per [72] evaluation, the best performing out-of-the-box tool is GROBID, followed by CERMINE and ParsCit, respectively.

Several book-search engines are available [73-75] that mainly rely on IE approaches for better searching and indexing of books in digital libraries [76-80]. A summary of related literature review of hybrid approaches and related work, a taxonomy of similar approaches is provided in Table 7.

Table 7. Summary of hybrid approaches

Ref.	Approach	Features	Datasets	Metric
[64]	OCR, PAXAT	Authors details and affiliations	Arvix, ACM, ACL journals	P,R,F
[65]	CERMINE, ML	Header metadata and reference metadata		P,R,F
[66]	CBR	Paper metadata		P,R, F
[67]	Text features	Author details and its affiliations	Elsevier Database	P, R, F
[71]	DL, NLP, CV	Extraction of header metadata of papers	SEYMORE and OURS	P, R, F
[69]	FLAG-PDFe, LLS	Title, abstract, headings, references and caption of figures and tables	ESWC datasets	P, R, F
[70]	Text features	Tables with captions		P, R, F

It is observed that hybrid techniques are way better than their simple counterparts in terms of not only the metrics like F-score, precision and recall but in terms of the coverage and managing many unseen document formats. This is mainly because in the hybrid techniques people employ the best combinations of the schemes where each is playing a role in the IE in its own way. For instance, one scheme is good at extracting title but no other metadata, other is good at extracting authors and details but no other metadata, third is good at references but no other metadata. Combining the three can be good at extracting title, authors details as well as the references which was not possible individually.

6. INFORMATION EXTRACTION TOOLS

IE refers to extracting metadata that can be read by a machine, such as author names, title, abstract, keywords and

references. Several approaches, including RE, RBS, ML, SVM, HMM, and CRF etc. have been proposed to this problem develop many open-source metadata parsers and tools. That include Anystyle-Parser, Biblio, Cérmin, Citation, GROBID, ParsCit, PDFSSA4MET, Science Parse and Citation Tagger. Several approaches have also been proposed [81-90]. Table 8 enlists resources along with their primary algorithm and links.

Table 8. Summary of metadata tools for IE

Name	Approach	Extracted Fields
Anystyle-Parser	CRF	authors, booktitle, date, DOI, edition, editor, genre, ISBN, journal, location, pages, publisher, title, URL, volume
Biblio	RE	authors, date, editor, genre, issue, pages, publisher, title, volume, year
BibPro	Template Matching	authors, editor, institution, issue, journal, pages, volume, year
CERMINE	CRF	authors, DOI, issue, pages, title, volume, year
Citation	RE	authors, title, URL, year
Citation-Parser	RE	authors, booktitle, issue, journal, pages, publisher, title, volume, year
Free_Cite	CRF	authors, booktitle, date, editor, institution, journal, location, pages, publisher, title, volume
GROBID	CRF	authors, date, editor, issue, journal, organization, pages, title, volume
Neural-ParsCit	LSTM	authors, booktitle, date, editor, institution, journal, issue, location, pages, publisher, volume
Pars-Cit	CRF	authors, booktitle, date, editor, institution, journal, issue, location, pages, publisher, volume
PDFSSA 4 MET	RE	pages, title, volume, year
Reference Tagger	CRF	authors, issue, journal, pages, title, volume, year
Science Parse	CRF	author title, volume, year, journal

7. CONCLUSIONS

This study is dedicated to review and evaluate various IE techniques, approaches, and tools in the literature. Detailed overview is summarized in the form of tables about the IE from scientific publications. The area of reference in table header reflects the respective analysis of research. Type field reflects what sort of information is being extracted i.e., either research performs extraction of metadata from headers or extraction of sections from the body or the extraction of acknowledgement and references. Format refers to the input format needed for further processing e.g., by the proposed methodology like doc, plain text, XML. Improvement refers to major identifying contribution or features that improve efficiency introduced into the report. The dataset refers to the name of the dataset used for measurement purposes. Lastly, Metric reflects measure(s) of assessment added respectively to performance outcomes. Below, in the Metrics column: A, P, R, and F, precision, accuracy, recall and F1-score are expressed, respectively. It is evident in the light of Table 5 that most studies use CRF for the extraction of metadata. In CRF, HMM of the highest order are generated to catch the possibility of different segments having variable lengths. Other enhancements include smoothing methods, better error functions and optimization algorithms. In IE, among the key problem in logical extraction of structural information especially in the presence of errors that occur during the conversion process since various libraries result in errors and affect IE performance. Nevertheless, for all studies dealing with PDF format it is a critical component. On the other hand, studies using OCR to define visual format blocks appear to work very well, and typically exploit knowledge about layout and font style to boost performance. There are now a variety of open-source platforms which help extract this knowledge from scientific articles automatically. Such systems are currently suffering mainly due to format conversion with the layout and formatting problems. Recent comparative study [72] shows that GROBID, CERMINE and ParsCit pose best results among the various open-source extractors. In future, the schemes involving deep learning models and other hybrid intelligent approaches such as federated learning and transfer learning [91-100] can be anticipated more successful and worthy to explore for the IE from published scholarly articles of diverse nature.

REFERENCES

[1] Lim, C.G., Jeong, Y.S., Choi, H.J. (2019). Survey of temporal information extraction. Journal of Information Processing Systems, 15(4): 931-956. https://doi.org/10.3745/JIPS.04.0129

[2] 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. https://doi.org/10.1007/978-3-319-67056-0_18

[3] Zaman, G., Mahdin, H., Hussain, K., Rahman, A. (2020). Information extraction from semi and unstructured data sources: A systematic literature review. ICIC Exp. Lett., 14(6): 593-603. https://doi.org/10.24507/icicel.14.06.593

[4] Agrawal, K., Mittal, A., Pudi, V. (2019). Scalable, semi-supervised extraction of structured information from scientific literature. In Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications, Minneapolis, Minnesota, pp. 11-20. http://dx.doi.org/10.18653/v1/W19-2602

[5] Mannai, M., Karâa, W.B.A., Ghezala, H.H.B. (2018). Information extraction approaches: A survey. In Information and Communication Technology, pp. 289-297. https://doi.org/10.1007/978-981-10-5508-9_28

[6] Adnan, K., Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data, 6(1): 1-38. https://doi.org/10.1186/s40537-019-0254-8

[7] 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. https://doi.org/10.1016/j.techfore.2015.12.019

[8] Adnan, K., Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. International Journal of

Engineering Business Management, 11: 1847979019890771. https://doi.org/10.1177%2F1847979019890771
[9] Popovski, G., Seljak, B.K., Eftimov, T. (2020). A survey of named-entity recognition methods for food information extraction. IEEE Access, 8: 31586-31594. https://doi.org/10.1109/ACCESS.2020.2973502
[10] Niklaus, C., Cetto, M., Freitas, A., Handschuh, S. (2018). A survey on open information extraction. arXiv preprint arXiv:1806.05599.
[11] 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. https://doi.org/10.1109/ACCESS.2021.3063181
[12] Safar, M. (2020). Digital library of online PDF sources: An ETL approach. International Journal of Computer Science and Network Security, 20(11): 173-181. https://doi.org/10.22937/IJCSNS.2020.20.11.21
[13] Biniam, P. (2020). Ontology-based information extraction from legacy surveillance reports of infectious diseases in animals and humans. Digitala Vetenskapliga Arkivet.
[14] 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): 18. https://doi.org/10.1186/s13673-016-0074-1
[15] Konys, A. (2018). Towards knowledge handling in ontology-based information extraction systems. Procedia Computer Science, 126: 2208-2218. https://doi.org/10.1016/j.procs.2018.07.228
[16] Constantin, A., Peroni, S., Pettifer, S., Shotton, D., Vitali, F. (2016). The document components ontology (DoCO). Semantic Web, 7(2): 167-181. https://doi.org/10.3233/SW-150177
[17] Martínez-Romero, M., O'Connor, M.J., et al. (2017). Supporting ontology-based standardization of biomedical metadata in the CEDAR Workbench. In Proceedings of the Int Conf Biom Ont (ICBO), pp. 1-6.
[18] 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. https://doi.org/10.1007/978-3-662-44188-6_34
[19] Elizarov, A., Khaydarov, S., Lipachev, E. (2017). Scientific documents ontologies for semantic representation of digital libraries. In 2017 Second Russia and Pacific Conference on Computer Technology and Applications (RPC), Vladivostok, Russia, pp. 1-5. https://doi.org/10.1109/RPC.2017.8168064
[20] Qin, W., Elanwar, R., Betke, M. (2022). Text and metadata extraction from scanned Arabic documents using support vector machines. Journal of Information Science, 48(2): 268-279. https://doi.org/10.1177%2F0165551520961256
[21] Milosevic, N., Gregson, C., Hernandez, R., Nenadic, G. (2019). A framework for information extraction from tables in biomedical literature. International Journal on Document Analysis and Recognition (IJDAR), 22(1): 55-78. https://doi.org/10.1007/s10032-019-00317-0
[22] Suganya, G., Porkodi, R. (2018). Ontology based information extraction-a review. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, pp. 1-7. https://doi.org/10.1109/ICCTCT.2018.8551071
[23] Konys, A. (2016). A framework for analysis of ontology-based data access. In International Conference on Computational Collective Intelligence, Halkidiki, Greece, pp. 397-408. https://doi.org/10.1007/978-3-319-45246-3_38
[24] Sibarani, E.M., Scerri, S., Morales, C., Auer, S., Collarana, D. (2017). Ontology-guided job market demand analysis: A cross-sectional study for the data science field. In Proceedings of the 13th International Conference on Semantic Systems, Amsterdam, Netherlands, pp. 25-32. https://doi.org/10.1145/3132218.3132228
[25] Florrence, M. (2020). Building a multilingual ontology for education domain using Monto method. Computer Science and Information Technologies, 1(2): 47-53. https://doi.org/10.11591/csit.v1i2.p47-53
[26] Burch, M., Pompe, D., Weiskopf, D. (2015). An analysis and visualization tool for DBLP data. In 2015 19th International Conference on Information Visualisation, Barcelona, Spain, pp. 163-170. https://doi.org/10.1109/iV.2015.38
[27] Visser, M., van Eck, N.J., Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies, 2(1): 20-41. https://doi.org/10.1162/qss_a_00112
[28] Nasar, Z., Jaffry, S.W. (2018). Trust-based situation awareness: Agent-based versus population-based modeling a comparative study. In 2018 International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, pp. 1-7. https://doi.org/10.1109/ICACS.2018.8333494
[29] Ware, M., Mabe, M. (2015). The STM report: An overview of scientific and scholarly journal publishing. University of Nebraska, Lincoln
[30] 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. https://doi.org/10.1136/bmjopen-2016-012545
[31] Sajid, N.A., Ahmad, M., Afzal, M.T., Rahman, A. (2021). Exploiting papers' reference's section for multi-label computer science research papers' classification. Journal of Information & Knowledge Management, 20(01): 2150004. https://doi.org/10.1142/S0219649221500040
[32] Martín-Martín, A., Thelwall, M., Orduna-Malea, E., Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. Scientometrics, 126(1): 871-906. https://doi.org/10.1007/s11192-020-03690-4
[33] Björk, B.C. (2017). Growth of hybrid open access, 2009–2016. PeerJ, 5: e3878. https://doi.org/10.7717/peerj.3878
[34] Wu, J., Lee, C. (2020). Scholarly very large data: challenges for digital libraries. In Challenges for Large Scale Networking (LSN) Workshop on Huge Data: A Computing, Networking and Distributed Systems Perspective.
[35] Khabsa, M., Giles, C.L. (2014). The number of scholarly documents on the public web. PloS one, 9(5): e93949. https://doi.org/10.1371/journal.pone.0093949

[36] 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. https://doi.org/10.1080/03075079.2015.1060711
[37] Hartmann, W.M. (2005). Modern Acoustics and Signal Processing. Computational Ocean Acoustics. https://doi.org/10.1007/978-1-4419-8678-8
[38] Indrawati, A., Yoganingrum, A., Yuwono, P. (2019). Evaluating the Quality of the Indonesian Scientific Journal References using ParsCit, CERMINE and GROBID. University of Nebraska - Lincoln.
[39] Lauscher, A., Eckert, K., Galke, L., et al. (2018). Linked open citation database: Enabling libraries to contribute to an open and interconnected citation graph. In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, Texas, USA, pp. 109-118. https://doi.org/10.1145/3197026.3197050
[40] Tang, J. (2016). AMiner: Toward understanding big scholar data. In WSDM 2016: Ninth ACM International Conference on Web Search and Data Mining, San Francisco, California, USA, pp. 467-467. https://doi.org/10.1145/2835776.2835849
[41] Sleeman, J., Finin, T., Joshi, A. (2015). Entity type recognition for heterogeneous semantic graphs. AI Magazine, 36(1): 75-86. https://doi.org/10.1609/aimag.v36i1.2569
[42] Yadav, P., Remala, N., Pervin, N. (2019). Reccite: A hybrid approach to recommend potential papers. In 2019 IEEE International Conference on Big Data (Big Data), pp. 2956-2964. https://doi.org/10.1109/BigData47090.2019.9006220
[43] Romary, L., Lopez, P. (2015). Grobid-information extraction from scientific publications. ERCIM News, 100. https://hal.inria.fr/hal-01673305.
[44] 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. https://doi.org/10.1016/j.wpi.2017.11.002
[45] Azimjonov, J., Alikhanov, J. (2018). Rule based metadata extraction framework from academic articles. arXiv preprint arXiv:1807.09009.
[46] Tuarob, S., Mitra, P., Giles, C.L. (2015). A hybrid approach to discover semantic hierarchical sections in scholarly documents. In 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1081-1085. https://doi.org/10.1109/ICDAR.2015.7333927
[47] Rizvi, S.T.R., Mercier, D., Agne, S., Erkel, S., Dengel, A., Ahmed, S. (2018). Ontology-based information extraction from technical documents. In ICAART, (2): 493-500. https://doi.org/10.5220/0006596604930500
[48] Casali, A., Deco, C., Beltramone, S. (2016). An assistant to populate repositories: gathering educational digital objects and metadata extraction. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 11(2): 87-94. https://doi.org/10.1109/RITA.2016.2554018
[49] Clark, C.A., Divvala, S. (2015). Looking beyond text: Extracting figures, tables and captions from computer science papers. In Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.
[50] Ahmad, R., Afzal, M.T., Qadir, M.A. (2016). Information extraction from PDF sources based on rule-based system using integrated formats. In Semantic Web Evaluation Challenge, pp. 293-308. https://doi.org/10.1007/978-3-319-46565-4_23
[51] Kovriguina, L., Shipilo, A., Kozlov, F., Kolchin, M., Cherny, E. (2015). Metadata extraction from conference proceedings using template-based approach. In Semantic Web Evaluation Challenges, pp. 153-164. Springer, Cham. https://doi.org/10.1007/978-3-319-25518-7_13
[52] 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. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, pp. 515-516. https://doi.org/10.1145/3383583.3398590
[53] 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. https://doi.org/10.1007/s00799-018-0242-1
[54] Ojokoh, B., Zhang, M., Tang, J. (2011). A trigram hidden Markov model for metadata extraction from heterogeneous references. Information Sciences, 181(9): 1538-1551. https://doi.org/10.1016/j.ins.2011.01.014
[55] Tansazan, A., Mahdavi, M.A. (2017). Metadata extraction from Persian scientific papers using CRF model. Library and Information Science Research, 7(1): 304-321.
[56] Ramesh, S.H., Dhar, A., Kumar, R.R., KS, S., Pearce, J., Sundaresan, K.R. (2016). Automatically identify and label sections in scientific journals using conditional random fields. In Semantic Web Evaluation Challenge, pp. 269-280. https://doi.org/10.1007/978-3-319-46565-4_21
[57] Seol, J.W., Choi, W.J., Jeong, H.S., Hwang, H.K., Yoon, H.M. (2018). Reference Metadata Extraction from Korean Research Papers. In International Conference on Mining Intelligence and Knowledge Exploration, pp. 42-52. https://doi.org/10.1007/978-3-030-05918-7_5
[58] Körner, M. (2017). Reference String Extraction Using Line-Based Conditional Random Fields. arXiv preprint arXiv:1705.08154. http://arxiv.org/abs/1705.08154.
[59] Anzaroot, S., McCallum, A. (2014). A new dataset for fine grained citation field extraction (Author's Manuscript). University of Massachusetts, Amherst Amherst United States.
[60] Vilnis, L., Belanger, D., Sheldon, D., McCallum, A. (2015). Bethe projections for non-local inference. arXiv preprint arXiv:1503.01397.
[61] Rahnama, M., Hasheminejad, S.M.H., Nasiri, J.A. (2020). Automatic metadata extraction from Iranian theses and dissertations. In 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), pp. 1-5. 10.1109/ICSPIS51611.2020.9349570
[62] 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. In Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 61-64. https://doi.org/10.1145/2756406.2756946
[63] Amarnadh, S., Rao, V.S., Prasad, M.R., Rao, V.V. (2016). A study on meta data extraction systems and features of cloud monitoring. Journal of Computer Science IJCSIS, pp. 131-135.
[64] 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. https://doi.org/10.4018/JDM.2018040101
[65] 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 (IJDAR), 18(4): 317-335. https://doi.org/10.1007/s10032-015-0249-8
[66] Prasath, R.R., Öztürk, P. (2016). An approach to content extraction from scientific articles using case-based reasoning. Res. Comput. Sci., 117: 85-96. https://doi.org/10.13053/rcs-117-1-7
[67] Chenet, M. (2017). Identify and extract entities from bibliography references in a free text. Master's thesis, University of Twente. https://purl.utwente.nl/essays/73817.
[68] Liu, R., Gao, L., An, D., Jiang, Z., Tang, Z. (2017). Automatic document metadata extraction based on deep networks. In National CCF Conference on Natural Language Processing and Chinese Computing, pp. 305-317. https://doi.org/10.1007/978-3-319-73618-1_26
[69] Ahmed, M.W., Afzal, M.T. (2020). FLAG-PDFe: Features oriented metadata extraction framework for scientific publications. IEEE Access, 8: 99458-99469. https://doi.org/10.1109/ACCESS.2020.2997907
[70] 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. https://doi.org/10.1177/0165551514551903
[71] Zhou, Q., Jiang, Z., Yang, F. (2020). Sentences similarity based on deep structured semantic model and semantic role labeling. In 2020 International Conference on Asian Language Processing (IALP), pp. 40-44. https://doi.org/10.1109/IALP51396.2020.9310496
[72] 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. In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, pp. 99-108. https://doi.org/10.1145/3197026.3197048
[73] Alamoudi, A., Alomari, A., Alwarthan, S. (2021). A rule-based information extraction approach for extracting metadata from PDF books. ICIC express letters. Part B, Applications: An International Journal of Research and Surveys, 12(2): 121-132. https://doi.org/10.24507/icicelb.12.02.121
[74] Alghamdi, H., Dawwas, W., Almutairi, T.H. (2022). Extracting ToC and metadata from PDF books: A rule-based approach. ICIC Express Letters, Part B: Applications, 13(2): 1-10. https://doi.org/10.24507/icicelb.12.02.121
[75] Ahmad, M., Qadir, M.A., Rahman, A., Zagrouba, R., Alhaidari, F., Ali, T., Zahid, F. (2020). Enhanced query processing over semantic cache for cloud based relational databases. Journal of Ambient Intelligence and Humanized Computing, pp. 1-19. https://doi.org/10.1007/s12652-020-01943-x
[76] Rahman, A., Alhaidari, F.A. (2018). Querying RDF data. Journal of Theoretical and Applied Information Technology, 26(22): 7599-7614.
[77] Alhaidari, F.A. (2019). The digital library and the archiving system for educational institutes. Pakistan Journal of Information Management and Libraries, 20: 94-117.
[78] Faisal, H.M., Tariq, M.A., Alghamdi, A., Alowain, N. (2019). A query matching approach for object relational databases over semantic cache. In Application of Decision Science in Business and Management. IntechOpen. ISBN: 978-1-83880-099-4.
[79] Khan, S.N., Nawi, N.M., Imrona, M., Shahzad, A., Ullah, A., Rahman, A.R. (2018). Opinion mining summarization and automation process: A survey. International Journal on Advanced Science Engineering Information Technology, 8(5): 1836-1844. https://doi.org/10.18517/ijaseit.8.5.5002
[80] Faisal, H.M., Ahmad, M., Asghar, S., Rahman, A. (2017). Intelligent quranic story builder. International Journal of Hybrid Intelligent Systems, 14(1-2): 41-48. https://doi.org/10.3233/HIS-170241
[81] Rahman, A., Ahmed, M., Zaman, G., Iqbal, T., Khan, M.A.A. et al. (2022). Geo-Spatial Disease Clustering for Public Health Decision Making. Informatica, 46(6): 21-32. https://doi.org/10.31449/inf.v46i6.3827
[82] Ahmed, M.I.B., Rahman, A.U., Farooqui, M., Alamoudi, F., Baageel, R., Alqarni, A. (2021). Early identification of COVID-19 using dynamic fuzzy rule-based system. Mathematical Modelling of Engineering Problems, pp. 805-812.
[83] Ur, A., Rahman, S., Naseer, I., et al. (2021). Supervised machine learning-based prediction of COVID-19. Computers, Materials and Continua, 69(1): 21-34. https://doi.org/10.32604/cmc.2021.013453
[84] Mahmud, M., Haq, I.U. (2021). Information security in business: A bibliometric analysis of the 100 top cited articles. Library Philosophy and Practice, pp. 1-49.
[85] Alhaidari, F.A., Musleh, D., Mahmud, M., Khan, M.A. (2019). Synchronization of virtual databases: A case of smartphone contacts. Journal of Computational and Theoretical Nanoscience, 16(5-6): 1740-1757. https://doi.org/10.1166/jctn.2019.8115
[86] Atta-ur-Rahman, F.A.A. (2019). An Electronic Data Interchange Framework for Educational Institutes. ICIC Express Letters, 13(9): 831-840. https://doi.org/10.24507/icicel.13.09.831
[87] Ahmad, M., Farooq, U., Rahman, A., Alqatari, A., Dash, S., Luhach, A.K. (2019). Investigating TYPE constraint for frequent pattern mining. Journal of Discrete Mathematical Sciences and Cryptography, 22(4): 605-626. https://doi.org/10.1080/09720529.2019.1637158
[88] Rahman, A. (2018). Efficient decision based spectrum mobility scheme for cognitive radio based V2V communication system. Journal of Communications, 13(9): 498-504. https://doi.org/10.12720/JCM.13.9.498-504
[89] Alhiyafi, J. (2018). Health level seven generic web interface. Journal of Computational and Theoretical Nanoscience, 15(4): 1261-1274. https://doi.org/10.1166/jctn.2018.7302
[90] Musleh, D., Ahmed, R., Alhaidari, F. (2019). A novel approach to Arabic keyphrase extraction. ICIC express letters. Part B, Applications: An International Journal of Research and Surveys, 10(10): 875-884. https://doi.org/10.24507/icicelb.10.10.875
[91] Rahman, A.U., Alqahtani, A., Aldhafferi, N., Nasir, M.U., Khan, M.F., Khan, M.A., Mosavi, A. (2022). Histopathologic oral cancer prediction using oral squamous cell carcinoma biopsy empowered with

transfer learning. Sensors, 22(10): 3833. https://doi.org/10.3390/s22103833
[92] Rahman, A.U., Abbas, S., Gollapalli, M., et al. (2022). Rainfall prediction system using machine learning fusion for smart cities. Sensors, 22(9): 3504. https://doi.org/10.3390/s22093504
[93] Gollapalli, M.A., Chabani, S. (2022). Modeling and verification of aircraft takeoff through novel quantum nets. Computers, Materials and Continua, 72(2): 3331-3348. https://doi.org/10.32604/cmc.2022.025205
[94] Ibrahim, N.M., Gabr, D.G.I., Rahman, A., Dash, S., Nayyar, A. (2022). A deep learning approach to intelligent fruit identification and family classification. Multimedia Tools and Applications, pp. 1-16. https://doi.org/10.1007/s11042-022-12942-9
[95] Gollapalli, M., Rahman, A., Musleh, D. et al. (2022). A neuro-fuzzy approach to road traffic congestion prediction. Computers, Materials and Continua, 72(3): 295-310.
[96] Rahman, A., Mahmud, M., Iqbal, T., et al. (2022). Network anomaly detection in 5G networks. Mathematical Modelling of Engineering Problems, 9(2): 397-404. https://doi.org/10.18280/mmep.090213
[97] Ghazal, T.M., Al Hamadi, H., Umar Nasir, M., Gollapalli, M., Zubair, M., Adnan Khan, M., Yeob Yeun, C. (2022). Supervised machine learning empowered multifactorial genetic inheritance disorder prediction. Computational Intelligence and Neuroscience. https://doi.org/10.1155/2022/1051388
[98] Nasir, M.U., Ghazal, T.M., Khan, M.A., et al. (2022). Breast cancer prediction empowered with fine-tuning. Computational Intelligence and Neuroscience. https://doi.org/10.1155/2022/5918686
[99] Alhaidari, F., Almotiri, S.H., Al Ghamdi, M.A., Khan, M.A., Rehman, A., Abbas, S., Khan, K.M. (2021). Intelligent software-defined network for cognitive routing optimization using deep extreme learning machine approach. Computers, Materials & Continua, 67(1): 1269-1285. https://doi.org/10.32604/cmc.2021.013303
[100] Alhaidari, F., Shaib, N.A., Alsafi, M., et al. (2022). ZeVigilante: Detecting Zero-Day malware using machine learning and sandboxing analysis techniques. Computational Intelligence and Neuroscience. https://doi.org/10.1155/2022/1615528