ICIC Express Letters
Volume **14**, Number **6**, June **2020**

 ISSN 1881-803X
pp. **593–603**

INFORMATION EXTRACTION FROM SEMI AND UNSTRUCTURED DATA SOURCES: A SYSTEMATIC LITERATURE REVIEW

Gohar Zaman[1], Hairulnizam Mahdin[1], Khalid Hussain[2] and Atta-ur-Rahman[3,*]

[1]Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
Parit Raja, Batu Pahat, Johor 86400, Malaysia
Gi180024@siswa.uthm.edu.my; hairuln@uthm.edu.my

[2]Barani Institute of Sciences (Sahiwal)
PMAS Arid Agriculture University
Rawalpindi 46000, Pakistan
Dr.khalid@baraniinstitute.edu.pk

[3]Department of Computer Science
College of Computer Science and Information Technology (CCSIT)
Imam Abdulrahman Bin Faisal University
P.O. Box 1982, Dammam 31441, Saudi Arabia
[*]Corresponding author: aaurrahman@iau.edu.sa

Received September 2019; accepted December 2019

Abstract. *Millions of structured, semi structured and unstructured documents have been produced around the globe on a daily basis. Sources of such documents are individuals as well as several research societies like IEEE, Elsevier, Springer and Wiley that we use to publish the scientific documents enormously. These documents are a huge resource of scientific knowledge for research communities and interested users around the world. However, due to their massive volume and varying document formats, search engines are facing problems in indexing such documents, thus making retrieval of information inefficient, tedious and time consuming. Information extraction from such documents is among the hottest areas of research in data/text mining. As the number of such documents is increasing tremendously, more sophisticated information extraction techniques are necessary. This research focuses on reviewing and summarizing existing state-of-the-art techniques in information extraction to highlight their limitations. Consequently, the research gap is formulated for the researchers in information extraction domain.*
Keywords: Information extraction, Semi structured, Unstructured documents, Digital libraries, Retrieval

1. **Introduction.** Millions of structured, semi structured and unstructured documents have been produced around the globe on a daily basis [1]. Thousands of research societies like IEEE, ACM, and Springer exist and publish scientific documents tremendously [2]. For example, until 2017 IEEE contains 4.5 million [3] documents in their database, while Elsevier publishes more than 430,000 articles annually in 2,500 journals and its archives contain over 13 million [4] documents and Wiley Online Library has more than 4 million articles [5]. They all contain some piece of information that is needed by research community and interested parties. However, due to the massive volume and verifying document formats, search engines are facing problems in indexing such documents, thus making retrieval of information inefficient and time consuming [6-8]. There is a strong need of research to overcome this problem, as the documents volume is piling up rapidly and with the incoming of many new sources of the publications [9-11].

DOI: 10.24507/icicel.14.06.593

Various techniques have been investigated in this regard. Among these techniques ontology based information extraction techniques are becoming popular. The desired structure is defined in terms of ontology [12]. Ontology is collection of related concepts about an object. In terms of information extraction, desired structure can easily be defined in terms of ontology [12]. Fuzzy systems, on the other hand, have gained a lot of attention of the researchers in various fields like science and engineering. These systems are popular due to their suitability in the situations that involve approximations and less accuracy [13,14]. Such systems may play a vital role in the information extraction. Moreover, information extraction involves natural language processing (NLP) and its module word sense disambiguation (WSD) to mitigate the inherent ambiguity of natural languages [14]. Many researchers have been investigating various techniques in information extraction in various fields. In this regard machine learning and data mining, CRF (conditional random field) and hybrid techniques are related to extraction of structural information from unstructured/semi structured published scientific articles [1,2,15].

This paper focuses upon chronologically reviewing the work done in information extraction from semi and unstructured scientific documents for sake of creation of a digital library for and archiving system to help the search engines and researchers.

The rest of the paper is organized as follows. Section 2 contains the literature review on information extraction and the approaches/techniques that have been employed in this regard. Section 3 narrates the potential applications of information extraction. Section 4 highlights the common issues in information extraction while Section 5 concludes the paper.

2. Existing Techniques in Information Extraction.

Information extraction (IE) is the process of automatically extracting structured information from unstructured and/or semi structured machine-readable documents. In most of the cases this activity involves processing human language texts by means of text mining, pattern matching and natural language processing (NLP) or similar techniques [15].

From the brief literature survey [12,16,17] it is apparent that the majority work done in the field of information extraction is based on a specific format, specific set rules and specific types of documents mostly available in unstructured and semi structured format [16]. Most of the documents for information extraction are web-based documents. However, according to [17] most of the published work is comprised of PDF and text documents. Although various techniques exist [9,17,18] that address this issue, they are developed to address a narrow domain and with very specific rules that are only applicable to limited formats specific to that community. Nonetheless, when these techniques are investigated over slightly modified formats, their performance is compromised [11].

According to the scientific and research communities a variety of published work is available in unstructured and semi structured form, mainly in text files like PDF and word documents [11]. Nevertheless, the concerning systematic part of the extraction for the most wanted knowledge (usually complete structure of the article or a custom set of desired attributes) from such documents is not a simple and easy task. This is due to diverse formatting standards followed by different research communities. For that reason, the compulsory step is the knowledge about documents pattern. It also required constructing the algorithms which help to minimize the variance among the undergoing text documents with their system identifiable depiction. This work is summarized in Table 1 with clarification about the technique used and the primary objectives achieved.

In this field of research, various methods have been developed to extract information either from XML documents [21-25] or from plain-text document [26-29]. None of these approaches utilize the patterns in both XML and text formats to identify the desired information of the published research articles. In support of a robust solution, using only one of these formats is not enough.

In [30], authors proposed an RBS based method for information extraction from scientific documents in the form of XML or simple text. The empirical tests were performed on XML files with the help of PDFx online tool. Authors built an ontology and utilized a rule-based approach after crafting the rules by observing the given dataset documents. The technique possessed an accuracy of 77.5%. However, the major limitation of the technique was that it worked for only one specific conference format, specific corpus and with slight modification in the format the algorithm might not be effective.

2.1. **Information extraction (IE) approaches.** In various empirical applications, like IE in medical field, the main task is related to a known series of examples. IE categorizes the sub fields of a specified example which describes the significant knowledge. There are some advantages of IE such as incorporation about the product knowledge obtained from several sources, multiple question answers sessions, research about contact knowledge, search out the main focusing parameters in medical field and delete the raucous data from the original data for achieving good results.

Specifically, there are three advanced approaches which are rule-based knowledge, classification-based structure and labeling with chronological pattern. These are supervised learning methods in which two levels operate for any performing tasks (Figure 1). First level is extraction in which sub categories of any provided example dig out through structure of learning models according to the numerous designs. Now the mined data is known as interpreted form of data. In interpreted data, explicit knowledge about the performing task is pre-defined meta-data. The second level of IE is training. Different types of models are designed for exposure of sub categories in training phase. In every designed model, stimuli can be considered as a series of examples. A document may be examined like a series of words or text lines in any application. The taxonomy of information extraction is shown in Figure 1. Among these techniques, ontology-based methods are more appropriate for sake of information extraction from scientific semi-structure documents. This is mainly because ontological frameworks represent the exact document format. That is why this approach is considered in this research.

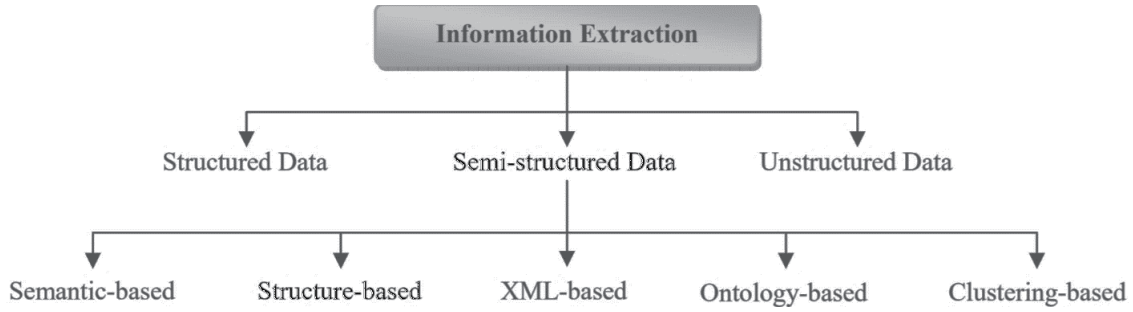

FIGURE 1. Taxonomy of information extraction

Following section highlights the advanced techniques.

2.2. **Rule based information extraction methods.** In these methods some rules are crafted and based on those rules' information is extracted from the given sources. Therefore, these approaches are divided into three classes:

1) Vocabulary support approach
2) Rule support approach
3) Wrapper orientation

2.2.1. *Vocabulary support approach.* Classical knowledge taking out systems built a pattern of vocabulary. Researchers used this stencil vocabulary for extraction process of required knowledge from novel data. These types of systems including AutoSlog [31],

AutoSlog-TS [32], and CRYSTAL [33] are known as dictionary support or sometimes pattern recognition systems. In such systems, a major question arises that how these vocabularies map the pertinent prototypes about knowledge, which is the most critical part of desired task. The initial structure was AutoSlog which recognized the vocabulary about text from different training models. It designs a vocabulary that was about taking out the advance trends of text as concept nodes. This vocabulary is known as concept nodes in which every node keeps some type of concept about any word. The concept is having an anchor for each node which turns on with two major parts like a linguistic prototype and a set of facilitating circumstances. These two parts provide the applicability of the given system. An anchor is a word that behaves as a trigger. Facilitating circumstances correspond to some set of restrictions which are applicable on linguistic advance trends with their parts. In [34], authors proposed a new approach for designing a vocabulary along some texts, called seed words. These seed words are useful for classification of upcoming words which goes to the accurate class with identical pattern. The vocabulary trends may be examined increasingly.

2.2.2. *Rule support approach.* In this category, rules are concerning area rather than vocabulary. Such kinds of approaches are normally used for semi-supervised data such as web-pages. There are two algorithms which are designed for explicit semi-supervised data. Firstly, bottom-up considers the extraordinary issues and converts these issues into common one. Secondly, top-down focus considers common issues and learns rules about this category. Several researchers provide better results regarding this area of research including [35-38]. The main issue with the rule support approach is that it is highly specific to the document type.

2.2.3. *Wrapper orientation.* This is another subcategory of rules in which supervised and semi-supervised data consider for work at the same time. A wrapper is a data mining process while it keeps a set of rules associated with extraction and needs a piece of program for deployment of these rules. It is an automatic method in which a training dataset is used in orientation of wrapper algorithm as detection of target knowledge. Few authors present their ideas about this approach along distinctive wrapper model designs like WEIN [39], Stalker [40] and BWI [41].

2.3. **Categorization supports extraction approaches.** Under this class, IE novel approaches are discussed through supervised learning. The primary thought is about IE issue just like perfect categorization. Inside this segment, author explains the approaches for IE categorization with covering all aspects. Here is an example of two class classification issue. Let us first consider a two-class classification problem. Let $\{(\boldsymbol{a_1}, b_1) \ldots (\boldsymbol{a_n}, b_n)\}$ be a training dataset in which $\boldsymbol{a_p}$ represents a feature vector and $b_p \in \{-1, +1\}$, $1 \leq p \leq n$ belongs to a classification tag. As a rule of classification design, there are two levels, that is, knowledge and forecasting. According to knowledge, a design can be searched for labeled dataset which is split with training data. On the other hand, forecast level is used for well-read design and that design classifies the unlabeled dataset sometimes this prediction provides numerical results and sometimes results are in the form of rules series.

The mainly famous approach for categorization is support vector machine (SVM) while this method is used for designing phase in [42]. Supposedly, linear classifier attains results according to the described model in [42] must keep some generality anomalies. With the passage of time, linear SVM enhanced its working for non-linear conditions regarding problems so this type of linear SVM is known as non-linear SVM. Non-linear SVM is having different types of functions according to these references [43-45]. This is all about two class problems. If problem exceeds from two classes, then researchers exploit other approaches such as "one class versus all others". Later, many variants of SVM were introduced to enhance the process like,

- Boundary recognition by categorization structure
- Improvement in IE via a two-class margin
- Improvement in IE as a result of unbalance categorization design

2.4. **Chronological labeling approaches for extraction.** For any extraction activity, some set of rules is a primary need for IE for example according to the meta-data extraction processes on research articles [46] and some labels are also considered as a primary task. Here a document is judged like a surveillance set of series x. The unit of this surveillance set is a small part of a document that can be a word, a text line or any other part of the document. So, the activity is searching a label series of y. Therefore, conditional probability $p(y|x)$ gives the maximum results through designed approach mentioned above. Meta-data extraction task along conditional random fields (CRFs) can be used like features with all aspects. Consequently, dependent and random features are capable of partiality design. Sometimes these features directly work with several features. On the other hand, sometimes these features execute fewer operations in designing level. Orientation of a feature may be performing simultaneously with training session [47]. Further work in this regard can be categorized as:

- Non-linear CRFs
- CRFs used for relational knowledge
- 2-D CRFs used for web extraction of specific knowledge
- Active CRFs & Tree-structure CRFs

2.5. **Some futuristic approaches.** Machine learning and artificial intelligence have been the setting new standards in every field of study. IE is also among such domains. The major techniques are given as follows:

- Deep leaning/neural network
- Fuzzy system
- Ontologies

2.5.1. *Deep leaning/neural network.* Artificial neural networks (ANN) along with their deep learning (DL) counterparts has been tremendously used in many fields of study like data mining, prediction, classification and optimization. The traditional ANN algorithms require more samples and slow learning times and can overfit the learning model [70]. The idea of DL specified by [71] learns fast and it is efficient in the cost of computational complexity.

2.5.2. *Fuzzy systems.* Fuzzy systems have been proposed, investigated and proven their versatility for various areas of research over past many decades. Their applications are limitless and are varying from control system [13], engineering [14] to data mining [55] and much more. Fuzzy systems are promising in terms of efficiency where the information is fully available and where a factor of uncertainty is involved.

2.5.3. *Ontologies.* Ontologies are generic formal specification of the terms (words/concepts/objects) in a specific domain and their relations. In recent years the development of ontologies has been moving from the realm of artificial-intelligence laboratories to the desktops of domain experts [56]. Ontologies have become common on the web also referred to as semantic web [57]. The ontologies on the web range from large taxonomies categorizing web sites (such as on Google) to categorizations of products for sale and their features (such as on Amazon). Some of the reasons for creating ontology are:

- Sharing common understanding of the structure of information among the stakeholder
- Enabling reuse of domain knowledge
- Making domain specific rules, generic and analysis

There are several techniques in IE with their own pros and cons as well as their own target area and domain in terms of the scientific document or reports, etc. The most common approaches developed over the time in the literature are enlisted in Table 1.

TABLE 1. The work related to information extraction

Paper Title	Techniques	Author Name	Main Idea	Primary Objectives
A Hybrid Approach for Scholarly Information Extraction	Hybrid Technique of Clustering and Classification	Bodó and Csató (2017) [9]	To extract the meta-data of the research paper using dictionary & font-based information	Extract the title and author name of the research paper
Using Text Mining Techniques for Extracting Information from Research Articles	Clustering, Word Cloud, ASRM, Visualization, Similarity Measure, Term Frequency	Salloum et al. (2018) [1]	Extract the interesting topics from 300 journals of 6 different major database	Focus on extracting interesting topics from research papers based on the distance base mobile learning in HE
Framework for Automatic Information Extraction from Research Papers on Nanocrystal Devices	NaDevEx AIE System	Dieb et al. (2015) [20]	Making rules using CRF techniques and finding patterns to extract information	Extracting meaningful information from research papers based on nanocrystal devices
Logical Structure Recovery in Scholarly Articles with Rich Document Features	CRF Technique	Luong et al. (2012) [6]	Extracting the information by identifying the font size of research paper using OCR	Extract the logical structure of paper like author name, affiliation, and section
Parsing Citations in Biomedical Articles Using Conditional Random Fields	Conditional Random Field	Zhang et al. (2011) [18]	Extract the citation of the research paper using conditional random field	Extraction of citation include author name, title, source of publication, volume, pages, year
Extracting and Matching Authors and Affiliations in Scholarly Documents	Enlil (name of technique) Information Extraction System using CRF and use of SVM	Do et al. (2013) [7]	First extract author name & affiliation using CRF and connect them using SVM	Focus on extracting authors name and their affiliation
Semi-automatic Meta-data Extraction from Scientific Journal Article for Full-text XML Conversion	Rule Base Method & CRF	Kim et al. (2014) [8]	Making rules using CRF techniques and finding patterns to extract information	Extracting the meta-data on first and last pages of paper like title, author details, abstract and references

3. **Applications of Information Extraction.** There is a huge number of applications of information extraction. Here few of them are enlisted that are most frequently and widely used.

3.1. **Information extraction in digital libraries (DL).** Meta-data is a form of ordered data which is used for searching process of different types of documents, images, patterns and trends for users in DL. According to the meta-data, search-engines provide the facility of information recovery with high accuracy. For the generation of meta-data, many researchers, scientists and librarians spend their precious time on this creation by hand while this manual meta-data conversion process into automatic form is another hard task which depends on IE [48,59-69].

3.2. **Extraction process on individual report.** Every community keeps the organization system of individual information and considers this subject as a key note. Individual information reviews the important parts of the information related to profile, contact, social circle and personal webpage [49].

3.3. **Table extraction through CRFs.** Tokens are arranged in the combination of rows and columns which is known as table. In this manner, table consists of tokens which are in the form of text. Researchers represent their work in [50] in the shape of table extraction design in which verification step executes through CRFs for content and its design. In the description of the reference, proposed structure places the tables in the form of simple text with statistical summaries and their labels along with their tokens [30].

3.4. **Shallow parsing with CRFs.** In the simple text, shallow parsing classifies several phrases in non-recursive form. It is promising to fulfill the requirement of parsing or may be IE in [51]. The basic idea of NP chunking is, along some standard datasets with their assessment metrics described in [52]. The enhanced version of above mentioned work was a type of shared activity for CoNLL-2000 [53]. The work in [54] makes use of CRFs into shallow parsing and obtains a detailed pragmatic literature on various forms of chronological labeling methods.

4. **Common Issues in Information Extraction.** From the literature review, it is apparent that the rule-based information extraction methods are better from the others; however, they still lack in terms of incorporation of diversity of the document formats. From the literature review, we have figured out some common issues with information extraction and where still researchers are encouraged to work.

1) **No generalized mechanism for all the domains**
   It is observed that each approach is a customized approach and highly domain specific, rather than generic. For sake of information extraction, one must define a separate approach every time the document type is changed. This is one of the major limitations observed. Even a scheme performing excellent in one way, may not be good at all for a slightly modified version of the document being provided.
2) **Every time need to build a domain specific model/framework for IE.**
   Building a model is very important in the process of information extraction because it is generally comprised of several segments like parsing, tokenizing, and pre-processing that may vary from domain to domain, document to document, format to format (in which the document is presented like XML, and text). No generic model exists that addresses more than one domain. So, the models or frameworks are highly domain specific.
3) **There are no unified IE goals. They may vary over the time and usually custom based.**
   As far as the goals of IE concerned, they may vary from domain to domain, document to document. Sometimes, we are just looking for the structural information, sometimes a specific piece of information from the whole document like just keywords, sometimes just references, etc. For example, in emails, we might be looking for different information, in reports there could be a different outcome.
4) **No technique is win-win.**
   As far as evaluation of the techniques is concerned, they are highly subjective. One cannot say one technique is universally good or bad. The criteria for evaluation of the techniques may vary and same is the case of figure of merits.
5) **Type of document (text, PDF, email, HTML) is a critical factor.**
   Techniques of information extraction are highly depending upon the type of document whether it is text, PDF or HTML. Even more critical, if the PDF version is old and/or the document is a scanned version with a lot of noise induced by the scanner, etc. In this case addition measures must be taken like optical character recognition (OCR) that comes with its own limitation to certain fonts and often ends up with compromised accuracy due to noise and other factors induced by the scanner, camera or related device.

6) **Overwhelming type and nature of documents and their corpora made the job even difficult.**
In the era of information, hundreds and thousands of new documents, with different formats have been producing on the daily basis, information extraction is becoming more and more complex. Scientific communities have been producing structured, semi structured and unstructured documents with a tremendous speed. This is making the job very difficult for the search engines to index such rapidly growing documents and for the researchers to explore the precise documents over the web.
7) **Ambiguities in NLP**
It is obvious that the documents being considered for whatever type of information extraction consist of text written in natural languages. Natural languages are inherently ambiguous. Even with quite sophisticated techniques available for processing, it is still an open area of research, especially, when it comes to the non-native writers.
8) **Multi-Lingual documents**
English is not the only language being used nowadays for producing the documents of interest, but other languages are also being rapidly used for scientific and other types of writing. Even there are journals that publish in Korean, Chinese, Arabic and other languages. Information extraction of such documents is even challenging. Even within one language there are many dialects being followed as a standard practice.
9) **Nonstandard languages**
Internet being a source of voluminous data production may not comprise only standard documents written in a well-defined natural language. Instead, the social media related documents or texts like tweets in Twitter, users' feedbacks about a product in Amazon online store, and reviews on some issue in Facebook mainly comprise naïve, non-standard and slang text format. Now information extraction from such types of sources is even tedious. Nonetheless, this is one of the hottest areas of research because this information is critical for many reasons like for e-commerce, finding trends about a product, a person, electoral campaign and a team.

5. **Conclusion.** Information extraction is one of the hottest areas of research in data and text mining for digital libraries. This technique is mainly used to extract structural and other important information from semi structured and unstructured documents mainly over the web. This paper is dedicated to overviewing the state-of-the-art techniques that are in practice and to highlight their limitations. The objective is to find the areas in information extraction that need improvement; to help the researchers and scholars of the field. In future, we are aiming to propose our own technique for improved information extraction for digital libraries and information retrieval using an intelligent hybrid technique.

REFERENCES

[1] S. A. Salloum, M. Al-Emran, A. A. Monem and K. Shaalan, Using text mining techniques for extracting information from research articles, in *Intelligent Natural Language Processing: Trends and Applications*, Springer, 2018.
[2] F. Peng and A. McCallum, Information extraction from research papers using conditional random fields, *Inf. Process. Manag.*, vol.42, no.4, pp.963-979, 2006.
[3] *IEEE*, Available: https://ieeexplore.ieee.org, [Accessed: 12-Aug-2018].
[4] RELX Group, *RELX Group Company Reports*, 2017 RELX Group Annual Report, 2017.
[5] *Wiley*, Wiley Online Library, [Accessed: 12-Aug-2018].
[6] 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*, IGI Global, 2012.
[7] H. H. N. Do, M. K. Chandrasekaran, P. S. Cho and M. Y. Kan, Extracting and matching authors and affiliations in scholarly documents, *Proc. of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp.219-228, 2013.

[8] S. Kim, Y. Cho and K. Ahn, Semi-automatic metadata extraction from scientific journal article for full-text XML conversion, *Proc. of the International Conference on Data Mining (DMIN)*, p.1, 2014.
[9] Z. Bodó and L. Csató, A hybrid approach for scholarly information extraction, *Stud. Univ. Babes-Bolyai, Inform.*, vol.62, no.2, 2017.
[10] P. Groth, M. Lauruhn, A. Scerri and R. Daniel, Open information extraction on scientific text: An evaluation, *Proc. of the 27th International Conference on Computational Linguistics*, pp.3414-3423, 2018.
[11] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data*, Cambridge University Press, New York, NY, United States, 2006.
[12] S. T. R. Rizvi, D. Mercier, S. Agne, S. Erkel, A. Dengel and S. Ahmed, *Ontology-Based Information Extraction from Technical Documents*, 2018.
[13] 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.1, pp.801-810, 2016.
[14] 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.
[15] K. Jayaram and K. Sangeeta, A review: Information extraction techniques from research papers, *Proc. of IEEE Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, pp.56-59, 2017.
[16] J. Chen, C. Zhang and Z. Niu, A two-step resume information extraction algorithm, *Math. Probl. Eng.*, vol.2018, 2018.
[17] 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, 2014.
[18] Q. Zhang, Y.-G. Cao and H. Yu, Parsing citations in biomedical articles using conditional random fields, *Computers in Biology and Medicine*, vol.41, no.4, pp.190-194, 2011.
[19] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek and Ł. Bolikowski, CERMINE: Automatic extraction of structured metadata from scientific literature, *Int. J. Doc. Anal. Recognit.*, vol.18, no.4, pp.317-335, 2015.
[20] 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, p.1872, 2015.
[21] X. Li, *The Comparison of QlikView and Tableau: A Theoretical Approach Combined with Practical Experiences*, UHasselt, 2015.
[22] 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, 1996.
[23] S. Jebbara and P. Cimiano, Aspect-based sentiment analysis using a two-step neural network architecture, *Semantic Web Evaluation Challenge*, pp.153-167, 2016.
[24] 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, 2009.
[25] 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, 2011.
[26] A. Agarwal, F. Biadsy and K. R. Mckeown, Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams, *Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp.24-32, 2009.
[27] A. C. E. S. Lima, L. N. de Castro and J. M. Corchado, A polarity analysis framework for Twitter messages, *Appl. Math. Comput.*, vol.270, pp.756-767, 2015.
[28] M. Hu and B. Liu, Mining and summarizing customer reviews, *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.168-177, 2004.
[29] A. Krouska, C. Troussas and M. Virvou, The effect of preprocessing techniques on Twitter sentiment analysis, *The 7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pp.1-5, 2016.
[30] R. Ahmad, M. T. Afzal and M. A. Qadir, Information extraction from PDF sources based on rule-based system using integrated formats, *Semantic Web Evaluation Challenge*, pp.293-308, 2016.
[31] E. Riloff, Automatically constructing a dictionary for information extraction tasks, *Proc of the 11th National Conference on Artificial Intelligence*, 1993.
[32] E. Riloff, Automatically generating extraction patterns from untagged text, *Proc. of the National Conference on Artificial Intelligence*, pp.1044-1049, 1996.
[33] S. Soderland, D. Fisher, J. Aseltine and W. Lehnert, CRYSTAL: Inducing a conceptual dictionary, *arXiv Prepr. C.*, 1995.
[34] E. Riloff and R. Jones, Learning dictionaries for information extraction by multi-level bootstrapping, *AAAI/IAAI*, pp.474-479, 1999.
[35] F. Ciravegna, $(LP)^2$, an adaptive algorithm for information extraction from Web-related texts, *Proc. of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*, 2001.

[36] J. Tang, J. Li, H. Lu, B. Liang, X. Huang and K. Wang, iASA: Learning to annotate the semantic Web, *Journal on Data Semantics IV*, pp.110-145, 2005.
[37] R. Mooney, Relational learning of pattern-match rules for information extraction, *Proc. of the 16th National Conference on Artificial Intelligence*, vol.334, 1999.
[38] D. Freitag, Information extraction from HTML: Application of a general machine learning approach, *AAAI/IAAI*, pp.517-523, 1998.
[39] N. Kushmerick, D. S. Weld and R. Doorenbos, *Wrapper Induction for Information Extraction*, University of Washington, Seattle, WA, United States, 1997.
[40] I. Muslea, S. Minton and C. Knoblock, Stalker: Learning extraction rules for semistructured, Web-based information sources, *Proc. of AAAI-98 Workshop on AI and Information Integration*, pp.74-81, 1998.
[41] D. Freitag and N. Kushmerick, Boosted wrapper induction, *AAAI/IAAI*, pp.577-583, 2000.
[42] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
[43] B. E. Boser, I. M. Guyon and V. N. Vapnik, A training algorithm for optimal margin classifiers, *Proc. of the 5th Annual Workshop on Computational Learning Theory*, pp.144-152, 1992.
[44] R. Soentpiet, *Advances in Kernel Methods: Support Vector Learning*, MIT Press, 1999.
[45] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Science & Business Media, 2013.
[46] F. Peng and A. McCallum, Accurate Information Extraction from Research Papers Using Conditional Random Fields, *Information Processing and Management: An International Journal*, 2006.
[47] A. McCallum, Efficiently inducing features of conditional random fields, *Proc. of the 19th Conference on Uncertainty in Artificial Intelligence*, pp.403-410, 2002.
[48] L. Steve, G. Lee and B. Kurt, Digital libraries and autonomous citation indexing, *IEEE Comput.*, vol.32, no.6, pp.67-71, 1999.
[49] R. C. Berwick, Principles of principle-based parsing, in *Principle-Based Parsing*, Springer, 1991.
[50] N. Ducheneaut and V. Bellotti, E-mail as habitat: An exploration of embedded personal information management, *Interactions*, vol.8, no.5, pp.30-38, 2001.
[51] S. P. Abney, Parsing by chunks, *Principle-Based Parsing*, Springer, pp.257-278, 1991.
[52] L. A. Ramshaw and M. P. Marcus, Text chunking using transformation-based learning, in *Natural Language Processing Using Very Large Corpora*, Springer, 1999.
[53] E. F. Tjong Kim Sang and S. Buchholz, Introduction to the CoNLL-2000 shared task: Chunking, *Proc. of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning*, vol.7, pp.127-132, 2000.
[54] F. Sha and F. Pereira, Shallow parsing with conditional random fields, *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, vol.1, pp.134-141, 2003.
[55] A. Rahman and S. Das, Data mining for student's trends analysis using Apriori algorithm, *Int. J. Control Theory Appl*, vol.10, no.18, pp.107-115, 2017.
[56] D. Allemang and J. Hendler, *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*, Elsevier, 2011.
[57] L. Yu, Social networks and the semantic Web, in *A Developer's Guide to the Semantic Web*, Springer, 2014.
[58] E. Agirre, O. L. De Lacalle, A. Soroa and I. Fakultatea, Knowledge-based WSD and specific domains: Performing better than generic supervised WSD, *IJCAI*, pp.1501-1506, 2009.
[59] Atta-ur-Rahman and F. A. Alhaidari, Querying RDF data, *Journal of Theoretical and Applied Information Technology*, vol.96, no.22, pp.7599-7614, 2018.
[60] Atta-ur-Rahman and F. A. Alhaidari, The digital library and the archiving system for educational institutes, *Pakistan Journal of Library and Information Science*, vol.20, no.1, pp.94-117, 2019.
[61] H. M. Faisal, M. Ahmad, S. Asghar and Atta-ur-Rahman, Intelligent quranic story builder, *International Journal of Hybrid Intelligent Systems*, pp.1-8, 2017.
[62] Atta-ur-Rahman and S. Das, Big data analysis for teacher recommendation using data mining techniques, *International Journal Control Theory and Applications*, vol.10, no.18, pp.95-105, 2017.
[63] N. Shahzadi, Atta-ur-Rahman and M. J. Sawar, Semantic network based classifier of Holy Quran, *International Journal of Computer Applications (IJCA)*, vol.39, no.5, pp.43-47, 2012.
[64] N. Shahzadi, Atta-ur-Rahman and A. Shaheen, Semantic network based semantic search of religious repository, *International Journal of Computer Applications (IJCA)*, vol.36, no.9, pp.1-5, 2011.
[65] Atta-ur-Rahman, Knowledge representation: A semantic network approach, in *Handbook of Research on Computational Intelligence Applications in Bioinformatics*, 1st Edition, Chapter 4, IGI Global, 2016.
[66] Atta-ur-Rahman, Teacher assessment and profiling using fuzzy rule based system and Apriori algorithm, *International Journal of Computer Applications (IJCA)*, vol.65, no.5, pp.22-28, 2013.

[67] Atta-ur-Rahman, K. Sultan, N. Aldhafferi and A. Alqahtani, Educational data mining for enhanced teaching and learning, *Journal of Theoretical and Applied Information Technology*, vol.96, no.14, pp.4417-4427, 2018.

[68] Atta-ur-Rahman, S. A. Alrashed and A. Abraham, User behavior classification and prediction using FRBS and linear regression, *Journal of Information Assurance and Security*, vol.12, no.3, pp.86-93, 2017.

[69] Atta-ur-Rahman, D. N. Zaidi, M. H. Salam and S. Jamil, User behavior classification using fuzzy rule based system, *The 13th International Conference on Hybrid Intelligent Systems (HIS'13)*, Tunisia, pp.118-123, 2013.

[70] J. Cheng, Z. Duan and Y. Xiong, QAPSO-BP algorithm and its application in vibration fault diagnosis for hydroelectric generating unit, *Journal of Vibration and Shock*, vol.34, pp.177-181, 2015.

[71] G.-B. Huang, D. H. Wang and Y. Lan, Extreme learning machines: A survey, *Int. J. Mach. Learn. Cybern.*, vol.2, pp.107-122, 2011.