EXTRACTING MEANINGFUL INFORMATION FROM SEMI AND UNSTRUCTURED DATA SOURCES WITH DIFFERENT TECHNIQUES : A SCIENTIFIC LITERATURE REVIEW

Shital M Chaniyara Atmiya University Rajkot [1],
Dr. Dipti H Domadiya National Computer College Jamnagar [2],
Dr. Stavan C Patel Atmiya University Rajkot [3],

Abstract.

Millions of structured, semi-structured, and unstructured documents are produced around the globe a day today. Several research societies like IEEE, Elsevier, Springer, and Wiley that we use to publish the scientific documents enormously and some individual's documents are sources of such data. Thanks to their massive volume and ranging document formats, search engines face problems in indexing such documents, thus retrieving inefficient, tedious, and time-consuming data. Information extraction from such documents is among the most well-liked areas of research in data/text mining. Because the number of such documents is increasing tremendously day by day on a large scale that's why proper and more sophisticated information extraction techniques are necessary to find out, this research focuses on reviewing and summarizing existing practices in information extraction to highlighting their limitations.

Keywords: Information extraction, unstructured documents, Semi-structured, Digital libraries, Retrieval

1. Introduction. Millions of structured, semi-structured, and unstructured documents are produced around the globe each day today. Several research societies like IEEE, Elsevier, Springer, and Wiley that we use to publish the scientific documents enormously and a few individual documents are sources of such data. For instance, until 2020, IEEE contains 2TB documents in their database, while 430,000 articles annually published by Elsevier in 2,500 journals and include over 13 million copies and Wiley Online Library has approx 4 million articles in it, However, due to huge volume and verifying document formats, search engines face problems in indexing such documents, thus making retrieval of data inefficient and time-consuming and documents volume is piling up rapidly and with the incoming of the many new sources of the publications.

Various techniques have been investigated in this regard of information extraction from the large volume of collected data. Ontology-based information extraction techniques are becoming popular from all these techniques. Ontology is a collection of related concepts about an object, and with the help of these techniques, the desired structure can easily be defined. In

the science and engineering field, fuzzy systems have is used by a lot of researchers. These systems are popular due to their suitability in situations that involve approximations and less accuracy.

Such systems may play an essential role in information extraction, which involves tongue processing (NLP) and its module word meaning disambiguation (WSD) to mitigate the inherent ambiguity of natural language. In this regard, machine learning and data processing, CRF (conditional random fields), and hybrid techniques are associated with extracting structural information from unstructured/semi-structured published scientific articles.

This paper focuses on chronologically reviewing the work wiped out information extraction from semi and unstructured scientific documents for the sake of creation of a digital library for and archiving system to assist the search engines and researchers. The rest of the paper is organized as follows. Section 2 contains the literature review on information extraction and, therefore, the approaches/techniques that are employed in this regard. Section 3 narrates the potential applications of data extraction, while Section 4 concludes the paper.

2. Existing Techniques in Information Extraction. Extracting structured information from unstructured and semi-structured machine-readable documents automatically this process is known as Information extraction (IE). This process involves human language text processing using text mining, pattern matching, and natural language processing (NLP) or similar techniques in most cases. The brief literature survey majority of work done in the fields of information extraction is based on a specific format, specific set rules, and specific types of documents mostly available in unstructured and semi-structured formats Majority web-based forms are used for information extraction. Although various techniques exist that address this issue, they're developed to deal with a narrow domain and very specific rules applicable to limited formats specific to its community.

According to the scientific and research communities, a spread of published work is out there in unstructured and semi-structured form, mainly in text files like PDF and word documents. Nevertheless, the systematic concerning part of the extraction for the foremost wanted knowledge (usually complete structure of the article or a custom set of desired attributes) from such documents isn't easy and straightforward. This is often thanks to diverse formatting standards followed by different research communities. For that reason, the critical step is that the knowledge about documents pattern. It also required constructing the algorithms which help to attenuate the variance among the undergoing text documents with their system identifiable depiction.

This work is summarized in Table 1 with the technique used and the primary objectives

achieved

2.1 Information extraction (IE) approaches. IE categorizes the sub fields of a specified example which describes the significant knowledge. There are some advantages of IE like the incorporation of the merchandise knowledge obtained from several sources, multiple question answers sessions, research about contact knowledge, obtain the most focusing parameters for achieving good results.

There are three main approaches: rule-based knowledge, classification-based structure, and labeling with the chronological pattern.

This approach can be categorized into four groups, as shown below.

1. Supervised learning systems approach is proposed and used in many works and requires large amounts of data to set as training data. It then uses the machine learning rules and techniques to extract the required information;

2. Semi-supervised learning systems approach (e.g., Mutual Bootstrapping);

3. Unsupervised learning systems approach is quite different from supervised learning systems for it uses corpus bootstrapping methods that depend on small seed rules that are learned from an annotated system;

4. Hybrid NER systems; this approach uses a combination of a dictionary base and Conditional Random

A document could also be examined sort of a series of words or text lines in any application. The taxonomy of ex- data traction is shown in Figure 1. Among these techniques, ontology-based methods are more appropriate for the sake of data extraction from scientific semi-structured documents. This is often mainly because ontological frameworks represent the precise document format. That's why this approach is taken into account during this research.

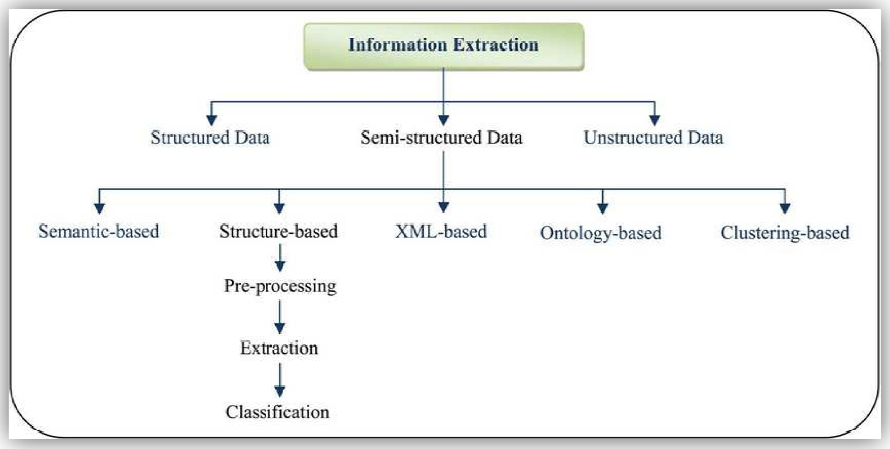

FIGURE 1. Taxonomy of information extraction

Advance techniques for information extraction.

2.2 Rule-based information extraction methods.

Based on some designed rules' information is extracted from the given sources.

Developers were encoding human knowledge into computer systems as rules that get stored during a knowledge domain. These rules define all aspects of a task, typically within the sort of "If" statements ("if A, then do B, else if X, then do Y").

While the number of rules is written based on actions you would like a system to handle (for example, 15 acts means manually writing and coding a minimum of 15 controls), rules-based systems are generally lower effort, less expensive, and fewer risky since these rules won't change or update on their own. These approaches are divided into the following categories:

1. Vocabulary support approach

2. Rule support approach

3. Wrapper orientation

2.2.1 Vocabulary support approach: In this Vocabulary support approach system built a pattern of vocabulary. Researchers used this vocabulary for the extraction process. These systems include AutoSlog, AutoSlog-TS, and CRYSTAL referred to as dictionary support or pattern recognition systems. The most crucial part of the desired task in this system is how these vocabularies map the pertinent prototypes about knowledge. The initial structure was AutoSlog which recognized the vocabulary about the text from different training models. It designs a vocabulary that was about removing the advanced trends of text as concept nodes. This vocabulary is known as concept nodes, during which each node keeps some concept about any word. The idea has an anchor for each node that activates with two major parts: a linguistic prototype and a gaggle of facilitating circumstances. These two parts provide the applicability of the given system. An anchor could also be a word that behaves as a trigger. Facilitating cases correspond to some set of restrictions that are applicable on linguistic advance trends with their parts. The authors proposed a replacement approach for designing a vocabulary and some texts, called seed words. These seed words are helpful for classification of upcoming words, which matches the precise class with identical patterns. The vocabulary trends could even be examined increasingly.

2.2.2 Rule support approach: Rules are concerning area rather than vocabulary in this approach, generally used for semi-supervised data such as web pages. Two algorithms are designed for detailed semi-supervised data. First, it will consider special bottom-up issues and converts these issues into a common one. The second thing is to focus on top-down common issues and learns rules about this category. It is highly specific that the document type is the main issue of this approach.

2.2.3 Wrapper orientation: In this approach, supervised and semi-supervised data consider for work on an equivalent time. A wrapper may be a data processing process while keeping a group of rules related to the extraction and wishes a bit of program to deploy those rules. it's an automatic method during which a training dataset is employed in the orientation of wrapper algorithm as detection of target knowledge

2.3 Categorization supports extraction approaches. In this category, through supervised learning, IE novel approaches are discussed. The first thought is about IE issues, a bit like perfect categorization. Inside this segment, converting with all aspects author explains the procedures for IE categorization. An example of two class classification issue is given here, which allows us to consider a two-class classification problem first. Let (a1, b1) . . . (an, bn) be a training dataset in which ap represents a feature vector and $b_p \in \{-1,+1,\}1 \leq p \geq$ n belongs to a classification tag. As a rule of classification design, there are two levels, that is, knowledge and forecasting. Consistent with knowledge, a method is often looked for the labeled dataset, split with training data. On the opposite hand, forecast level is employed for well-read design which design was classifies the unlabeled dataset. Sometimes, this prediction provides numerical results, and sometimes products are within the sort of rules series.

Support vector machine (SVM) is a famous approach for categorization for designing introduce. Linear classifies attains results consistent with the described model in must keep some generality anomalies. With the passage of your time, linear SVM enhanced its working for non-linear conditions regarding problems, so this sort of linear SVM is understood as non-linear SVM. Non-linear SVM has different kinds of functions consistent with these references. This is often all about two-class problems. Researchers exploit other approaches like "one class versus all others if the problem exceeds two classes." Later, many variants of SVM were introduced to reinforce the method 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 rules may be an immediate need for IE, for instance, consistent with the meta-data extraction processes on research articles. A few labels are also considered a primary task. Here a document is judged sort of a surveillance set of series x. This surveillance set may be a small part of a document which will be a word, a text line, or the other a part of the document. This activity is searching. As a result, a label series of y. Therefore, contingent probability p(y x) gives the maximum results through the designed approach mentioned above. Meta-data extraction tasks along conditional random fields (CRFs) are often used like features with all aspects. Consequently, dependent and random components are capable of partiality design. Sometimes these features directly work with several features. On the opposite hand, periodically, these features execute fewer operations at the designing level. Orientation of an element could also be performing simultaneously with a training session. Further, add this regard are often 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. In the everyday field of study for information extraction, Machine learning and artificial intelligence have been setting new standards. IE is also among such domains. Essential techniques are given as below:

- Deep leaning
- neural network
- Fuzzy system
- Ontologies

2.5.1 Deep leaning. Deep Learning – it's a branch of Machine Learning that leverages a series of non-linear processing units comprising multiple layers for feature transformation and extraction. Its several layers of artificial neural networks perform the ML process. The primary layer of the neural network processes the data input and passes the knowledge to the second layer.

The second later then processes that information further by adding additional information (for example, user's IP address) and passes it to subsequent layer. This process continues throughout all layers of the Deep Learning network until the specified results are achieved.

2.5.2 Neural Networks – a structure consisting of ML algorithms wherein the synthetic neurons make the core computational unit that focuses on uncovering the underlying patterns or connections within a dataset, a bit like the human brain does while deciding.

2.5.3 Fuzzy systems. The term fuzzy refers to things that aren't clear or are vague. We repeatedly encounter a situation within the world once we can't determine whether the state is true or false; their symbolic logic provides precious flexibility for reasoning. In this way, we will consider the inaccuracies and uncertainties of any situation.

2.5.4 In a Boolean system, the truth value is represented by 1.0 and false by 0.0. But within the fuzzy system, there's no logic for the absolute truth and absolute false value. But in symbolic logic, there's an intermediate value too present, which is partially true and partially wrong.

2.5.5 Ontologies. Ontologies having a specific domain and their relation are generic formal specification of the terms (words/concepts/objects. The event of Ontologies has been moving from the realm of artificial-intelligence laboratories to the desktops of domain experts in recent years. On the online and in the semantic web, Ontologies have become common nowadays. The online range from large taxonomies categorizes internet sites (such as on Google) to categorizations of products purchasable and their features (such as on Amazon). several explanations for creating ontology are:

• Sharing a common understanding of the structure of data among the stake-holder

• Enabling reuse of domain knowledge

• Making domain specific rules, generic and analysis

I have several techniques with their pros and cons as their target and domain regarding the scientific document or reports. The foremost common approaches developed over time within the literature are enlisted in Table 1.

TABLE-1. Information extraction work

Paper Title	Techniques	Author Name	Main Idea
A Hybrid Approach for Scholarly Information Extraction	Hybrid Technique of Clustering and Classification	Bod'o and Csat'o (2017) [5]	To extract the meta-data of the research paper using dictionary & font-based information
Using Text Mining Techniques for Extracting Information from Research Articles.	Text Mining Processing Framework	Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K.(2018)[8]	Text Mining Processing Framework for synergy between information extraction and data mining techniques helps to discover different interesting text patterns in the retrieved articles
Information Extraction from Semi and Unstructured Data Sources: A Systematic Literature Review	chronologically reviewing techniques	Gohar Zaman, Hairulnizam Mahdin, Khalid Hussain and Attaur-Rahman (2020)[9]	upon chronologically reviewing the work done in information ex- traction from semi and unstructured scientific documents
Big Data: Survey, Technologies, Opportunities, and Challenges	A hybrid approach for metadata extraction	Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam,, Muhammad Shiraz, and Abdullah Gani[2014][10]	It includes increase in data, the progressive demand for HDDs, and the role of Big Data in the current environment of enterprise and technology. To enhance the efficiency of data management, they devised a data-life cycle that uses the technologies and terminologies of Big Data.
Using Text Mining Techniques for Extracting Information from Research Articles	Clustering, Word Cloud, ASRM, Visualization, Similarity	Salloum et al. (2018) [1]	Extract the exciting topics from 300 journals of 6 different central database

	Measure, Term Frequency		
Framework for Automatic Information Extraction from Research Papers on Nanocrystal Devices	NaDevEx AIE System	Dieb et al. (2015) [7]	Making rules using CRF techniques and finding patterns to extract information
Logical Structure Recovery in Scholarly Articles with Rich Document Features	CRF Technique	Luong et al. (2012) [2]	Extracting the information by identifying the font size of the research paper using OCR
Parsing Citations in Biomedical Articles Using Conditional Random Fields	Conditional Ran- dom Field	Zhang et al. (2011) [6]	Extract the citation of the research paper using the conditional random field.
Extracting and Matching Authors and Affiliations in Scholarly Documents	Enlil (name of the technique) Information Extraction System using CRF and use of SVM	Do et al. (2013) [3]	First extract author name & Affiliations using CRF and connect them using SVM
Semi-automatic Meta- data Extraction from Scientific Journal Article for Full-text XML Conversion	Rule Base Method & CRF	Kim et al. (2014) [4]	Making rules using CRF techniques and finding patterns to ex- tract information

4. **Conclusion.** One of the most liked areas of the researcher is Information extraction in data and text mining for digital libraries. This system is especially wont to extract structural and other important information from semi-structured and unstructured documents, mainly online. This paper is devoted to over viewing the state-of-the-art techniques in practice by the existing system and giving some idea about their limitations. The target is to find the areas in information extraction that require improvement; to assist the researchers and
5. students of the field. In the future, we are getting to propose our technique for improved information extraction for digital libraries and knowledge 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] 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.

[3] H. H. N. Do, M. K. Chandrasekaran, P. S. Cho and M. Y. Kan
I am 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.

[4] 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.

[5] Z. Bodo and L. Csat'o, A hybrid approach for scholarly information extraction, Stud. Univ. Babes- Bolyai, Inform. vol.62, no.2, 2017.

[6] 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.

[7] 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.

[8] Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K.
Information_extraction_from_semi_and_unstructured_data_sources_a_systematic_literature_review from https://www.researchgate.net/publication/341030837

[9] Gohar Zaman, Hairulnizam Mahdin, Khalid Hussain and Attaur-RahmanUsing Text Mining Techniques for Extracting from research paper of
https://link.springer.com/chapter/10.1007/978-3-319-67056-0_18

[10] Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam,, Muhammad Shiraz, and Abdullah Gani, Big Data: Survey, Technologies, Opportunities, and Challenges from https://www.hindawi.com/journals/tswj/2014/712826/abs/