A Survey on Information Retrieval Using Various Techniques

[1]Florence Vijila . S, [2]Nirmala . K
[1]Research Scholar, [2]Associate Professor
[1]Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India,
[1]Department of Computer Science, Quaid-E- Millath Govt. College for Women, Chennai, India

Structured data, typically, is predefined data. Semi-structured and unstructured data are not predefined data that includes documents, emails, social media posts, images, videos, etc. Text extraction is a critical stage of analyzing Journal papers. Journal papers generally are in PDF format which is semi structured data. Journal papers are presented into different sections like Introduction, Methodology, Experimental, Result, Conclusion etc. It makes easy to analyze based on readers interested topic. The main importance on section extraction is to find a representative subset of the data, which contains the information of the entire set. To extract research papers, we can approach machine learning, NLP, etc. In this paper we present review of various extraction techniques from a PDF document. Data consolidation is used to combine the extracted data to obtain structured data from papers. This will make the knowledge extraction process easy to manage and analyze.

***Index Terms*— Information extraction, Text Mining, NLP, Machine Learning Methods**

I. INTRODUCTION

Data is becoming the focus in today's business scenarios due to the dramatic growth of internet and online storages. This data can be structured, semi-structured and unstructured. Structured data is organized and can be easily analyzed. Semi structured and unstructured data are not predefined and complex that makes gaining knowledge from them a great challenge. Most organizations prefer to work with structured data because it has the advantage of being easily stored, managed, and analyzed. However, most of today's decision making data comes from external sources that are unstructured which makes up more than 90% of today's data [1]. Text mining and natural language processing (NLP) can be applied to texts to structure and extract meaningful information from it.

Data required for the flow of information such that, it provides a platform for expressing the views and specific perspectives of an individual. It is the essential unit of communication over the generations which has facilitated in the need for preserving and utilizing the data. Over the years, there has been a rapid escalation in the information gradient supported by convenient methods of data storage and representation based on the underlying application. The main challenge encountered by a user is to capture data of his interest. Retrieving a part of the data or extracting a relevant section from a given document is an active area of research and lead to the introduction of search techniques based on Key terms and pattern matching.

The research document which are usually stored as textual documents are in a Portable Document Format (PDF) consists of huge amount of significant data, which is the primary source of information reserves for the research scholars. These documents provide a base on which existing information is extracted, analyzed, retrieved in case of manipulations if necessary, and stored as well for further reference. In order to increase the readability of the document, it is being categorized into different sections where a very brief insight into the document is provided in terms of abstract and introduction. This is further followed by several other sections like review, methodology applied, results and conclusion.

Key terms are the exact data items which are explicitly defined in terms of other data units. The use of key terms has enabled the retrieval of required data in much better manner. More complex applications resulted in more number of key terms, finally resulting in the usage of key terms with associated index terms called as values. The key - value pairs increased the effectiveness of data retrieval in a proficient manner. The selection of key terms plays a very significant role as the selection process is based on the data semantics and syntax which varies over textual, audio and video data representation. This enables in highlighting the characteristic features of the data document and providing the fundamentals towards Information Extraction or Information Retrieval.

A list of key terms associated with every document enable the user in streamlining the documents based on their applicability. The noble example of information extraction based on key terms is the Google Search Engine which works on the principles of PageRank algorithm using a logarithmic scale [2]. The main advantages of PageRank are less query time, less susceptibility to localized links, more efficiency. On the other hand, disadvantages of PageRank are may retrieve documents less relevant to user query, rank sinks, it is a static algorithm, dangling links [3].

The real time applications of today not only need information retrieval to happen in a very efficient manner but always contemplate more on the gathering data on qualitative basis than on quantitative levels. As the data size has increased tremendously, manual information retrieval has become impossible. Hence it provides concepts to model this data so that they can return all the documents that are relevant to the query term and ranked based on certain importance measures. This paper provides the discussion of methods for information extraction in general and then focus on the review of section extraction

methods in particular research documents. The proposed survey provides a detailed insight into the existing information extraction, and their shortcomings.

II. GENERAL EXTRACTION METHODS

A much broader categorization of the documents can be given in terms of free-text, semi-structured text and structured text [4][5]. A free-text is a loose collection of scripts and stories which are not well formatted and have very minimal connectivity established. Extracting information from such documents is a very tedious job as each unit of data that is present in the document must be analyzed to understand its existence. The data that is maintained as a structured text is stored in an organized manner and retrieved by their key values. The semi-structured text on the other hand follows a domain level categorization based on automated method.

Information Extraction (IE) is the process of extracting useful data from the already existing data by employing the statistical techniques of Natural Language Processing (NLP) [6]. It is defined as the act of identifying, collecting and regularizing relevant information from the given text and producing the same in a suitable output structure [7]. Although the extraction process has been automated over years, the need for training the system to work as per the rapid changes within the specified time range is very much important.

The process of extraction can be decomposed to make its functionality simpler by following a modular approach for each of its task of selection, reordering and composition. Each of the sub-process thus identified can in turn adopt a flexible technique based on the underlying application and this is advantageous in case of modular approach. It also facilitates in improving the validity of the system, relationship and level of coupling between the components

The major modules identified under Information Extraction are: segmentation, classification, association, normalization and co-occurrence resolution. For the text given as the input, segmentation focuses on dividing, based on the semantics and syntax of the data format. Lexical analyzers are commonly used to define the semantic rules to enable an effective data segmentation process. The basic data units are identified by segmentation and of all the identified data units, specific data is given primary importance based on the priority, number of times it is being repeated and their usage according to the semantics of the language defined.

The most commonly used segmentation technique is by the Viterbi Algorithm [8] which works on the concept of state machines. Machine learning approaches [9] are employed based on the probabilistic rules followed by the Context Free Grammars. Classification methods follow segmentation which categorize similar data units into a common group such that meaningful relationships can be constructed between the grouped entities. The normalization module is used to confirm that the extractions are according to the specified formats to enable synchronization between the forms of data obtained at each module. In case of repetition, co-occurrence [10] is used such that identity of each data entity is maintained.

Classification in extraction enables us to classify the major domains into which the segmented data is finally stored by using machine learning approaches. Common scenario used for representing the classified data is by the use of decision trees. In order to extract the related entities, association rules are used to extract the desired relations of various categories. Machine learning plays a very critical role in IE as we are able to achieve most accurate results with very less error propagation rate. In case if any improvements are required then Machine Learning with NLP provides a basis for the proofs that are obtained through empirical methods. Extraction [11] through Machine Learning if automated helps in the extraction of patterns covering different areas in very minimal time duration

Commonly used Machine Learning based Extraction methods can be categorized as supervised learning and unsupervised learning. Supervised learning is used when there is user interaction in defining the rules for the extraction process, or in defining the sample training example based on which the supervision task can be carried out. The categories of supervised learning are propositional learning and relational learning. Using the core mathematical foundations, propositional learning is practiced where the examples are represented in terms of zero order logic or attribute-value logic. On the other hand, relational learning uses the first order logic to represent the examples for the learning process and is mostly useful in case of textual data type.

The need for human interaction to provide the training example is one of the major drawback of the supervised learning technique which has paved way for the unsupervised learning mechanisms which are developed by employing corpus bootstrapping method [12] to obtain the seed rules from annotated systems. Yet another category is the semi supervised learning technique based on mutual bootstrapping and finally Hybrid NER [13] (Name Entity Recognition) system which works by combining the randomized conditional variable and a supporting base variable. However, these techniques are just not enough to solve the current tribulations faced during extraction as data identification, classification and management as the core activities which have to be given due importance [14].

III. RELATED WORK

An integration of semantic technology (ST), NLP and Information extraction (IE) are used in [15] to provide a new method for knowledge extraction from research documents. After pre-processing of the data, keywords are extracted from the documents using regular expression. They used two triple-store on sentences and words based on three type formats: Subject, Predicate and Object to extract useful information from the documents. Then, inference rules are applied on triple-store data to extract knowledge from the processed data. This thesis [16] proposed a new methodology for knowledge discovery from large unstructured text data.

An ‘Ontology-based Knowledge Discovery in Text (On-KDT)’ is presented to exploit the encrypted semantic information in ontologies to improve the process for knowledge extraction. This methodology was applied in three different areas: software requirements to extract the outline from the text files, PubMed abstracts to derive important medical information and business warehouses to extract business rules.

An effective methodology was presented in [17] to extract structured data from a large corpus of unstructured business documents based on two major steps. In the first step, they searched for the similar documents in the corpus, and they made families of similar documents. The second step involves the statistical quantization of object instances from these families. Then, they measured the attributes of a specific object quantitatively to extract structured knowledge.

The basic idea of the traditional machine learning method is to transform the relation extraction problem into a binary classification problem, and then train the supervised system with labeled text [18]. Further, the supervised methods of relation extraction include features based method and kernel method. Feature based method is needed to manually set the feature-matrix of text, while kernel method does not manually label. Feature based method may have better results in specific areas.

The process of dealing with the problem of relation extraction with machine learning method contains the following steps:

1) Find sentences that contain relations and entities in the texts. Giving a sentence S = group1(ws1 e1 ws2 e2 ws3, r1), group2(ws1 e1 ws2 e2 ws3, r2), . . ., where e1 and e2 are entities, ws1, ws2 and ws3 are the word sequences around the entities, and r1 and r2 are the relations between e1 and e2.

2) Transform relation groups into corresponding feature matrix, that is, F(group).

3) Set a hypothesis h(F(group)), if the value of h () is positive then it means that the group implies relation r, otherwise does not imply relation r.

In this way, the task of extracting relation is turned into the task of identifying a specific relation in a sentence. With a number of text with positive and negative labels as the training data, the supervised relation extraction system can be constructed by using SVM, Logistic or other binary classifiers

IV. EXTRACTION PROCESS OF PARTICULAR DOCUMENT

This section highlights how extraction process for scientific literature is different from general information extraction form PDF document. Paul Buitelaar et.al [15] in their work on Topic Extraction from scientific literature have suggested automated system that is dynamic in nature to support critical systems by extracting data from scientific applications. A pattern based approach has been employed to organize the data according to the research requirements. The extraction process depends on the pertinent skills and the semantic relations between the existing phonetics and finally statistical approaches are employed for information retrieval along with learning scheme sets defined by machine learning aspects. To bring in the support for cross domains, domain-specific linguistic patterns are used. The representation of the extracted units is done here by the use of an association network which signifies the interconnections between the units there by highlighting the area of proficiency and capability between the researchers.

The work Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers, proposed by Sonal Gupta in [19] characterizes the given research document into key areas like focus, domain of application and different underlying techniques used. The data extracts are obtained using the semantic extraction patterns which uses bootstrapping as the learning technique and finally representing them in the tree like structure. The different domains identified here were termed as communities and this became a novel approach as it influenced in extracting the characteristics at the semantic levels by employing bootstrapping learning technique and finally representing the output in the form of dependency tree. This work facilitated in the dynamic extraction of rich information which is worth the search as it categorizes into respective communities making the search more effective.

The task of identifying the relevant topic of importance is very difficult even in case of providing the keywords. Streamlining the major key terms is itself a tedious job and upon that moving along the related data in large texts is quite a challenge. In case if we consider multi-document data set, an automated key phrase extraction is more useful as it helps in obtaining a precise subset of categories generally termed as clusters. The algorithm for text extraction used in this case could provide exact clusters which were much better compared to the previously used key phrase extraction techniques irrespective of working for various domains there by providing domain independency.

The cluster-to-cluster similarity index could also be calculated using the proposed algorithm. But the proposed algorithm did not employ any ranking schemes which would have enhanced the removal of unwanted permuted combinations of phrases [20]. The keywords are decided based on their repeated occurrence and frequency which is termed as co-occurrence distribution factor. By using the proposed algorithm, key term [21] extraction is done very accurately even if the occurrence is rare but the term is important. The advantages of this method enabled text mining without the need for corpus and its effective usage in case of domain-independent keyword extraction.

Slobodan Beliga in [22] has provided a detailed review on keyword extraction and the different approaches used for keyword extraction. The importance of both supervised and unsupervised learning methods is understood along with the features of Croatian keyword extraction. Further categories of extraction include document-oriented and collection-oriented techniques through which more systematic extraction could be achieved. In case of applications with complex networks, Selectivity-based keyword extraction has been proposed as the new unsupervised keyword extraction technique which paves way for further research applications [23]. The advantages of Selective based keyword extraction is that it does not work on any linguistic knowledge of the text document, but the proofs are obtained by the employment of statistical methods because on which manual annotations required is minimal and beneficial in case of fast computing.

There is a new method in Bo Chen [24] nonlinear model using SVM with max-margin discriminate projection is used to fine the variable and augmented variables to extract information from data set. This model is developed for supervised data set. This model can be extended to unsupervised way to cluster the data. When we are working with data of a variety of different domains, it is very much necessary to have minimal domain specific knowledge for the automatic extraction of the key terms. If at all new domains are to be included into the system, then the earlier systems were not flexible enough for updating due to strenuous manual tuning of the domain-dependent semantics and syntax. Moreover, even though we have numerous extraction techniques, arriving at the most accurate model is purely based on the design decisions made at the early steps.

V. EXTRACTION METHODS

Extraction methods can be categorized into different classes depending on input data, analyzing techniques, information required [25]. These data can be differentiated into structured, semi-structured and unstructured data. Analyzing techniques of data analysis can be NOSQL Databases, BigQuery, MapReduce, WibiData, Skytree and Hadoop. The required information as the output can be grouped into Descriptive analytic (analytics that help in understanding what has happened), predictive analytics (analytics that help in anticipating what will happen) and prescriptive analytics (analytics that help to respond what next). We are analyzing and reviewing the existing techniques so that a new method can be developed to overcome the disadvantages.

VI. CONCLUSION

This paper discusses a few extraction methods to extract sections from documents. NLP and probabilistic extraction techniques are purely based on the design decisions made at the early steps. However, machine learning methods still have many shortcomings and limitations, such as the most of relation extraction systems The complexity also occurs in terms of the overall time period taken in the extraction process. Identifying the right key phrase and finalizing a right subset is often time consuming though automation is followed in the recent days. Analyzing and extracting data of different semantics have resulted in performance issues as well

REFERENCES

[1] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics", International Journal of Information Management, vol. 35, no. 2, pp. 137-144, 2015.

[2] Amy Langville & Carl Meyer, "Google's PageRank and Beyond". The Science of Search Engine Rankings Princeton University Press, 2006.

[3] Pooja Devi, Ashlesha Gupta, Ashutosh Dixit, "Comparative Study of HITS and PageRank Link based Ranking Algorithms", IJARCCE, Vol 3, Issue 2, February 2014.

[4] S. Soderland, "Learning information extraction rules for semi-structured and free text," Machine learning, vol. 34, p 233-272, 1999.

[5] McCallum, A. (2005). "Information extraction: Distilling structured data from unstructured text", ACM Queue (Vol. 3, pp. 48{57). New York, NY, USA, 2005.

[6] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information", IBM Journal of research and development, vol. 1, pp. 309- 317, 1957.

[7] P. Cimiano, U. Reyle, and J. Šarić, "Ontology-driven discourse analysis for information extraction", Data & Knowledge Engineering, vol. 55, pp. 59- 83, 2005.

[8] Jerry R. Van Aken, "A Statistical Learning Algorithm for Word Segmentation", Miscrosoft Corporation, Redmond, WA 98052.

[9] Neil Ireson, Fabio Ciravegna, "Evaluating Machine Learning for Information Extraction", 22nd International Conference on Machine Learning, Bonn, Germany, 2005.

[10] Yutaka Matsuo, Mitsuru Ishizuka, "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information", AAAI 2003.

[11] Goncalo Simoes, Helena Galhardas, Luisa Coheur, "Information Extraction tasks: a survey", 2004.

[12] Sonal Gupta, Christopher D Manning, "Analyzing the Dynamics of Research by Extracting Key aspects of Scientific papers", Stanford University.

[13] Kanwalpreet Singh Bajwa, Amardeep Kaur, "Hybrid Approach for Named Entity Recognition", International Journal of Computer Application, Vol 118-No1, May 2015.

[14] Muawia Abdelmagid, Ali Ahmed and Mubarak Himmat, "Information Extraction Methods and Extraction Techniques in the Chemical Document's Contents: Survey", ARPN Journal of Engineering and Applied Sciences, 2015.

[15] R. Upadhyay and A. Fujii, "Semantic Knowledge Extraction from Research Documents", Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, vol. 8, pp. 439-445, IEEE, 2016.

[16] Polpinij, "Ontology-based knowledge discovery from unstructured and semi-structured text," University of Wollongong thesis Collection, 2014.

[17] G. Pandey and R. Daga, "On Extracting Structured Knowledge from Unstructured Business Documents" In: Proc IJCAI Workshop on Analytics for Noisy Unstructured Text Data, pp 155-162, 2007.

[18] C. W. Xiang, T. Liu, and L. I. Sheng, "Automatic entity relation extraction," *Journal of Chinese Information Processing*, 2005.

[19] Sonal Gupta, Christopher D. Manning, "Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers", Department of Computer Science Stanford University, 2010.

[20] Khaled M. Hammouda, Diego N. Matute, and Mohamed S. Kamel, "CorePhrase: Keyphrase Extraction for Document Clustering", Pattern Analysis and Machine Intelligence (PAMI) Research Group University of Waterloo, 2005.

[21] Kavitha Jayaram, Sangeetha K, "A Review: Information Extraction Techiniques from Research papers", International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2017), 978-1-5090-5960-7/17.

[22] Slobodan Beliga, "Keyword extraction: a review of methods and approaches", University of Rijeka, Department of Informatics, 2014.

[23] J. Saratlija, J. Šnajder, B. Dalbelo-Bšić, "Unsupervised topic-oriented keyphrase extraction and its application to Croatian", T ext, Speech and Dialogue, pp. 340-347, 2011.

[24] Bo Chen, Hao Zhang, Xuefeng Zhang, Wei Wen, Hongwei Liu and Jun Liu, "Max-Margin Discriminant Projection via Data Augmentation", IEEE Transactions on Knowledge and Data Engineering, Vol 27, N0 7, July 2015.
[25] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody, "Critical analysis of Big Data challenges and analytical methods. Journal of Business Research 70 (2017) 263-286.