International Institution for Technological Research and Development
Volume 1, Issue 1, 2015

Effective Information and Metadata Extraction from Web

Nidhi Goyal[1] and Dr. Shyamal Tanna[2]

[1] Information & Technology Department, LJIET
Ahmedabad, Gujarat, India
goyalnidhi2710@gmail.com

[2] Information & Technology Department, LJIET
Ahmedabad, Gujarat, India
drsmtanna@gmail.com

Abstract

Web is a great source of information today. A lot of information is available over the internet and a lot of information is added and updated to it every day hence web data extraction systems are necessary to use. While Internet takes up by far the most significant part of our daily lives, finding jobs/employees on the internet has started to play a crucial role for job seekers and employees. Online recruitment websites and human resources consultancy and recruitment companies enable job seekers to create their resume in order to find and apply for the desirable jobs, whereas they enable for the companies to find the qualified employees they are looking for. But the resumes are written in many ways that make difficult for the online recruitment companies to keep these data in their relational databases. So, in this mentioned project, a system enables free structured format of resumes to transform into an ontological structure model. The proposed system will be kept in Semantic web approach that provides companies to find expert finding in an efficient way.

Keywords: *Ontology, Semantic Web, Information Extraction, Resume and curriculum Vitae.*

1. Introduction

The web has become the major source of information, bearing the potential of being the world's largest encyclopedic source of all the news, data, etc. It brings up the interesting idea of converting this sheer volume of unstructured textual data into useful information available for everyone. But the accurate information extraction from web pages is an intensive and time consuming task which requires important background knowledge. Thus the development of efficient and robust information extraction is a big challenge.

Information extraction broadly refers to extracting knowledge from unstructured text sources. The main goal behind it is to allow semantic tagging of the text source as well allowing the possibility of machine reading of the text source.
Semantic web is an extension of World Wide Web that aims to enable computers to discover, search, infer and collect Web's information without human effort. Semantic web allows efficient way of representing data on the World Wide Web. Ontology is the term that refers to define and make connections between information.
Web Ontology Language (WOL) is a standard ontology language from World Wide Consortium that processes and instantiates Ontology.

Techniques used for Web Data Extraction
Web data extraction techniques uses different approaches like markov chains, graph theory, neural network approaches, statistical techniques and association mining and different extraction tools to extract the information from the web sources. Following methods are used for the web data extraction purpose:

1.1 Tree-based techniques

One of the most important features in web data extraction is the semi-structured or unstructured nature of web pages i.e. no specified structure most of the times like in databases. This type of data can be represented by labeled ordered rooted trees, where labels represent the tags of the HTML mark-up language syntax, and the tree hierarchy represents the different levels of nesting of elements constituting the web pages. The representation of a web page by using a labeled ordered rooted tree is known as DOM (Document

Object Model). The general idea behind the Document Object Model is that HTML Web pages are represented by means of plain text, which contains HTML tags, i.e., particular keywords defined in the mark-up language that can be interpreted by the browser to represent the elements specific of a Web page (e.g., hyper-links, buttons, images and so forth). HTML tags may be nested one into another, forming a hierarchical structure. This hierarchy is captured in the DOM tree, whose nodes represent HTML tags. This technique is easy and cheaper to implement than other techniques.

1.2 Web wrappers

Web wrappers are the programs or procedure, that might implement one or different classes of algorithms, which seeks and finds data required by a human user, extracting them from unstructured (or semi-structured) web sources, and transforming them into structured data, merging and unifying this information for further processing, in a semi-automatic or fully automatic way. The main limitation of web wrapper is that for every website or script we have to develop a different program which extract data according to web source hence these are costly as compared to other techniques but faster than other techniques. Web wrappers are characterized by a life-cycle which constitutes wrapper generation, wrapper execution and wrapper maintenance like in life cycle of software.

1.3 Machine learning approaches

Machine Learning techniques fit well to the purpose of extracting domain-specific information from web sources, since they rely on training sessions during which a system acquires a domain expertise. These techniques are applied on semantic web which is based on machine learning systems. These techniques are performed by automatic systems instead of manually done. Statistical Machine Learning systems are also developed, relying on conditional models or adaptive search as an alternative solution to human knowledge and interaction.

1.4 Web data mining

Web data mining is an application of data mining technique which is used for searching hidden information & patterns from the WebPages. As Web has grown exponentially along with its strengths and its weakness, the strength is that one can find out information on just about anything even if the quality varies. The weakness is that there is the problem of abundance and types of information. Standard data mining techniques may be applied for mining information on the Web, but data mining mainly deals with structured form of data organized in well formed databases while web mining deals with unstructured form of data. So mining of web data is one of the most challenging tasks for the data mining.

2. Literature Survey

Duygu Celik, Askin Karakas, Gulsen Bal, Cem Gultunca, Atilla Elci, Basak Buluz, Murat Can Alevli[1], has proposed a system that enables free structured format of resumes to transform into an ontological structure model. The proposed system is based on ontological structure model and called Ontology based Resume Parser (ORP) and tested on a number of Turkish and English resumes. This proposed system is kept in Semantic Web approach that provides companies to find expert finding in an efficient way. In these system, they applied the algorithms like Sentence End, N gram and Jaro Wrinkle algorithm to achieve the efficient results.
They have tested the proposed system with 250 Turkish resumes and acquire average of 80% information extraction form resumes. The calculation of information extraction is presented below:

$$\frac{Extracted\ Word\ Numbers}{Total\ Word\ Numbers} \times 100$$

And the system runtime is 5 second averagely. To prevent wrong information extraction, the system applies defined rules on specific parts. To illustrate, if a paragraph is an education experience paragraph then only education experience information extraction rules will be applied not work experience rules. The system defines some main parts/sections which have to be defined in a resume such as Work Experience, Education Information, General Information etc. Based on these sections, the system calculates percentage completeness of a resume to indicate how many parts are completed averagely.

Yang Xiudan and Zhu Yuanyuan[2] has used the concept of ontology to analyze the structure and content of

the website to build ontology model in order to extract the information based on ontology from the e-commerce website for the users. In this paper, they used the ontology technology to build the wrapper and then extract information from the e-commerce site.

In this paper, they propose to put the domain ontology into the e-commerce information extraction; they use the semantic extraction algorithm in the system, generate the extraction rules based on OWL ontology and construct the wrappers for the web information extraction.

Ontology composition can be explained using the following formula:

Ontology = Classes + Relations + Axiom + Instances + Functions.

First, the paper downloads a batch of e-commerce web pages (html), inputs the ontology and the web resources into GATE, and the uses the ontology to extract the information from the web resources, and finally outputs the results. GATE has the function to load the ontology and extract using ontology. The results are as follows. The paper inputs the extracted data into a database table, and queries from it when the users make a request.

Baraa Jebali and Ramzi Farhat[3] has described the approach for automatic generation of learning objects' semantic metadata. The extraction process is based on the OBIE systems' principles. The input of the approach is a set of IEEE LOM metadata elements in conformance with two requirements. First, each data element must describe the educational content of the learning object. Second, it must be one of the data elements frequently filled by the learning objects' authors and required by most of the LOM application profiles. Concerning the outputs, each one is a couple of a domain concept (from domain ontology) and a degree of pertinence. Moreover they present the details concerning the integration of their approach to learning objects' repositories by taking the COLORS repository as example. In fact the ultimate goal behind their approach is the improvement of repositories' services by offering semantic metadata.

3. OBIE System

The figure present combinations of Ontology with IE system. It distinguishes between Knowledge base and ontology. Information Extractors populates the Knowledge base and is queried by users. The beauty of this architecture is that Human intervention as a Domain expert is allowed to manipulate the internal extraction logic. The information extractors in this work utilize RDF graphs from Semantic Web sources, extract information from text, and return extracted information as RDF graphs. SPARQL queries provide means for additional filtering mechanisms. The role of semantic lexicons is fulfilled by applying linguistic resources such as text segmenters, POS taggers, and text chunkers.

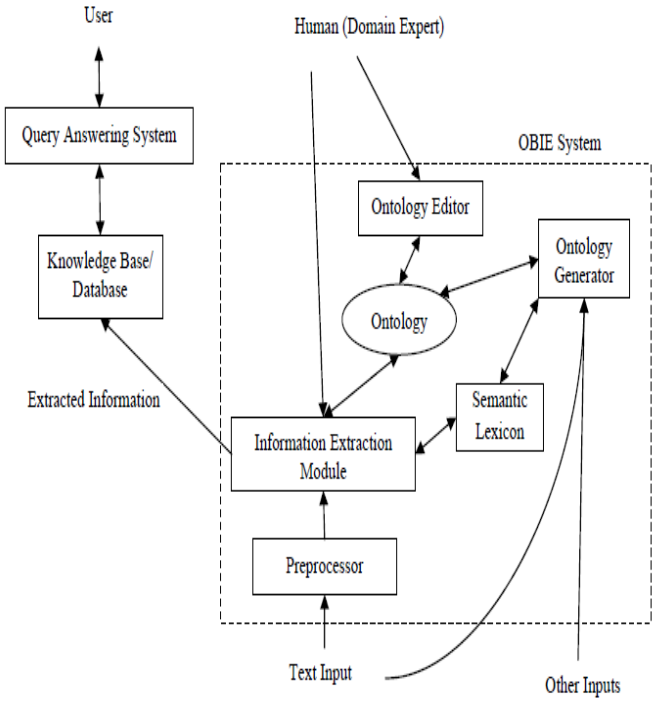

Figure 1: General Architecture of an OBIE system

4. Conclusion

In this paper, we have reviewed the new field of ontology-based information extraction and a number of systems that are categorized under it. We review the some of the Ontology Based Information Extraction method. Ontology Based Information Extraction method only guides the system that how to pull out efficient and relevant information using the Information Extraction methods. There are several directions for future work with OBIE System like improving the efficiency of IE process to improve the precision and recall. Generating the semantic contents for the Semantic Web is one of the major factors that make OBIE an interesting research field. OBIE system can be used for identified the semantic content for

semantic web and also implemented Ontology Based web service for result. Most of the OBIE system uses single Domain specific ontology. However, there is no rule to not to use multiple Ontology. Among other things, we have provided a definition for the field, identified a common architecture for OBIE systems and classified the existing systems along different dimensions. We believe that these will be useful for future research work in this area.

References

[1] Duygu Celik, Askin Karakas, Gulsen Bal, Cem Gultunca, Atilla Elci, Basak Buluz and Murat Can Alevli, "Towards an Information Extraction System Based on Ontology to Match Resumes and Jobs", IEEE 37th Annual Computer Software and Applications Conference Workshops, 978-0-7695-4987-3/13 ,2013

[2] Yang Xiudan and Zhu Yuanyuan, " Ontology-based information extraction system in E-commerce websites", IEEE, 978-1-4577-0860-2/11/,2011.

[3] Baraa Jebali and Ramzi Farhat, "Ontology-based semantic metadata extraction approach",IEEE, 2013.

[4] Ritesh Shah and Suresh Jain, "Ontology-based Information Extraction: An overview and a study of different approaches", International Journal of Computer Applications, Volume-87-No.4, February 2014.

[5] Ashraf Uddin, Rajesh Priyani and Vivek Kumar Singh, "Information and Relation Extraction for Semantic Annotation of eBook Texts", Springer International Publishing Switzerland, DOI: 10.1007/978-3-319-01778-5_22, 2014.

[6] Rinaldo Lima, Hilario Oliveira, Fred Freitas, Bernard Espinasse, Laura Pentagrossa, "Information Extraction from the Web: An Ontology-Based Method using Inductive Logic Programming", IEEE 25th International Conference on Tools with Artificial Intelligence, 1082-3409/13, 2013.

[7] Neeraj Raheja and Dr. V.K. Katiyar, "A Survey on Data Extraction in Web Based Environment", International Journal of Software and Web Sciences, ISSN: 2279-0063, 2013.

[8] S.C. Gowri, Dr. K. Meenakshi Sundaram, "A Study on Information Retrieval and Extraction for Text Data Words using Data Mining Classifier", International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 10, 2015.