Journal of Engineering and Applied Sciences 13 (Special Issue 3): 3284-3288, 2018
ISSN: 1816-949X
© Medwell Journals, 2018

Data Extraction in Semantic Web: Literature Review

Rishabh Bhandari, Vikas Deep and Naveen Garg
Amity University, Nodia, Uttar Pradesh (UP), India

Abstract: An enormous amount of data exists on the internet; there are millions of web pages online which contain some sort of content. WWW is the biggest database on the planet and every hour peta bytes of data is being uploaded stored in databases of various websites. Now, this data is not stored in a defined way, it is just stacked up. The data online is semantically unstructured and that is why sometimes it is difficult to extract a particular data from such a giant stack. Most of the content on the web is in the form of HTML coding. There are various ways in which data can be extracted by keeping HTML coding as the base. This study discusses some of these methods and gives a literature review over them.

Key words: Information retrieval, semantic web, ontology, content, coding, data

INTRODUCTION

Semantic web is an addition of the web through W3C in which the content in the pages of the web is structured in a way that they are available to the users in a defined format. The W3C states that, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise and community boundaries" (WWWC W3C., 2011). Information extraction is the process of taking out well defined data from the unstructured content. Usually this process is performed by taking human understandable linguistic documents and then sorting them out by the process of NLP (Freitag, 2000). At present stage, the information extraction techniques that we use are not so efficient that they can make full use of semantic proficiency of the document and give specific results. "Hypertext Mark-up Language (HTML) is the language for creating web pages and applications". HTML tags are used to add pictures, videos, etc., to a web page (Flanagan, 2011). HTML provides the way to create structured documents by defining the whole content in form of tags such as heading, paragraphs, etc. This way of defining the content in a structured way is called Semantic HTML (Berners-Lee, 2000). In this study our objectives are to study the different methods for information extraction of the structured data in semantic web.

Literature review: Structuring HTML documents Saikat Mukherjee, Guizhen Yang, Wenfang Tan and I.V. Ramakrishnan knew all the HTML documents that work through a template have the same schema with their relationships in a hierarchical order. So, they wanted to find elements in the HTML document that have some relations with each other. Through this they worked on generating and algorithm that would separate the items in tree structure according to their semantics and this would show the same schema.

They took in web pages of news websites. They stated that most of the items on the page are of no use but these websites have a fixed schema. The method they proposed has 2 algorithms for structuring the information semantically (Fig. 1).

First is partition algorithm, now in this algorithm there are two components: Building blocks and maximal repeating substrings. In this method the main aim is to find syntactic similarity between the elements. Every node is given a type in the DOM tree structure. The type of nodes tells about the path that is subtree rooted from that node. Two kinds of types are given primitive type to those who are in sequence and compound type, it is given to the whole sequence. This is called building blocks. In maximal repeating substrings when a substring s of string A is repeated m times it becomes MPS under some specific conditions. The conditions state firstly the substring should have most of the elements and secondly, its radius should be large with its own length small. Partition algorithm is called to convert the HTML document's DOM tree into an organised structure. The algorithm restructures the DOM tree bottom-up after going through it top down. All the type information about the children nodes of an internal node is collected in this step.

The second algorithm is find partition. If an internal node has many children nodes then this algorithm is called upon that node to do pattern discovery on the children nodes. This algorithm selects an internal node. Now, the main function of this algorithm is to find the structural equivalence between all the children nodes and

Corresponding Author: Rishabh Bhandari, Amity University, Nodia, Uttar Pradesh (UP), India

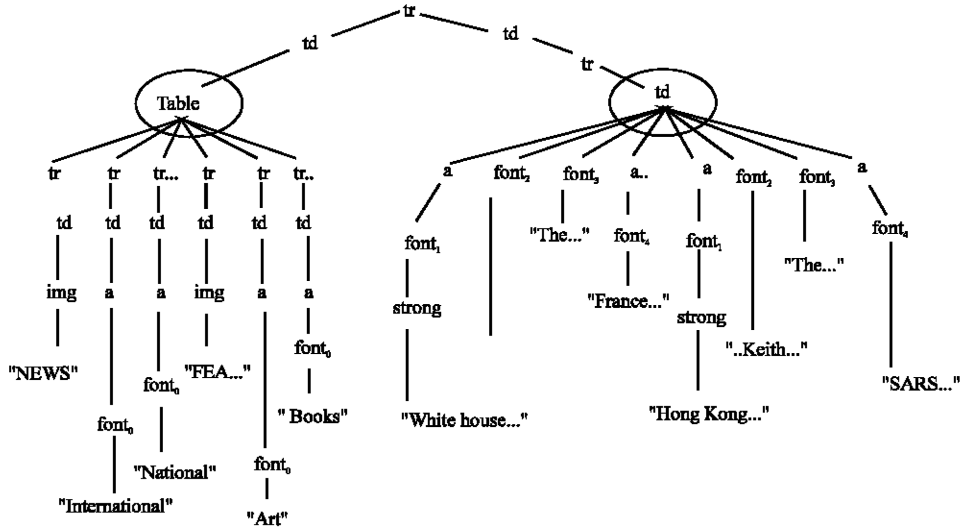

Fig. 1: Dom tree fragment of New York Times front page (Mukherjee *et al*., 2003)

restructure the tree. They for experiment selected 3 different websites. The page layout of these websites was totally different yet there method was able to collate all the information of a specific topic from these 3 websites into one (Mukherjee *et al*., 2003).

MATERIALS AND METHODS

Using clustering and distribution of nouns: Rezaei *et al*. (2015) knew that analysis of a web document that is in text is much more complex than analysis of a normal one which is in HTML because there is much more information like, HTML codes. They research on an unsupervised way to retrieve keywords from a web documents. The process takes unigram nouns by implying POS tagging on the content, then groups the nouns according to their semantic equivalence. There are six steps in the process:

Pre-processing: First the HTML source of web page is downloaded and parsed as DOM tree. All the JS and CSS codes are rejected as they are used for the purpose of styling. XPath is used to extract text nodes from the XML document. If the length of the text ode is <6 g followed by a text node of a same length or less, then the text of the preceding node is deleted.

Parts of speech: When keywords are selected by people on their own then they mostly select nouns or phrases of nouns. Then unigrams nouns are extracted by applying POS tagging on them.

Lemmatization: This main objective of this process is to bring the text into its basic form called lemma. All the candidate nouns are lemmatized using stanford lemmatization.

Similarity measure: Semantic equivalence within all the pairs is measured using Wu and Pulmer measure and if the lemma is not found in WordNet then all the nouns are deleted from the list.

Clustering: The nouns are grouped by applying agglomerative algorithm only when the sameness within the lemmas is greater than or equal to a threshold.

Selecting keywords: Nouns are ranked according to their frequency in the content.

Page segmentation using type analysis: Jain and Prasad (2014a, b) explained a way to parse web documents in form of blocks. These blocks contain some semantic information. This method of undertaking the problem is automatic and rooted on a more specific typing technique which tightly couples type analysis with vital signals to create tree structure out of blocks with the objective to get higher rate of coherence in both semantic and visual views. The basic idea behind the method leads from the following points.

- HTML tags of objects with homogeneous semantics are same
- Similar semantic blocks usually contain items with similar HTML tag sequences
- The location in the tree structure and the parent is also same with blocks of similar semantics

Firstly, a DOM tree is created of a HTML document. It is then parsed into semantic structural tree via. a 2 step process in which firstly the original DOM structure is traced through various processes namely, type analysis, filtering, generating nodes and pattern discovery. After it is done the outcome again goes through a top down specification process.

The HTML DOM tree might look ordered in presentation but it still is disordered when looked through semantic way. In the above bottom up analysis the nodes were assigned a precise type using type analysis. This is done by Type Recognition. Now, to perform type recognition there are few rules that have to be kept in mind. There are basically 7 rules that should be performed in priority. These rules assigned the type to the nodes according to their tag, height, font, text, width, link information and visibility.

Pattern discovery is finding out some sequential patterns of the child nodes which come under a single internal node. Pattern discovery works parallel with type recognition. This process is performed on the nodes that have many child nodes. Pattern discovery has its own algorithm and also some more rules that should be kept in mind to enhance the performance. The effectiveness of pattern discovery depends on capability of sorting maximal repeating continuous substrings. The worst case complexity of the algorithm is "O(n^2)", where n is the length of original string. After all these process a rough tree is generated which is sound in semantic structure and each node is a semantic block. Lastly one more specification is done just to get the structure similar to the presentation style.

About 24 HTML documents were experimented under the method. These web documents were from various different websites. It was observed that the algorithm is able to attain fulfilment even when VIPS (Vision based segmentation algorithm) fails to do that.

Analysis of XML schema using grid computing: Kim *et al*. (2009) stated that a lot of e-Businesses are using XML schemas nowadays. This schema mapping plays an important role in getting together these e-Business applications. As the XML schemas use high end and more compound equivalence measures, the time required to go through and perform them is also more. So, a more revolutionary similarity algorithm is required to manage the complexity. They based their work on grid computing and tried creating a SOA for schema mapping to increase the efficiency of the mapping algorithm. They compared their method with technologies likes Hadoop and Globus that are also based on grid computing. They worked on an architecture that performs parallel to Java system, it is called MPJ (Message Passing for Java). Hadoop's software requirements are Java and SSHD. Its system setup is specific, it has got SSH security. The data is managed by DFS and it is a clustering type algorithm. Globu Software requirements are Java and Ant. Its system setup is system independent. It has got WS-security, the data is managed by GridFTP and it performs grid computing. Now, MPJ requires only Java, its system setup is system independent. It has no support in security and in data management and it performs grid computing. Though a basic environment is required and all the machines and processors that perform are "independent of each other in terms of configuration and resource management".

This is a multi layered proposal. In this approach according to the semantic similarity results the data items are put into a target schema as mapping candidate for each and every item in the source schema. To put it in simple language, let's say there is a task to map items. The source schema has n elements and m elements are in the target schema. Now all the possible pairs are calculated between source and target for similarities. It is calculated by n*m and a new similarity matrix is generated. Now ranking are given to the candidates according to their similarity (Fig. 2).

The architecture consists of SOAP client, grid enhanced XML schema mapping web services (GridSM-WS) and a Universal Description Discovery and Integration (UDDI) directory service. GridSM-WS is of four parts namely, Schema-to-Schema (S2S) mapping service, a Grid Computing Manager (GCM) an Element to element (E2E) mapping service and a schema repository service. So, initially S2S writes its Web Service Description Language (WSDL) on UDDI then a SOAP client calls the schema mapping analysis. S2S generates the similarity matrix and sends jobs for all the cells in E2E. Now E2E calculates the semantic similarity finally S2S conjures the results and according to the semantic similarity returns the candidates to their source. The main aim for generating MPJ was to enhance high performance computing with the help of Java. Now, this algorithm is written in Java code and performed. The three methods (Hadoop, Globus and MPJ) were tested to get the

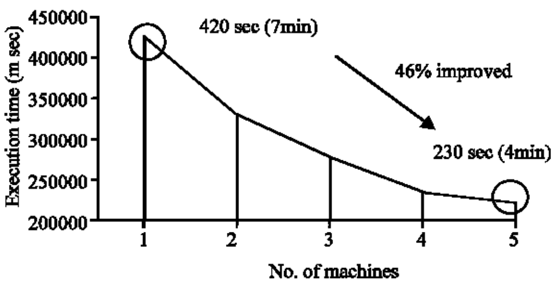

Fig. 2: The number of machines vs. execution time (Chen *et al*., 2015)

execution time. It was found that by increasing the number of processors the time to find the semantic similarity decreases. The good thing about this method is that more future research is being conducted to make it even more efficient. The makers are trying more ways and other schema mapping approaches to map the efficiency to Hadoop (Kim *et al*., 2009).

Rule based information extraction system for HSR: Chen *et al*. (2013) explained that for the last 20 years there has been a drastic increase in the data over the net which has given the rise for studying the structuring of the HTML and XML documents. Semi structured documents are everywhere even outside the web which way more difficult to process. The difference between semi structured data and HTML and XML is that the semi structured data is understandable by human not only that it also has its own internal schema. Now because of the explosion of the data the orthodox ways to processing these data is not possible because these method are not as high end and flexible as required. So, the researchers objectified two things. To bring out the semantic structure in HSR and developing a conversion system that transforms human readable scientific records into structured relational database. There were not trying to present the best method but to give and efficient frame work the semi structured HSR analysis tasks (Fig. 3).

"Human readable Scientific Records (HSR) are the documents from scientific tools and computer generated simulations". Human efforts are needed to vibrantly connect with the data in order to define the important schema for identifying the data when one is aiming to extract information from HSR. Now human efforts are needed to understand and use the semantically sound schema to understand HSR and extract information for learning and all this process requires a lot of human labour and time consumption.

In the method an essential thing is structure identification and this is done at two level 'document level structure' that consists of objects and convey the top view schema and 'region level schema' for drilling down into the internal data blocks (Table 1). In the region level schema there are various object types namely, key values pair which relate to the data value of a named attribute, quantitative data, dissociative data and descriptive text. While creating a model HSR there are few

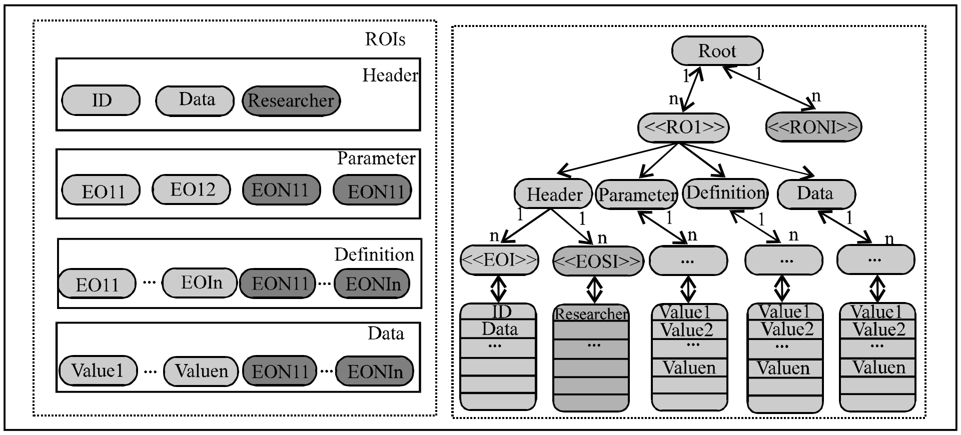

Fig. 3: The extracted HSR (Chen *et al*., 2015)

Table 1: Analysis of different techniquea

Technique	Purpose of algorithm	Experimental results
Structuring HTML documents	Automatically discover semantic structures	Proposed method can bring all similar link together (Mukherjee *et al*., 2003)
Using clustering and distribution of nouns	Extract unigram nouns and applying POS	Performs better than text rank and TF with a significant margin (Rezaei *et al*., 2015)
Page segmentation using type analysis	Parse web pages into blocks	Method can achieve thorough fullness in creating blocks
Analysis of XML schema using grid computing	Handle to complexity of XML schema	Proposed solution could become better with further research (Kim *et al*., 2009)
Rule based information extraction system for HSR	To convert HSR to modelled relational database	Present limitations hinder its total development. Future research (Chen *et al*., 2015) required

things that one should know about. The actual content of HSR is divided into multiple regions. Region of interest is a particular paragraph on which further study can be done. Entity of Interest is the small logical unit with enough semantic information for different analytical tasks. The operator V and U determines the relation of join within a region and data regions, respectively.

The strategy to retrieve information has some characteristics. To avoid the redundant whitespace reject the tokens of specific types and perform stemming algo. To retrieve ROIs use rule based approach and use filter operators to put in annotation to text data. The process is as follows:

- Extract ROI
- Extract EOIs
- Mark the EOIs into respective relational entities
- Repeat (1-3) till all are done

For the result various records from a number of fields were collected FEMR, LTR, CSR, PPR and TER. The result after the thorough research was that the errors in finding the entities are pre-dominant because EOIs are very complicated and prone to errors. But keeping in mind the shortcomings of the method the researchers are trying to expand their capability and trying out more methods to match the efficiency of the entities (Chen *et al*., 2015).

CONCLUSION

In this study, we reviewed different methods of information extraction. The methods studied are closer in achieving their defined goal. In method 1, the technique was able to group all the links together in method 2 the technique performs better than the previous techniques, method 3 can achieve fullness in creating blocks and method 4 and method 5 can become even better in performance by further research. But present information extraction techniques are not as efficient to be able to fully use the semantic information and give more accurate result. Furthermore, research is to be done to improve the efficiency of the systems.

REFERENCES

Berners-Lee, T., 2000. Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. 1st Edn., Harper Paperbacks, London, ISBN: 10: 006251587X, pp: 256.

Chen, G., B. An and S. Zeng, 2015. A rule-based information extraction system for human-readable semi-structured scientific documents. Proceedings of the 4th International Conference on Computer Science and Network Technology (ICCSNT'15) Vol. 1, December 19-20, 2015, IEEE, Harbin, China, ISBN:978-1-4673-8173-4, pp: 75-84.

Flanagan, D., 2011. JavaScript: The Definitive Guide; Activate Your Web Pages. 6th Edn., O'Reilly Media, Inc., Sebastopol, California, USA., ISBN:978-0-596-80552-4, Pages: 1079.

Freitag, D., 2000. Machine learning for information extraction in informal domains. Mach. Learn., 39: 169-202.

Jain, V. and S.V.A.V. Prasad, 2014a. Mining in ontology with multi agent system in semantic web: A novel approach. Intl. J. Multimedia Appl., 6: 45-45.

Jain, V. and S.V.A.V. Prasad, 2014b. Ontology based information retrieval model in semantic web: A review. Intl. J., 4: 837-842.

Kim, J., S. Lee, M. Halem and Y. Peng, 2009. Semantic similarity analysis of XML schema using grid computing. Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI'09), August 10-12, 2009, IEEE, Las Vegas, Nevada, ISBN:978-1-4244-4114-3, pp: 57-62.

Mukherjee, S., G. Yang, W. Tan and I.V. Ramakrishnan, 2003. Automatic discovery of semantic structures in html documents. Proceedings of the 7th International Conference on Document Analysis and Recognition, August 6, 2003, IEEE, Edinburgh, UK., ISBN:0-7695-1960-1, pp: 245-249.

Rezaei, M., N. Gali and P. Franti, 2015. ClRank: A method for keyword extraction from web pages using clustering and distribution of nouns. Proceedings of the 2015 IEEE-WIC-ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'15) Vol. 1, December 6-9, 2015, IEEE, Singapore, ISBN:978-1-4673-9618-9, pp: 79-84.

WWWC W3C., 2011. W3C semantic web activity. World Wide Web Consortium W3C, Cambridge, Massachusetts, USA.