A Survey on Knowledge Extraction Techniques for Semantic Web

Karishma
M.Tech CSE, VIT University

Abstract— Nowadays the Web has demonstrated as a wealthy, remarkable and marvellous data source of information. More than one domains can be viewed and mined in the web. Data mining in web is known as web mining. Objectives of web data mining is taken in searching relevant and reliable and meaningful knowledge from web learn about the particular user and web synthesis also. Extracting the exact information from a large volume repository of unstructured [9-10] or semi structured web is a big challenge. Still researchers are having interest in web data mining. From the various solution of this problem one is Semantic Web. Semantic web is enhancement of current web in where information is provided along with meaning also, hence it enable computers and human being to work together in better coordination. The main deal in web mining techniques is to extract the exact information from web data. This article provides an overview of various Semantic web mining techniques.

Keywords: - Web mining, Semantic Web, Ontology, RDF, Fuzzy Clustering, WUM (Web Usage Mining).

I. Introduction

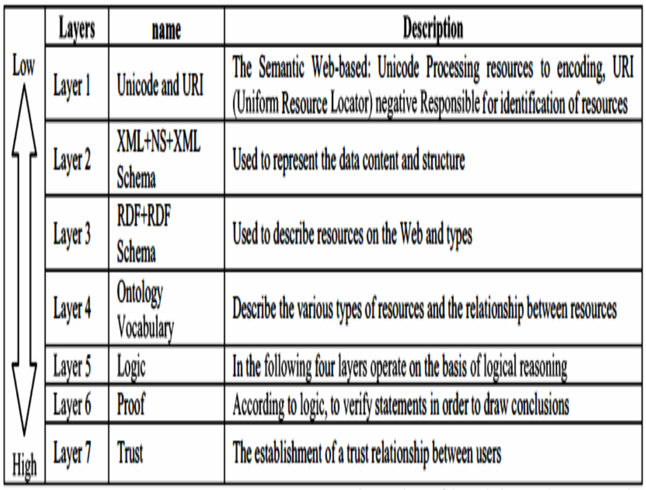
As stated by the Tim Berners-Lee about the World Wide Web (WWW) has given increase for vast volume of useful data to be gained in the digital form [1]. The growth rate of data is increasing exponentially in the web that has raised many new challenges. By the observation the data on the web described in appropriate manner and linked in the way that it can be used by automata, not only for display purpose. Generally these type of documents are not sufficiently descriptive and covered all over the web. As a consequence it is a big problem to retrieve the necessary and efficient information from the web. For this problem, from the possible solutions one is Semantic Web.

The semantic web is enhancement of the current web in which with the information meaning of data is also defined. So, both people and machines can do work together with better coordination [1]. Semantic web is cooperative and beneficial movement for application of machine – understandable environment that is led by World Wide Web Consortium (W3C). The Semantic web [2] is an advanced evolution of the World Wide Web in which meaning of stored information and services on the web is clarified. The main goal of semantic web is convert unstructured [9-10] and semi-structured document into a "web of data" and evolution of present web by allowing users to find, publish, combine information more easily.

Semantic web can efficiently applicable in health sectors, business and social networking. Standard languages suggested by W3C for semantic web are Resource Description Framework (RDF) [4, 5], RDF-Schema (RDFS) [5], Extensible Mark-up Language (XML) [3][11], Web Ontology Language (OWL) [6] to retrieve and exchange the information. A XML document contains document's content and corresponding DTD composition to meet with the effective requirements of structural exchange. With the increasing number of XML Vocabularies, the DTD expansion become complex, so XMLS is developed on the basis of W3C standards. Through these languages ontology [7] can be constructed that is used in semantic web. Each system specifies its own ontology different from the other system. So there is a possibility of presence of various ontologies in a particular domain. Various ontologies with proper interoperability in a domain mapped to each other [TABLE 2].

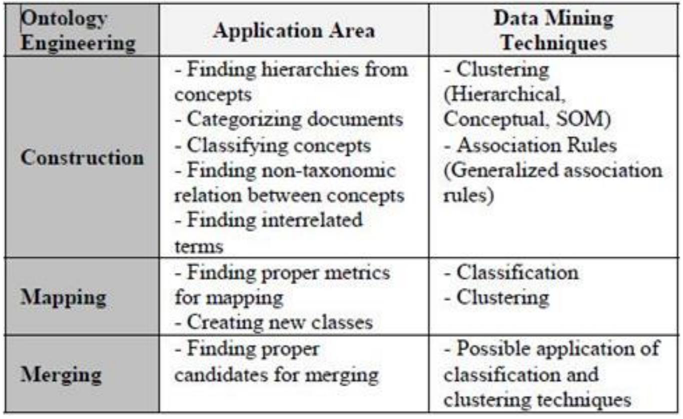
Still few challenges are present in semantic web field are Vastness, Vagueness, Uncertainty, Inconsistency, Deceit. To create record stores of data on the web, build vocabularies, apply rules for handling data, linked the data such type of functionalities are empowered by various technologies in semantic web. To extract the information from the web currently search engines are used. These search engines are working on the keyword-based search strategy, which make information extracting efficiency very low.

Table 1 Semantic Web Architecture



	Layers	name	Description
Low	Layer 1	Unicode and URI	The Semantic Web-based: Unicode Processing resources to encoding, URI (Uniform Resource Locator) negative Responsible for identification of resources
	Layer 2	XML+NS+XML Schema	Used to represent the data content and structure
	Layer 3	RDF+RDF Schema	Used to describe resources on the Web and types
	Layer 4	Ontology Vocabulary	Describe the various types of resources and the relationship between resources
	Layer 5	Logic	In the following four layers operate on the basis of logical reasoning
	Layer 6	Proof	According to logic, to verify statements in order to draw conclusions
High	Layer 7	Trust	The establishment of a trust relationship between users

Table 2 Data Mining Applications In Ontology



Ontology Engineering	Application Area	Data Mining Techniques
Construction	- Finding hierarchies from concepts - Categorizing documents - Classifying concepts - Finding non-taxonomic relation between concepts - Finding interrelated terms	- Clustering (Hierarchical, Conceptual, SOM) - Association Rules (Generalized association rules)
Mapping	- Finding proper metrics for mapping - Creating new classes	- Classification - Clustering
Merging	- Finding proper candidates for merging	- Possible application of classification and clustering techniques

Use of semantic web mining to search the data will appreciably enhance the efficiency and output of web mining. Web mining can be illustrated as [17]: "Extract requirements, patterns and information from the web resources. Normally, Web mining is classified in three categories: Web content mining, Web structure mining and

Web usages mining." Figure 1 tells about the Web mining classification.

To extract various types of data like text, image information and knowledge attribute from the web resources, Web content mining is used. For example: Which sites give you source code of project? Which sites are in English? Which pages have weather information? For network topological information extraction, web structure mining is used, that is link between the pages of site, point the page in another site, sort the page, import an important page etc. Web usage mining is allowing user to use the browser as an inter-mediator and use the linked pages. It will extract the interested information from the repository. For example, which services can client use? How much is the session time of individual page? Web mining plays a pintle role in semantic web mining to extract the information according to the user need. To extract the required information on the basis of query meaning various semantic web mining techniques exist.

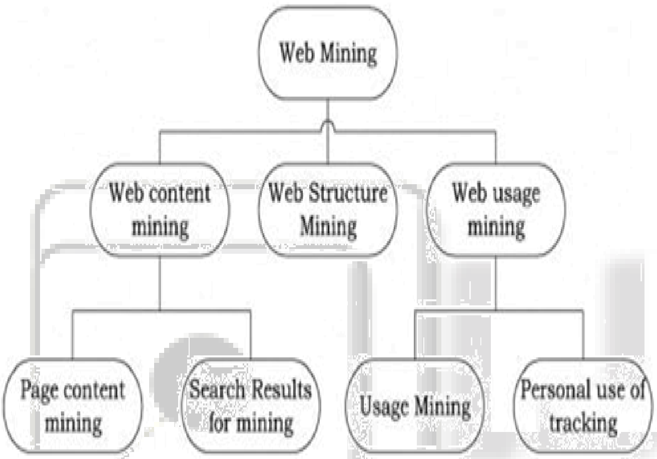


Fig. 1: Classified Tree of Web Mining

On the other hand side, after development of many technologies for semantic web mining still many conditions are there in which users are facing unreliable result. One of the most important reasons is identify exact requirement of user with the user defined query. Sometimes it happened because of ambiguity in the language itself and no context sensitive grammar present in that technology. Sometime the most dangerous case is users are not capable of describing the exact queries according to their requirement. In that case necessary condition is to learn users patterns [8] and determine their requirement to provide the exact information to them. In already present web data semantics details can not applicable. Many ideas and techniques are available in the literature portion for extracting the knowledge and information from the web.

II. Related Work

Tak-Lam Wong et al., [12] has developed a framework by using Bayesian learning approach. That is for adapting information extraction wrappers. In this paper, author gave an approach that can automatically take the extraction patterns of information from existing source web site and can transfer it to new sites, and at the same time bring new attribute in light with semantic properties. This text fragments generative model is related to the layout format and attributes. And it is designed to tackle the uncertainty. To tackle the wrapper adaption and new attribute discovery Bayesian learning approach and EM techniques are used. Experiments were conducted on more than 30 real world web sites and their domain was divided into three parts. And results indicate that this framework gives good performance. New attribute discovery can recognise the semantic label of that attributes. Major important characteristics of this model is that both site independent content and site dependent features are considered at the time of learning process, A limitation of this model is that the query that acts as a input of model is normally short sentences or phrases. Before the final step, a preprocessing step is required to extract the information from web document that increase the confidence in both the techniques-wrapper adaption and new attribute discovery job.

Gholamzadeh et al., [13] has proposed a novel approach to enhance the process of web service discovery and for development of service oriented applications. This discovery based approach is automatically searching semantic similarity between web services. This method used clustering methods application to check the similarity index. In this paper, a "fuzzy semantic clustering algorithm" is discovered. This algorithm makes the XML-structure of pre-exist WSDL file and establish ontology that will support semantic details. Experiments are performed on a 100 data set of web services with ontology on four different domains because of limitation of preprocessing. From the experimental result it is concluded that web service discovery job completed in reasonable time and improve the searching efficiency with performing well overall. One of limitation of this method is it will not work for automatic web service discovery.

Jayatilaka A.D.S et al., [14] has investigated the problem of retrieving knowledge from the big repository of web documents to develop ontologies. The proposed approach is a combination of web content mining and web usage mining. So both entities, web user's and web writer's view point can be captured which leads to extraction of conceptual relationships. The purposed approach is usable for large web area and it can be useful to create cross domain ontologies. And it will reduce price as well as time at the time of semantic web application development. This methodology can be used for search engine optimization threw which search engine can get a higher rank in web search environment. Still human involvement is there because ontology learning process is not converted into a fully automated process.

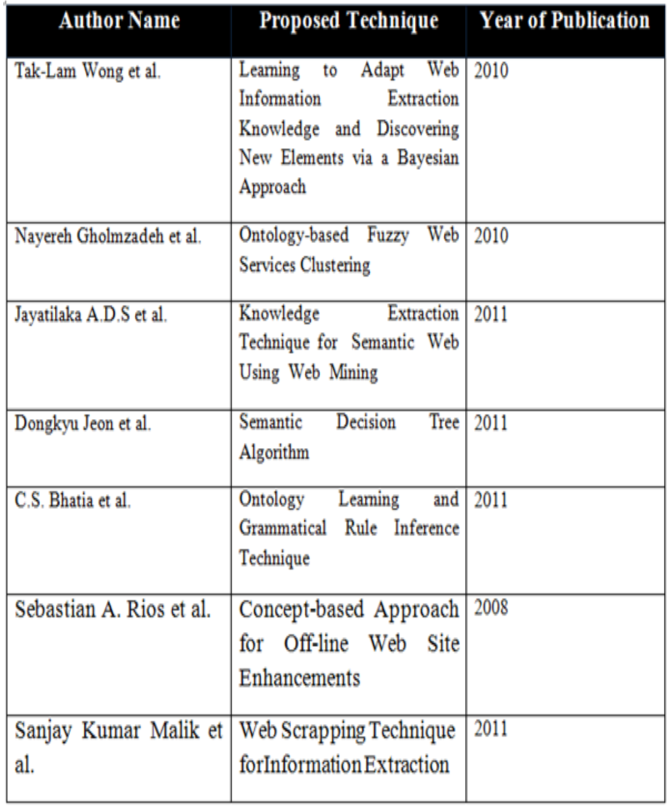
Dongkyu Jeon et al., [15] has proposed Semantic Decision Tree Algorithm for semantic web ontology. This algorithm can full fill the necessity of highly expected knowledge mining from big size ontology. Semantic web ontology has several characteristics to apply traditional decision tree algorithm which is the most popular in data mining classification. This algorithm will useful to mine the covered information and knowledge in the semantic web repository. This algorithm is useful to search variables for decision tree automatically based on network information of ontology. This algorithm contain an advance split condition named refinement like Cardinality Restriction Refinement, Domain Restriction Refinement, Concept Constructor Refinement, Qualification Refinement which make power of expression is more complex and rich. One major problem with this algorithm is sometime it create duplicate meanings for refinements. Target property of decision tree to relation is a data type property whose range is limited to Boolean typed value only.

C.S.Bhatia et al., [16] has invented a process of ontology learning to retrieve the information by using the Grammatical Rule Extraction Technique. This paper shows about the tough decomposition of a document into several individual segments. To identify the re-occurred features in web document Naïve Bays approach is used. Once the document is classified according to the ontology then it is easy to perform the critical semantic queries. This approach gives automatic processing to the tasks that perform the ontology development of particular document. Ontology learning by the Grammatical Rule Extraction Technique has obtained exact result of query. Use of Grammatical Rule Extraction Technique gives low percentage of error in index making. The result of this approach is not implemented on PDF and Postscript documents.

Sebastian A. Rios et al., [18] gave a concept based approach for offline web site enhancement. Because extracting of useful information from sites based on list of keyword searching techniques is somewhat difficult. Before this approach semantic WUM process used a concept-based approach in mining process of semantic web. The technique discussed in this paper can enhance the contents and structure of a web site offline. In this paper, proposed method was compared with four other WUM methods. After that a quality of enhancement was calculated using a survey to 100 visitors of the site, giving the effective result. This web mining process generates more semantically fit result according to the user's query. This approach is based on user's interest and correlation measurement, so by the result analysis it was proved that this approach obtain closer results for users according to their browsing preferences. Resultant performance is useful for 74% visitors which is better than the other methods(previously existed methods that were discussed in this paper) which are giving less than 50%. Classical WUM process took approximate 11 hours to complete the task whereas this approach took only 15 minutes for same task. Developed model is a very powerful method for solving many real time problems of classical WUM processes.

Sanjay Kumar Malik et al. [19] have discussed a web scraping technique for retrieving useful information from the HTML pages. This technique used a Prolog Server Pages (PSP). Semantic annotation technique is used here to add semantics rules and a valid structured to existed unstructured text documents. Knowledge Information Management (KIM) tools is used to extract semantic information. This paper highlights the web usage mining, semantic annotation and web scrapping and tells how to achieve better performance and efficiency at the information extraction time from the web repository. For further enhancement, semantic annotation in creation of ontology annotation may be used.

Table 3various Techniques On Semantic Web Mining



Author Name	Proposed Technique	Year of Publication
Tak-Lam Wong et al.	Learning to Adapt Web Information Extraction Knowledge and Discovering New Elements via a Bayesian Approach	2010
Nayereh Gholmzadeh et al.	Ontology-based Fuzzy Web Services Clustering	2010
Jayatilaka A.D.S et al.	Knowledge Extraction Technique for Semantic Web Using Web Mining	2011
Dongkyu Jeon et al.	Semantic Decision Tree Algorithm	2011
C.S. Bhatia et al.	Ontology Learning and Grammatical Rule Inference Technique	2011
Sebastian A. Rios et al.	Concept-based Approach for Off-line Web Site Enhancements	2008
Sanjay Kumar Malik et al.	Web Scrapping Technique forInformation Extraction	2011

III. SUMMARY

In recent world extracting the exact information from the web document in efficient manner is the main area of developer's and researcher's interest across the whole world. Without the help of semantic web, exact information extraction is impossible. Many techniques [Table 3] are suggested by the researchers for knowledge extraction from Semantic Web. Many semantic web mining techniques are discussed in the related work that will help the researchers to create the perfect semantic web mining technique in all the conditions. Because every technique has some limitation. Therefore, Combination of these techniques may help the researchers to develop the perfect technique for semantic web.

REFERENCE

[1] T. Berners-Lee, Weaving the Web: The Past, Present and Future of the World Wide Web by its Inventor, London, Texere.: 2000.

[2] Berners-Lee, T., "The Semantic Web," Scientific American, Vol.501., 2001.

[3] F. Harmelen, "THE SEMANTIC WEB: The Roles of XML and RDF", IEEE Internet Computing, vol. 15, no. 3, pp.63-74., 2000.

[4] Frank, M. & Eric, M., "RDF Primer", W3C Recommendation., 2004.

[5] D. Brickley, R.V. Guha, RDF vocabulary description language 1.0: RDF schema, W3C Recommendation, 2004.

[6] Deborah, L.M. & Frank, v.H., "OWL Web Ontology Language Overview", W3C Recommendation., 2004.

[7] T. Gruber., "A translation approach to portable ontology specifications". In: Knowledge Acquisition. 5: 199-199., 1993.

[8] D. Bollegala, Y. Matsuo and M. Ishizuka. 2010. A Web Search Engine-based Approach to Measure Semantic Similarity between Words. IEEE Transactions on Knowledge and Data Engineering. PP (99): 1.

[9] E. Hovy, Z. Kozareva, and E. Riloff, "Toward Completeness in Concept Extraction and Classification," 2009.

[10] K. Nakayama, T. Hara, and S. Nishio, "Wikipedia mining for an association web thesaurus construction," Berlin, Heidelberg: Springer-Verlag, 2007, pp. 322-334.

[11] C. Li and T.W. Ling, "From XML to Semantic Web," 2005, pp. 582-587.

[12] Tak-Lam Wong and Wai Lam "Learning to Adapt Web Information Extraction Knowledge and Discovering New Attributes via a Bayesian Approach," IEEE transactions on knowledge and data engineering, vol. 22, no. 4, 2010.

[13] Nayereh, Gholamzadeh, Fattaneh Taghiyareh "Ontology-based Fuzzy Web Services Clustering," 5th International Symposium on Telecommunications (IST'2010),2010.

[14] Jayatilaka A.D.S and Wimalarathne G.D.S.P "Knowledge Extraction for Semantic Web Using Web Mining," ICTer2011 : 089-094,2011.

[15] Dongkyu Jeon and Wooju Kim "Development of Semantic Decision Tree,"Data Mining and Intelligent Information Technology Applications (ICMiA), 2011, pp. 28-34.

[16] C.S.Bhatia and Dr. Suresh Jain "Semantic Web Mining: Using Ontology Learning and Grammatical Rule Inference Technique,"IEEE,2011.

[17] Wen-Wei Chen, "Data Warehouse and Data Mining Tutorial,"[M], Beijing: Tsinghua University Press, 2008, 4.

[18] Sebastian A. Rios and Juan D. Velasquez "Semantic Web Usage Mining by a Concept-based Approach for Off-line Web Site Enhancements," 2008 IEEE.

[19] Sanjay Kumar Malik, SAM Rizvi "Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation," 2011 IEEE.