A SEMANTIC BASED APPROACH FOR KNOWLEDGE DISCOVERY AND ACQUISITION FROM MULTIPLE WEB PAGES USING ONTOLOGIES

A.M.Abirami[1] and Dr.A.Askarunisa[2]

[1]Department of Information Technology, Thiagarajar College of Engineering, Madurai, Tamil Nadu, India
abiramiam@tce.edu
[2]Vickram College of Engineering, Enathi, Tamil Nadu, India
nishanazer@yahoo.com

ABSTRACT

The data from internet are dispersed in multiple documents or web pages. Most of them are not properly structured and organized. It becomes necessary to organize these contents in order to improve the search results by increasing the relevancy. The semantic web technologies and ontologies play a vital role in information extraction and new knowledge discovery from the web documents. This paper suggests a model for storing the web content in an organized and structured manner in RDF format. The information extraction techniques and the ontologies developed for the domain together discovers new knowledge. The paper also proves that the time taken for inferring the new knowledge is also minimal compared to manual effort when semantic web technologies are used while developing the applications.

KEYWORDS
Information Extraction, Semantic Web, Ontologies, RDF, SPARQL, Knowledge Discovery

1. INTRODUCTION

The World Wide Web has many documents which includes both relevant and irrelevant information. It becomes the user's responsibility to disregard unnecessary data and considers the necessary information. Also some web pages have unstructured or semi-structured textual content and the user needs to aggregate or segregate information based on his search needs. So Information Extraction from the web documents becomes predominant now-a-days.

The use of semantic web technologies like RDF, OWL, SPARQL, etc enables the information retrieval and extraction easier. A core data representation format for semantic web is Resource Description Framework (RDF). RDF is a data model for web pages. RDF is a framework for representing information about resources in a graph form. It was primarily intended for representing metadata about WWW resources, such as the title, author, and modification date of a Web page, but it can be used for storing any other data. It is based on triples subject-predicate-object that form graph of data [21].

All data in the semantic web use RDF as the primary representation language [16]. RDF Schema (RDFS) can be used to describe taxonomies of classes and properties and use them to create lightweight ontologies. Ontologies describe the conceptualization, the structure of the domain, which includes the domain model with possible restrictions [18]. More detailed ontologies can be created with Web Ontology Language (OWL). It is syntactically embedded into RDF, so like RDFS, it provides additional standardized vocabulary. For querying RDF data as well as RDFS and OWL ontologies with knowledge bases, a Simple Protocol and RDF Query Language

DOI : 10.5121/ijwest.2013.4306

(SPARQL) is available. SPARQL is SQL-like language, but uses RDF triples and resources for both matching part of the query and for returning results [17]. With the help of ontologies, the data is stored and organized in a meaningful way. It helps for context based search unlike the keyword based search, whereas the latter gives more irrelevant search results.

2. LITERATURE SURVEY

Raghu et. al [1] developed yellow pages service provider by using semantic web technologies and improved the search results by increasing the relevancy through feedback mechanism for geo-spatial and mobile applications. Richard Vlach et.al [2], developed a single schema to extract information from multiple web sources and handled ambiguous text. Wood et. al. [3] extracted the botanical data presented in the form of text and defined a model for correlating them using ontologies.

Rifat et. al [4], in his paper suggested the lazy strategy scheme for information extraction along with usual machine learning techniques by building the specialized model for each test instance. Jie Zou et. al [5] segmented HTML documents into logical zones for medical journal articles using Hidden Markov Model approach. In their technical report [6], Rifat Ozcan et. al used ontologies and latent semantic analysis technique to reduce the irrelevancy.

Ayaz et.al [7] discussed that the data from web sites can be converted to XML format for better information analysis and evaluation. Harish et. al [8] developed an interactive semantic tool to extract pertinent information from unstructured data and visualized it through spring graph using NLP and semantic technologies. James Mayfield et.al [9] discussed that the information retrieval could be tightly coupled with inference so that the semantic web search engines can lead to improvements in retrieval. Mohammed et.al [10], in his paper developed a web based multimedia enabled eLearning system for selecting courses to suit to students' needs using the dynamic mash-up technique.

Uijal et.al [11] presented a Linked Data approach to discover resume information enabling the task aggregation, sharing and reusing information among different resume documents. Chris Welty et.al [12], in his research paper, transformed the data into knowledge base and used deeper semantics of OWL to improve the precision of relation extraction. Maceij Janik et.al [15] in their paper classified Wikipedia documents using ontology and claimed that their model not required training set for classification, but can be done with the help of ontologies. Canan Pembe et. al [18] proposed a novel approach for improving the web search by representing the HTML documents hierarchically. They used semantic knowledge to represent the section and subsection of HTML documents.

3. METHODOLOGY

The general methodology adopted in this paper is diagrammatically represented in Figure 1 and explained in this section. Keywords are given to the search engines and the search results are analyzed further for relevant information extraction. The layout of different HTML pages is learnt and the pointer is moved to the section where the necessary data or information is present. We've conducted various experiments for different HTML's tags for retrieving the information, which in detail is explained in Section 4. Relevancy of web pages is determined with the help of synsets generated by the WordNet. Earlier, ontologies, or the vocabularies used for different domains are generated using Protégé tool [15]. The user query is mapped with the updated repositories and ontologies by

ontology mapping module and then the fine tuned results are given back to the user. This feedback mechanism introduced through ontology mapping increases the relevancy.

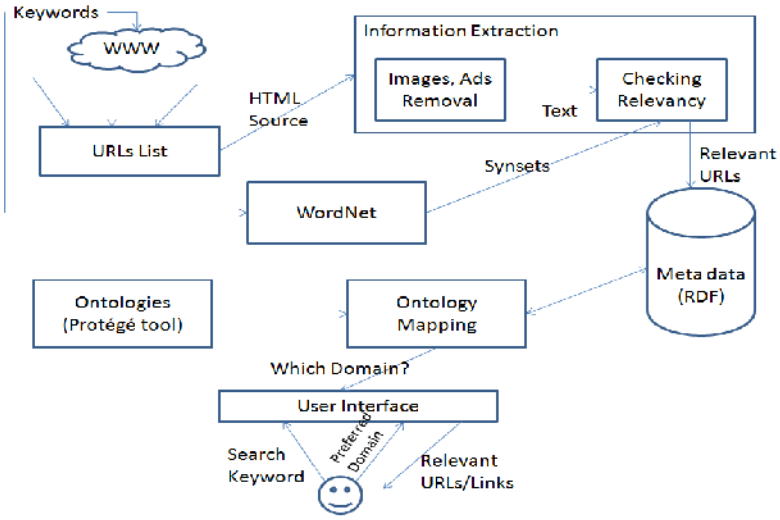

Figure 1. Generic model for relevant Information Extraction from web documents

Sample ontology representation [19] used for Experiment 1 is shown in Figure 2 where the relationship can be easily maintained for knowledge inference.

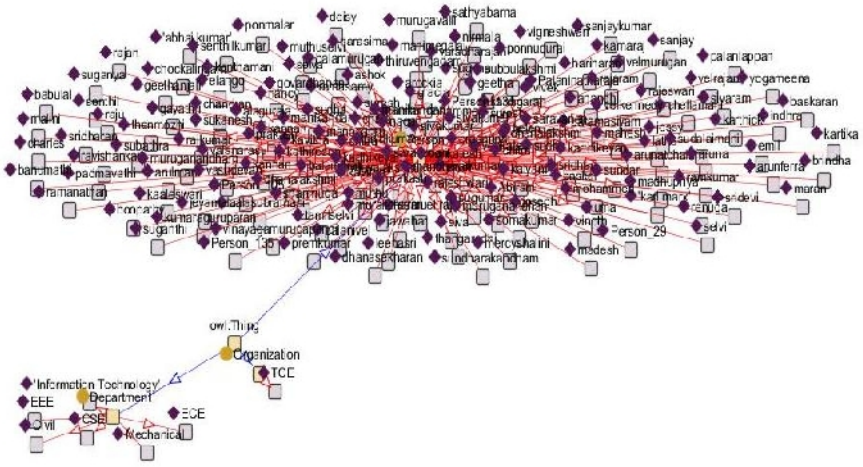

Figure 2. Ontology representation of College staff members

4. EXPERIMENTAL RESULTS

We've conducted two different experiments with HTML pages to extract information and to discover new knowledge from them. We've used RDF for storing the content and SPARQL and Jena [18] for querying RDF content. Ontologies created using Protégé tool is used for context based search. It establishes the relationship between the concepts and enables the increased relevancy.

4.1. Experiment 1

The staff profile in various formats like .html, .pdf, .doc is fed to the structured data extraction module after preprocessing and converted into RDF format. The model followed for this process is shown in Figure 3. For this experiment, we've taken our college web site (www.tce.edu) and collected the profiles of all staff members of all the departments. We've ignored the HTML pages in which the journal publication details may not present and have considered only the profiles which have the journal publication details. The number of documents taken for the experiment is shown in Table 1. We've used HTMLTidy parser to make the document syntactically clean document. By this process, all unclosed tags are closed and unwanted lines are removed.

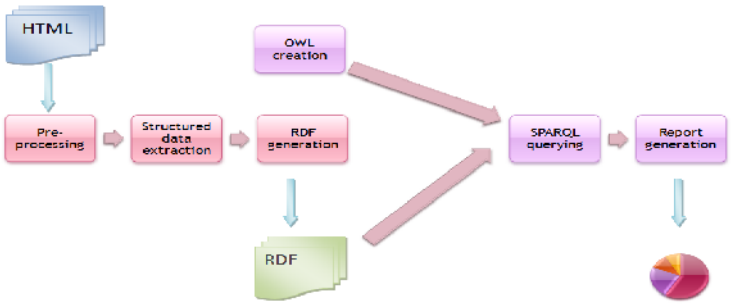

Figure 3. Model for Information Extraction from HTML pages (list tags)

Once the pre-processing phase gets completed, the clean html pages are fed into structured data extraction phase. This is accomplished by using the Wrapper Induction Technique [20], which is the supervised learning approach, and is semi-automatic. In this approach, a set of extraction rules is learned from a collection of manually labeled pages. The rules are then employed to extract target data items from other similarly formatted pages.

For example, the journal profile of a staff member after conversion may look like

<rdf:Description>
<staff:authors> A.M.Abirami </staff:authors>
<staff:title> An Enhanced ….. </staff:title>
<staff:journal>International Journal …. Technology </staff:journal>
<staff:year>2012</staff:year>
</rdf:Description>

Similarly profile of all staff members are converted into RDF format. The words related for each research area are stored in the RDF format. The exact domain in which each staff member is doing research is identified with the help of these words and their titles in the publications. The similarity between each staff member is measured with respect to their publications, thus the staff with similar profile is easily identified. Cosine similarity measure is used for this purpose. Latent semantic indexing can also be used as the alternate. These set of profiles can be recommended for new persons who fall in the same category of research. Or the staff member can easily find his/her research colleagues to strengthen their relationships between them. The model suggested in this paper thus is helpful for categorizing the web documents. We've developed a tool to give the visualization effect such that the tool gives report on the number of people working in the particular research field, number of publication from each department in each journal and the like. The Table 1 shows the new knowledge discovered from the set of profiles. Among the staff members of the department, the persons working in the same research category are easily discovered.

Table 1. Knowledge Discovered from staff member profile

Dept	Total Docs	Research Category	Matched Documents
CSE	14	Distributed Computing	CSE - 5, IT - 4, MCA- 1, ECE – 3
IT	11	Data Mining	CSE - 2, IT - 3, MCA – 2
MCA	7	Security	CSE - 3, IT – 3
ECE	26	Software Engineering	CSE - 2, IT - 1, MCA – 2
		Image Processing	CSE - 2, MCA - 1, ECE – 4

4.2 Experiment 2

Now-a-days all the colleges use web applications to maintain the students' personal and academic details, but it may be maintained and managed by one department. Not all department members are given access to the database and its tables. Sometimes it may become necessary to analyze the students' attendance and performance details, but we are left with data in the web pages. In this case, it is better if these data are converted into suitable machine learnable format so that inference and analysis can be easily made on the data. Otherwise, the human support is very much required for this process.

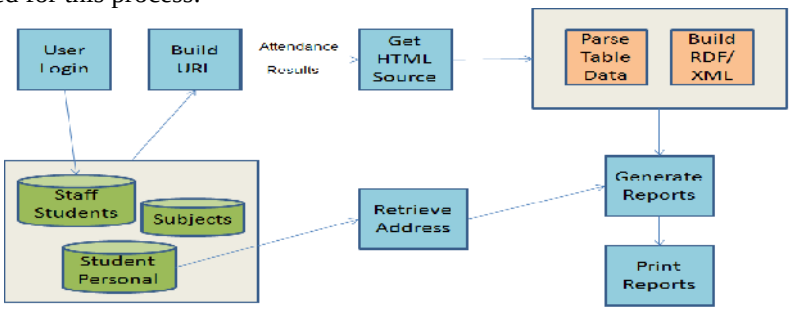

Figure 4. Model for Information Extraction from HTML pages (Tables)

We've followed the model given above to extract the relevant and required information about students from the different web pages as shown in Figure 4.

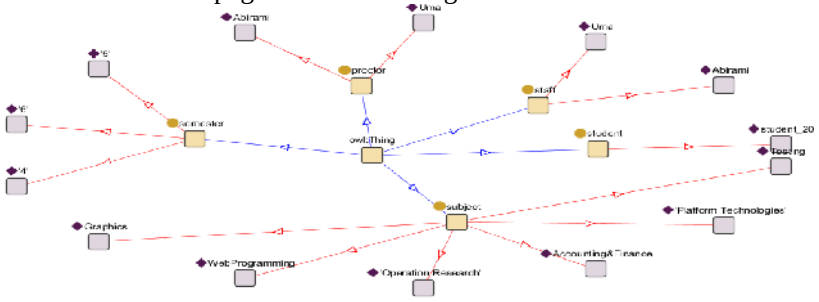

Figure 5. Ontology representation for Experiment 2

We've used DOM APIs to extract the content from HTML tables and converted into XML for querying. Ontology is used for building the URL and the tool developed learns the set of information to be retrieved from the login id and other databases. Ontologies are constructed for various concepts like student, staff member, proctor, subjects and semester. The relationship between each concept is established using OWL and the pictorial representation is shown in

Figure 5. We've extracted our students' data from the HTML pages into XML and RDF and used different XML query languages like XSLT, XPATH and SPARQL and compared the time taken for the inference and shown in Table 2.

Table 2. Knowledge Acquisiton from different web pages

HTML Records	Manual Effort (in mts)	Time taken (ms) Single XML/RDF file			Time Taken (ms) Multiple XML/RDF files	
		XSLT	XPath	SPARQL	XPath	SPARQL
60	30	90	78	207	6458	890
140	80	350	218	220	25272	1190
200	130	540	360	230	97984	1388
500	280	985	795	240	213783	2398

Here the student details like personal, attendance and test marks are displayed in different HTML tables. For example, the attendance of a single class is displayed in a web page. In order to collate details of a single student, we need to parse three or more web pages. But the information extraction module traverses all these web pages and parses the content into required format like XML or RDF and enables for easy query processing. Single XML file means that a group of students' details are stored in a file; multiple XML files means that each student detail is stored in separate files.

5. CONCLUSION

We have developed different semantic web applications to convert the unstructured or semi-structured text into XML/RDF format to enable easy machine processing. This approach leads to quick inferences and new knowledge discovery from the set of data. As a future enhancement, we will impart machine learning algorithms for classifying the web documents based on their content.

REFERENCES

[1] Raghu Anantharangachar1& Ramani Srinivasan, (2012) "Semantic Web techniques for yellow page service providers", International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.3.

[2] Richard Vlach & Wassili Kazakaos, (2003), "Using Common Schemas for Information Extraction for Heterogeneous Web Catalogs", ADBIS 2003, LNCS 2798, pp.118-132.

[3] M.M. Wood, S.J. Lydon, V. Tablan, D. Maynard & H. Cunningham, (2004), "Populating a Database from Parallel Texts Using Ontology-Based Information Extraction", LNCS 3136, pp. 254–264.

[4] Rifat Ozcan, Ismail Sengor Altingovde & Ozgur Ulusoy, (2012), "In Praise of Laziness: A Lazy Strategy for Web Information Extraction", ECIR 2012, LNCS 7224, pp. 565–568.

[5] Jie Zou, Daniel Le and George R. Thoma .: Structure and Content Analysis for HTML Medical Articles: A Hidden Markov Model Approach. DocEng'07, ACM 978-1-59593-776.

[6] Rifat Ozcan, Y & Alp Aslandogan, (2004), "Concept Based Information Retrieval Using

Ontologies and Latent Semantic Analysis", A Technical Report, CSE-2004-8.

[7] Ayaz Ahmed Ayaz Ahmed Shariff K, Mohammed Ali Hussain & Sambath Kumar, (2011), "Leveraging Unstructured Data into Intelligent Information – Analysis & Evaluation", IPCSIT Vol.4, International Conference on Information and Network Technology.

[8] Harish Jadhao, Dr. Jagannath Aghav & Anil Vegiraju, "Semantic Tool for Analysing Unstructured Data", International Journal of Scientific & Engineering Research, Volume 3, Is. 8.

[9] James Mayfield & Tim Finin, (2003), "Information Retrieval on the Semantic Web: Integrating inference and retrieval", SIGIR 2003 Semantic Web Workshop.

[10] Mohammed Al-Zoube & Baha Khasawneh, (2010), "A Data Mashup for Dynamic Composition of Adaptive Courses", The International Arab Journal of Information Technology, Vol. 7, No. 2.

[11] Ujjal Marjit, Kumar Sharma & Utpal Biswas, (2012), "Discovering resume information using Linked data", "International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.2.

[12] Chris Welty & William Murdoc. J, (2006), "Towards knowledge acquisition from Information Extraction", ISWC'06 Proceedings of the 5th international conference on The Semantic Web, pages 709-722.

[13] Maciej Janik & Krys Kochut, (2007), "Wikipedia in action: Ontological Knowledge in Text Categorization", University of Georgia, Technical Report No. UGA-CS-TR-07-001.

[14] Canan Pembe & Tunga Gungor, (2007), "Heading based Sectional Hierarchy Identification for HTML Documents". International Symposium on Conference in Computer and Information Sciences, ISCIS 2007.

[15] http://protege.stanford.edu/
[16] http://www.w3schools.com/rdf/rdf_example.asp
[17] http://www.w3.org/TR/rdf-sparql-query/
[18] http://jena.sourceforge.net/tutorial/RDF_API/
[19] http://www.obitko.com/tutorials/ontologies-semanticweb/ontologies.htm

[20] Bing Liu, (2007), "Web data mining - Exploring hyperlinks,contents,and usage data", Springer, New York.

[21] Grigoris Antoniou & Frank van Harmelen, (2003), "A semantic Web Primer". MIT Press, Cambridge, England.

Authors

A.M.Abirami is currently doing her research in information extraction using semantic web technologies.

Dr.A.Askarunisa completed her PhD in Software Testing. Currently she is guiding research scholars under software engineering and data mining.