

Using Ontologies in Web Mining for Information Extraction in Semantic Web: A Summary

Vishal Jain¹, Sanjay Kumar Malik²

¹M.Tech (CSE) IV Sem, University School of Information Technology, GGS Indraprastha University

²University School of Information Technology, GGS Indraprastha University

Abstract

A huge amount of data is available on the web which may be in structured, semi-structured or unstructured form. The need is to organize this data in a formal system which results in more relevant, useful and structured information. Ontology may be a mechanism for obtaining the information on web in a more structured way in semantic web. Web mining technique may be useful to discover and extract meaningful information from the Web documents. This paper focuses on presenting a theoretical framework of how Ontologies may be useful for data mining process for information retrieval in Semantic Web.

Keywords: Semantic Web, Ontology, Information Retrieval, Web Mining

1. Introduction

The Semantic Web addresses the first part of this challenge by trying to make the data also machine understandable in the form of Ontology, while Web Mining addresses the second part, by semi-automatically extracting the useful knowledge hidden in these data, and making it available. Information retrieval is synonymous with “determination of relevance”. Information retrieval is described as the task of identifying documents in the collection on the basis of properties approved to the documents by the user requesting the retrieval [13].

2. Semantic Web

Semantic Web technology is to address the problem by structuring the content of the web and extract maximum benefit from the processing power of machines and existing web. One of the most important things missing in current web is establishing links between the resources encoding semantic information [7]. This Semantic Web can be achieved by adding semantic structures to the current Web. Many candidate techniques have been proposed, such as semantic networks, conceptual graphs, the W3C Resource Description Framework (RDF) and XML Topic Maps [8]. Semantic Web plays an important role in extraction or retrieval of meaningful data using web mining.

3. Information Retrieval

Information Retrieval is the task of identifying documents in a collection on the basis of properties described to the documents by the user requesting the retrieval [2]. Current information retrieval techniques on web are not intelligent enough to exploit the meaning of data i.e.; semantic knowledge within documents and hence cannot give precise answers to precise questions. Information retrieval may be expressed in the form as shown in figure 1. Different types of data are available on the Web viz, structured, unstructured and semi structured. Structured data is in the form of text document. Semi structured data are not full and grammatical text.

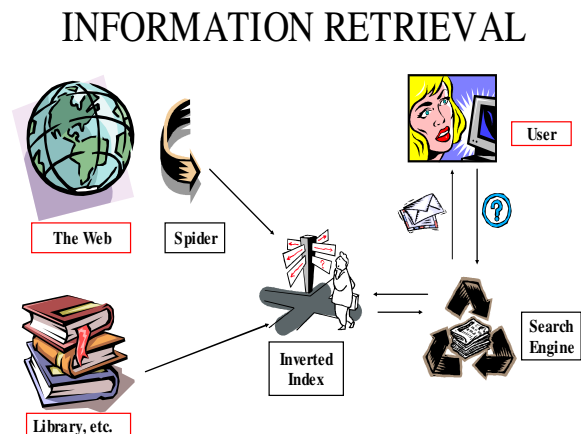


Figure 1 “Information Retrieval Method “[3]

Semantic web envisions the future web as pages text as well as semantic markup [6]. Different types of data available on the Web. Structured data are list, tree, and data in the form of table. Unstructured data is in the form of text document. Semi-structured data is a form of structured data that does not conform to the formal structure of tables and data models associated with databases but contains tags which are nonetheless or other markers to separate semantic elements and hierarchies of records and fields within the data. Information Retrieval is a field concerned with the structured analysis, organization, storage, searching, and retrieval of information [1]. For maintaining Semi Structured data, we may develop Ontology in Semantic Web.

4. Ontology

The term ontology can be defined in many different ways. Genesereth and Nilsson defined Ontology as an explicit specification of a set of objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold them [4]. Usually, Ontologies are defined to consist of abstract concepts and relationships (or properties) only. In some rare cases, Ontologies are defined

also to include instances of concepts and relationships[5]. Various algorithms may be proposed for extracting information from collection of web pages across different sites.

4.1 Ontology Design and Development

There are various tools to develop Ontology i.e Altova Semantic Works, Protégé, and Onto Lingua. Protégé is one of most widely used tool to develop Ontology. An Example of Bio Tech Department Ontology developed in Protégé version 3.4 is as follows:

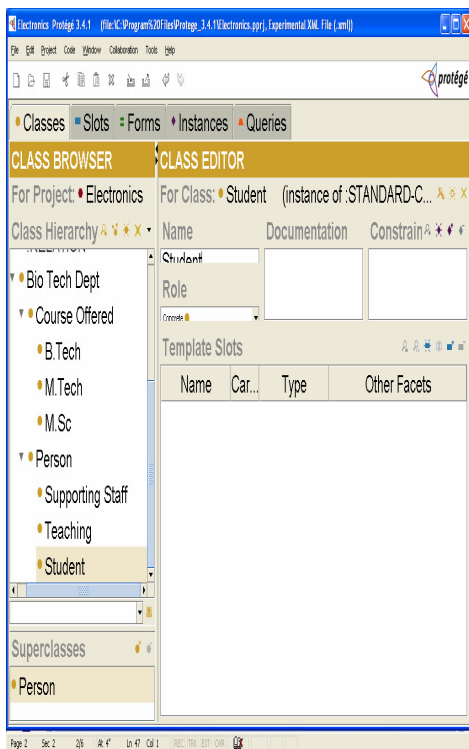


Figure 2 “Bio Tech Department Ontology”

Data Mining and Ontology may be related in two manners:

1. From ontologies to data mining, we can incorporate knowledge in the process with the use of ontologies, i.e. how the experts carry out the analysis tasks [12].
2. From data mining to Ontologies, we can include domain knowledge as the input information and use the ontologies to represent the results [12].

5. Web Mining

5.1 Introduction

Web mining is a Knowledge database process applied to Web data. A large amount of information is available on the Web which lacks structure where web mining may be useful. Web mining refers to discovery and analysis of useful information over the World Wide Web. The Web mining field encompasses a wide array of issues, primarily aimed at deriving actionable knowledge from the Web, and includes researchers from information retrieval, database technologies, and artificial intelligence [9].

5.2 Types of Web Mining

Web Mining can be classified into three categories: Web content mining (WCM), Web structure mining (WSM), and Web usage mining (WUM) [18].

5.2.1 Web Content Mining

Web Content Mining refers to mining of desired content over World Wide Web. Various search engines exist for the web content mining, such as AltaVista, Lycos, Web Crawler, Meta Crawler etc. There are various techniques to extract structured data using Web content mining. Some of the techniques are [10]:

- Web Crawler
- External crawler
- Internal crawler
- Wrapper Generation
- Page content mining

5.2.2 Web Structure Mining

Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level to generate structural summary about the Website and Web page.

5.2.3 Web Usage Mining

Web Usage Mining refers to automatic knowledge mining of user access patterns from web servers. Web usage mining is a kind of data mining that it mines the information of access routes kept back in servers, i.e. the information of access manners of user visits the web sites after users browse the web pages. The motive of mining is to find users' access models automatically and quickly from the vast web log data, such as frequent access paths, frequent access page groups and user clustering, etc [11]. Web Usage Mining has been defined as the application of data mining techniques to large Web data repositories in order to extract usage patterns.

5.2.3.1 Uses of Web Usage Mining in Search Engine

Queries and related clicks can be used to improve the search engine itself in different aspects: user interface, index performance, and answer ranking [14].

For Traversal Path Patterns

Traversal path pattern mining is based upon the availability of traversal paths that must be obtained from raw Web logs [15].

5.2.3.2 Log Analyzer

WebLog Expert is a fast and powerful access log analyzer. It will give information about site's visitors: activity statistics, accessed files, paths through the site, information about referring pages, search engines, browsers, operating systems, and more. WebLog Expert can analyze logs of Apache and IIS web servers. It can read GZ and ZIP compressed log files so that we won't need to unpack them manually [16]. Refer figure 3 for the usage pattern of the web server data logs as obtained by Weblog analyzer.

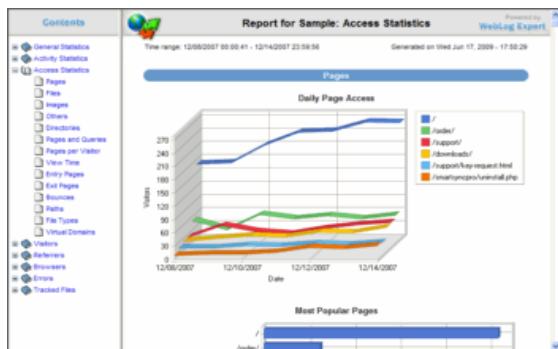


Figure 3 “Sample Log File “ obtained by WebLog Analyzer.

5.3. Semantic Web Mining

Semantic Web Mining is the integrating two technologies viz; semantic web and web mining. Following may be used as the framework of Data Mining with Ontology in Semantic Web:

5.3.1 Ontology for Data Mining

5.3.1.1 Metadata Ontologies

These Ontologies establish how this variable is constructed i.e. which was the process that permits us to obtain its value, and it can vary using another method. Of course this ontology must also express general information about the variable as is treated [12].

5.3.1.2 Domain Ontologies

These Ontologies explains the knowledge about application domain [12].

5.3.1.3 Ontologies for data mining process

These Ontologies codify all knowledge about the process, i.e. select features, select the best algorithms according to the variables and the problem, and establish valid process sequences [12].

Figure 4 shows the vision of Data Mining with Ontology Cycle.

6. Conclusion

This paper focuses on the significance of web mining in semantic web for information retrieval from web using an Ontology and refers a framework for the same. It may assist beginner researchers to start working in the area of web mining in semantic web.

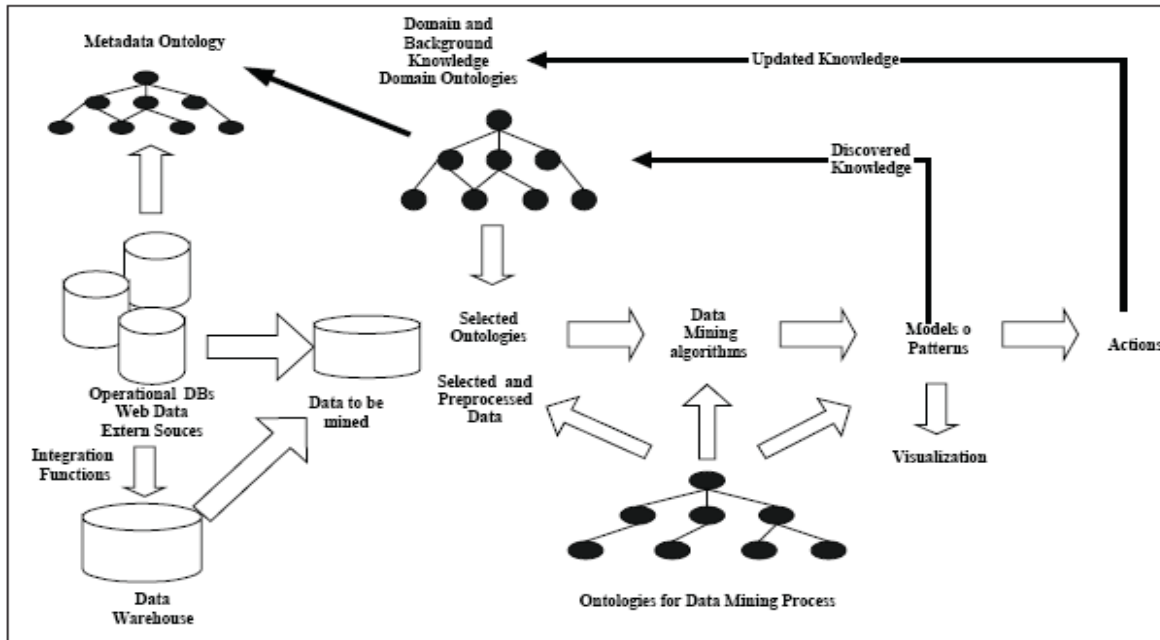


Figure 4 “Framework of Data Mining with Ontology in Semantic Web” [12]

7. References

- [1] G. Salton, “Automatic Information Organization and Retrieval”, McGraw-Hill, New York, 1968.
- [2] CARLO MEGHINI, FABRIZIO SEBASTIANI, AND UMBERTO STRACCIA, “A Model of Multimedia Information Retrieval”, Journal of the ACM, Vol. 48, No. 5, September 2001, pp. 909–970.
- [3] Jaime Carbonell, “Information Retrieval”, How to Power a Search Engine, 2003.
- [4] Genesereth, M. R., and Nilsson, N. J., Logical Foundations of Artificial Intelligence, San Mateo, CA: Morgan Kaufmann Publishers, 1987
- [5] R.M. Suresh, “A study on the ontology based web mining for digital library”, IET-UK International Conference on Information and Communication Technology in Electrical Sciences (ICTES 2007),
- [6] Urvi Shah, Tim Finin, Anupam Joshi, “Information retrieval on the semantic web”, ACM 1-58113-492-4/02/0011 *CIKM'02*, Nov 4-9, 2002
- [7] Ying Ding, Cornelis, Iadh Ounis and Joemon Jose, “ACM SIGIR Workshop on Semantic Web,” *SWIR 2003*
- [8] Benedicte Le Grand, Michel Soto, “XML Topic Maps and Semantic Web Mining”, Semantic Web Mining Workshop, Conférence ECML/PKDD 2001
- [9] Pranam kolari and Anupam joshi, “Web Mining : Research and Practice”, 2004 IEEE Copublished by the IEEE CS and AIP.
- [10] Kshitija Pol, Nita Patil, Shreya Patankar, Chhaya Das, “A Survey on Web Content Mining and extraction of Structured and Semistructured data”, First

- International Conference on Emerging Trends in Engineering and Technology, IEEE
- [11] Junjie Chen and Wei Liu, “Research for Web Usage Mining Model”, CIMCA-IAWTC'06, 0-7695-2731-0/06, IEEE
- [12] Héctor Oscar Nigro, Sandra Elizabeth González Císaro, Daniel Hugo Xodo, Data Mining with Ontologies: Implementations, Findings, and Frameworks, Information Science Reference
- [13] P.S. Bhatia, Sanjay Malik, Poonam Yadav , “ A Survey on Retrieval Methods”, Emerging Trends and applications in Computer Engineering [NCETA-2007] , Ajmer.
- [14] Ricardo Baeza-Yates, “Web Usage Mining in Search Engines”, Web Mining : Application and Techniques, Idea Group Publishing
- [15] Zhixiang Chen, Richard H. Fowler, Ada Wai-Chee Fu, Chunyue Wang, “Efficient Web Mining for Traversal Path Patterns”, Computational Science and Its Applications – ICCSA 2005
- [16] <http://www.weblogexpert.com/>
- [17] Rigitte Trousse, Marie-Aude Aufaure, Bénédicte Le Grand, Yves Lechevallier and Florent Masegla, “Web Usage Mining For Ontology Management », RP-LIP6-2007-10-21 Workshop
- [18] Olfa Nasraoui, Myra Spiliopoulou, Jaideep Srivastava, Bamshold mobasher, « Advances in Web Mining and Web Usage Analysis, 8th International Workshop on Knowledge Discovery on the Web, WebKDD 2006, Springer.