Mining the Semantic Web

Ajay Chakravarthy

Department of Computer Science, University of Sheffield,
Regent Court, 211 Portobello Street, Sheffield, S1 4DP, United Kingdom,
A.Chakravarthy@dcs.shef.ac.uk

Abstract: *In this paper we propose research on how semantic web technologies can be used to mine the web, for information extraction. We also examine how new unsupervised processes can aid in extracting precise and useful information from semantic data, thus reducing the problem of information overload .The Semantic Web adds structure to the meaningful content of Web pages; hence information is given a well-defined meaning; which is both human readable as well as machine-processable. This enables the development of automated intelligent systems, allowing machines to comprehend the semantics of documents and data. Here we propose techniques for automating the process of search, analysis and categorization of semantic data, further we examine how these techniques can aid in improving the efficiency of already existing information retrieval technologies by implementing reporting functionalities, which is highlighted in the future work and challenges.*

1. Introduction

The Semantic Web (SW) can be defined as an extension of the current web. Here the information is presented in a well-defined manner, better enabling computers and people to work in cooperation. Data in the Semantic Web is defined and linked in a way that can be used for more effective discovery, automation, integration and reuse across applications. This data can be shared and processed by automated tools as well as people. *The Semantic Web will provide an infrastructure that enables not just web pages, but databases, services, programs, sensors, personal devices, and even household appliances to both consume and produce data on the web.* (Hendler et al, 2002). There are various ways in which the applications using semantic data can communicate, hence in order to establish a common framework for representation of semantic information the Resource Description Framework (RDF) is introduced. The Semantic Web is known for being a web of Semantic Web Documents (SWDs) rather than for using various technologies; however, little is known about the structure or growth of the web of Semantic Web Documents. Today's search engines deal with SWDs poorly, since they have been developed to process text documents. Most make no attempt to parse XML documents into appropriate tokens and none take advantage of the structural and semantic information encoded in a SWD. This paper investigates the techniques for effective crawling, indexing, analysis and classification of semantic data.

2. Motivation

The process of mining for semantic data involves several processes namely, crawling the web for semantic web documents, extracting and analysing information from this data, clustering the semantic data for later retrieval purposes, and its scope for the future which involves enhancing the capability of information extraction systems by adding reporting functionality which involves tracking changes in information over time. We try to give an insight into these aspects in the following section.

The task of mining the Web for Semantic Data essentially consists of crawling the web and finding Semantic Web Documents, which are stored in the form of RDF, OWL, FOAF, RSS, etc at various locations. This leads us to the idea of designing a robust RDF crawler. *Crawling the semantic web is essentially identical to crawling the HTML content web - it's simply a case of choosing one or more starting points, downloading a resource and following the pointers in it to further resources.* (Biddulph, 2003). The difference between gathering HTML and RDF data is that RDF has a well defined mechanism for merging multiple RDF models. We many combine any number of RDF models to produce a single unified model. Hence instead of performing the task of building a database of keywords and links to locations where HTML representations related to those keywords can be found, the RDF crawler can create a combined model for all the semantic data found. The major advantage of this union of models is that the model now becomes a rich resource of information. That is one document contains the combined information of the all the separate documents which contain fragments of data. Some of the design considerations while implementing such a crawler could be Resource Pooling to avoid overload on the server, Gathering URLs from certain targets in RDF representation E.G the `<rdf:seealso>` triples that contain additional information about a document and mapping of the ontology to the data. After download of Semantic Data is complete, we now have to move to the second part of the process that is extraction and analysis of information from the data, one of the most efficient ways to extract information from RDF graphs is by using RDQL (RDF data query language). RDQL is now supported by many popular RDF API frameworks such as Jena[1]. Figure 1 shows the proposed design for the RDF crawler. The next section describes how this can be used in clustering documents according to similarity.

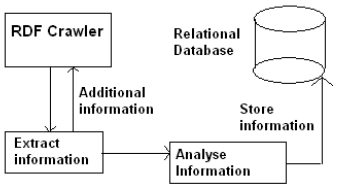

Fig. 1. RDF Crawler Design

Semantic Web Document clustering is an Open Research Topic and has not been experimented with until now, the advantage of this technique is that precision and recall rates of web searches can be significantly enhanced, thus reducing the problem of information overload. An enhanced version of the Suffix Tree Algorithm (Zamir and Etzioni) can be used to categorize documents according to their type. The high level structure hence produced when stored in the form of a tree it will have the advantages of faster fetch rates and a hierarchically ordered information structure. The information contained in such a high-level structure will be very precise and easy to retrieve.

[1] http://jena.sourceforge.net/

2. Related Work

Staab (Staab et al, 2004) presents work on the Ontotext RDF crawler which downloads interconnected fragments of fragments of RDF from the Internet and builds a knowledge base from its data. A host of URIs to be retrieved as well as URI filtering conditions are maintained at every phase of RDF crawling. This is done in order to download the resources containing RDF iteratively. To enable embedding in other tools, RDF Crawler provides a high-level programmable interface (Java API). Other work done in this field includes the Hackdiary [2]RDF Crawler which is a multithreaded java implementation capable of downloading simultaneously from many sources while the aggregation thread does the processing. It builds a model that remembers the provenance of the RDF and takes care to delete and replace triples if it hits the same URL twice. Hence the data is up-to-date all the times even after many runs. Our proposed work is different from the above mentioned in two ways, first and foremost our work is focussed in how to classify and categorize semantic data in-order to increase search precision. Secondly we propose a method for clustering semantic data and there is no reported work done in this area to date and is an open research topic. This process is better demonstrated by the experimental tool we developed the RDF Analyser as shown in Figure 2. The RDF Analyser is able to extract all information from structured data including information from anonymous nodes. In the next section we describe how information obtained from the semantic web can be used for reporting data, i.e. the ability for IE systems to be able to track changes in information over time and predict new information.

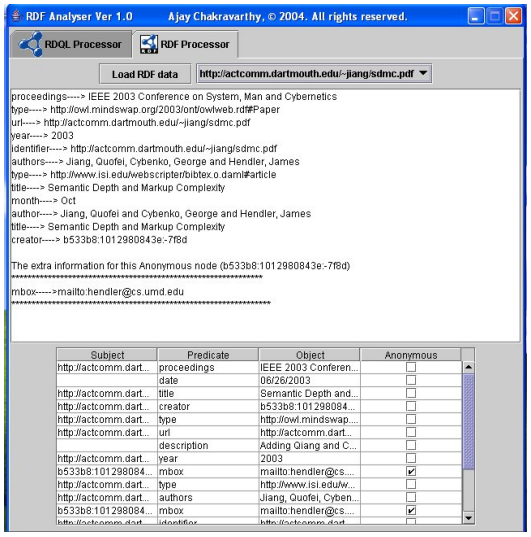

Fig. 2. RDF Analyser

3. Future Work

The Information Extraction systems E.G Armadillo[3], work on the principle of utilizing the redundant information on the web by using multiple citations as a way of validating the data. The valid data is then used to bootstrap the annotation process by using IE Annotation engines such as Amilcare (Ciravegna et al.). Hence producing machine-readable content for the Semantic Web i.e. Semantic Web Documents. Armadillo outputs RDF documents after crawling the web. Armadillo is currently able to learn over the HTML content of the World Wide Web. However if the IE system is able to learn from semantic data E.G RSS and XML feeds which are now increasingly being provided by most websites and implement a mechanism to track the changes information over this data. Say for example If an Article 'A' says that the

[2] http://www.hackdiary.com/archives/000030.html
[3] http://nlp.shef.ac.uk/wig/armadillo_home.html

cost of an Item ‘X’ is £ 300 in the year 1994 , and article ‘B’ says its cost has now risen to £ 350 in 1995 and article ‘C’ quotes that the cost of item ‘X’ is now £ 400 in 1996, now if the Intelligent System is able to track and match these changes in information , then it can successfully predict that the cost of the item ‘X’ in 1997 will be £ 450, after observing the trend. This is a simple example of what the system can aim to achieve. This concept can be used for more important tasks such as E-Commerce applications where the prices of goods need to be monitored at a regular basis or in the stock market where the stock quotes are updated frequently. If the system is able to give the user an idea of how the information might change in the near future after learning from changes in the past, then it can be of great help in the above mentioned areas.

This paper has investigated the various future challenges that the Semantic Web poses, in the areas of information extraction and retrieval. Also we have proposed ways in which semantic data can be mined from the web, for later analysis and categorization. Hence we are able to conclude that although significant success has been achieved in IE, there is still scope for interesting research in this area with the use of SW technologies.

References

D. Beckett, (2004) Redland RDF Application Framework, Institute for Learning and Research Technology, University of Bristol.

E. Miller (1998) An Introduction to the Resource Description Framework, Online Computer Library Centre, Inc.Office of Research, Dublin, Ohio.

F. Ciravegna, S. Chapman, A. Dingli, and Y. Wilks. Learning to Harvest Information for the Semantic Web. Department of Computer Science, The University Of Sheffield.

J. Hendler, T.B. Lee, E. Miller (2002) Integrating Applications on the Semantic Web, Computer Science Dept, University of Maryland, World Wide Web Consortium. Semantic Web Activity Lead

M. Biddulph (2003) Crawling the Semantic Web. BBC London, United Kingdom.

M. Steinbach, G. Karypis, V. Kumar. A Comparison of Document Clustering Techniques, Department of Computer Science and Engineering, University of Minnesota.

O. Zamir , O. Etzioni. Web Document Clustering: A Feasibility Demonstration, Department of Computer Science and Engineering, University of Washington.

R. Guha, R. McCool, E. Miller, (2004) Semantic Search. Stanford University.

R. Lee, (2004) Scalability Report on Triple Store Applications, Massachusetts institute of technology.

S. Staab, K. Apsitis, S. Handschuh, H. Oppermann, (2004) Specification of an RDF Crawler.