International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 1 Issue: 12 919 – 922

Semantic Web Mining: Review

Arti Shukla(Student)
AIM & ACT. Banasthali Vidyapith
Rajasthan, India
Shukla.arti11@gmail.com

Akanksha (Student)
AIM & ACT. Banasthali Vidyapith
Rajasthan, India
dearmonaaaaa@gmail.com

Priyanka Yadav
AIM & ACT. Banasthali Vidyapith
Rajasthan, India
engg.priyankayadav@gmail.com

***Abstract*—**This paper describes about Semantic Web Mining. The Purpose of this paper is to focus on how semantic web technologies can be used to mine the web , for relevant information extraction. Semantic Web Mining is about combining the two emerging research areas Semantic Web and Web Mining. Researchers work on improving the result off web mining by using semantic structure in the web and make use of Web Mining techniques for building the Semantic Web. In this manner both technologies are playing vital role to each other. 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 paper gives an overview of where the two areas meet today , and sketches ways of how a closer integration could be profitable.

Keywords*-Semantic Web;Web Mining ; Ontologies; Knowledge engineering; World Wide Web*

_____*****_____.

I. INTRODUCTION

On the Web most data are so unstructured that they can only be understood by humans , but the amount of data is so huge that they can only be processed efficiently by machines. Due to huge data it is very difficult to access relevant information when required. To overcome with this use of two fast-developing research area Semantic Web and Web Mining both are fruitful in this aspects. Semantic Web and Web Mining build on the success of the World Wide Web(WWW). Semantic Web is useful to make the data machine-understandable , while Web Mining extract the useful knowledge hidden in these data, and making it available as an aggregation of manageable proportions.

Web Mining techniques can be applied to help create the Semantic Web. A backbone of the Semantic Web are ontologies . The challenge is to learn ontologies and its concepts. In Semantic Web ontologies can be used to improve the process and result of Web Mining. Recent developments include the mining of sites that become more and more Semantic Web sites and the development of mining techniques that can use the power of Semantic Web knowledge representation . The tight integration of Semantic Web and Web Mining will greatly increase the understandability of the Web for machines, and will thus become the basis for further generations of the intelligent Web tools. In this manner both emerging technologies are complimenting each other to fetch out relevant information from large pool of data using various technologies. Type Style and Fonts

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

II. SEMANTIC WEB

Semantic Web is an extension of current web, in which information is given well-defined meaning, better enabling computer and people to work in cooperation [1] .

The Semantic Web is based on a vision of Tim Bemers-Lee , the inventor of the WWW . The great success of the current WWW leads to a new challenge : A huge amount of data is interpretable by humans only ; machine support is limited. Bermers-Lee suggests to enrich the Web by machine-processable information which supports the user in his tasks. For instance, today's search engines are already quite powerful, but still too often return excessively large or inadequate lists of hits. Machine processable information can point the search engine to the relevant pages and can thus improve both precision and recall. For instance , today it is almost impossible to retrieve information with a keyword search when the information spread over several pages.

Semantic web aims at offering a solution, capturing and exploiting the meaning of terms to transform current web from information-presentation platform to a platform that focuses on understanding and reasoning with the information [2].

The following steps show the direction where the Semantic Web is heading:
1. Providing a common syntax for machine understandable statements.
2. Establishing common vocabularies.
3. Agreeing on a logical language.
4. Using the language for exchanging proofs.
In Semantic Web Ontology approach is needed to describe. An ontology is "an explicit formalization of a shared understanding of a conceptualization"[3].

III. WEB MINING

The development of World Wide Web and its usage grows, it will continue to generate ever more content, structure, and usage data and the value of Web mining will keep increasing. Research needs to be done in developing the right set of Web metrics, and their measurement procedures, extracting process models from usage data, understanding how different parts of the process model impact various Web metrics of interest, how the process models change in response to various changes that are made-changing stimuli to the user, developing Web mining techniques to improve various other aspects of Web services, techniques to recognize known frauds and intrusion detection.

Web Mining is the application of data mining techniques to the content, structure, and usage of Web resources. It is thus "the nontrivial process of identifying valid, previously unknown, and potentially useful patterns" [4] in the huge amount of theseWeb data, patterns that describe them in concise form and manageable orders of magnitude. Like other data mining applications, Web Mining can profit from given structure on data (as in database tables), but it can also be applied to semi-structured or unstructured data like free-form text. This means that Web Mining is an invaluable help in the transformation from human-understandable content to machine-understandable semantics.
Three areas of Web Mining are commonly distinguished:
Content mining, structure mining, and usage mining [5,6,7]. In all three areas, a wide range of general data mining techniques, in particular association rule discovery, clustering,classification, and sequence mining, are employed and developed further to reflect the specific structures of Web resources and the specific questions posed in Web Mining

IV. SEMANTIC WEB MINING

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.

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[8]. 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 are 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.

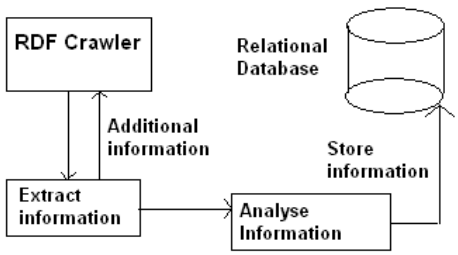

Fig. 1. RDF Crawler Design

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 [9] 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.

V. RELATED WORK

Staab [10] 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.

VI. FUTURE WORK

The Information Extraction systems E.G Armadillo 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 [11].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.

VII. CHALLENGES

Extracting an ontology from the Web is a challenging task. One way is to engineer the ontology by hand, but this is expensive. In [14], the expression *ontology learning* was coined for the semi-automatic extraction of semantics from theWeb. There, machine learning techniques were used to improve the ontology engineering process and to reduce the effort for the knowledge engineer. Ontology learning exploits many existing resources including texts, thesauri, dictionaries, and databases (see [15] as an example of the use ofWordNet). It builds on techniques from Web content mining, and it combines machine learning techniques with methods from fields like information retrieval [16] and agents [17,18], applying them to discover the 'semantics' in the data and to make them explicit. The techniques produce intermediate results which must finally be integrated in a machine understandable format, e.g., ontology.

VIII. CONCLUSION

In this paper, we have studied the combination of the two fast developing research areas Semantic Web and Web Mining. Semantic Web Mining aims at combining the two fast-developing research areas Semantic Web and Web Mining. This survey analyzes the convergence of trends from both areas: More and more researchers are working on improving the results of Web Mining by exploiting semantic Structures in the Web, and they make use of Web Mining techniques for building the Semantic Web.). Semantic Web is useful to make the data machine-understandable , while Web Mining extract the useful knowledge hidden in these data, and making it available as an aggregation of manageable proportions .Discussion of technology used in Semantic Web Mining which includes RDF crawler. Using Semantic Web Mining relevant information from various huge pools of data can be fetch. There are Various research work going on Semantic Web Mining which is discussed.

REFERENCES

[1] Schwartz, DG, 2003. From open is semantics to the semantic web: The Road Ahead, IEEE Intelligent systems, pp 52-58.

[2] McGuinness, DL, Fikes, R, Hender, J, Stein, LA, 2002. DAML+OIL:

An ontology language for the semantic web, IEEE Intelligent systems, pp 72-80.

[3] T.R. Gruber, Towards Principles for the Design of Ontologies Used for Knowledge Sharing, in: N. Guarino, R. Poli (Eds.), Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer, Deventer, Netherlands, 1993.

[4] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery, in: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, Cambridge, MA, 1996, pp. 1–34.

[5] O.R. Za¨ıane, From resource discovery to knowledge discovery on the internet. Technical Report TR 1998–13, Simon Fraser University 1998.

[6] R. Kosala, H. Blockeel, Web mining research: A survey, SIGKDD Explorations 2 (1) (2000).

[7] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, Web usage mining: Discovery and application of usage patterns from Web data, SIGKDD Explorations 1 (2) (2000) 12–23.

[8] Biddulph (2003) Crawling the Semantic Web. BBC London, United Kingdom

[9] Zamir , O. Etzioni. Web Document Clustering: A Feasibility Demonstration, Department of Computer Science and

Engineering, University of Washington.

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

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

[12] K. Elissa, "Title of paper if known," unpublished.Baader, F, Calvanese, D, Giacomo, GD, Fillottrani, P, Franconi, E,Grau, BC, Horrocks, I, Kaplunova, A, Lembo, D, Lenzerini, M, Lutz,C, Moller, R, Parsia, B, Patel-Schneider, P, Rosati, R, Suntisrivaraporn,B, Tessaris, S, 2006. Formalisms for Representing Ontologies: State of

the Art Survey, TONES.

[13] B. Le Grand, M. Soto, Xml topic maps and semantic web mining,in: G. Stumme, A. Hotho, B. Berendt (Eds.), Semantic Web Mining,Freiburg, 3rd September, 2001, 12th Europ. Conf. on Machine Learning(ECML'01)/5th Europ. Conf. on Principles and Practice of Knowledge.

[14] A. Maedche, S. Staab, Ontology learning for the semantic Web, IEEE

Intelligent Syst. 16 (2) (2001) 72–79.

[15] A. Maedche, S. Staab, Ontology learning for the semantic Web, IEEE

Intelligent Syst. 16 (2) (2001) 72–79.

[16] G. Paaß, J. Kindermann, E. Leopold, Learning prototype ontologies by hierarchical latent semantic analysis. In [166], 2004, pp. 49–60.

A. Maedche, Ontology Learning for the Semantic Web, Kluwer, 2002.

A.B. Williams, C. Tsatsoulis, An instance-based approach for identifying candidate ontology relations within a multi-agent system. In Proceedings of the First Workshop on Ontology Learning OL'2000, Berlin, Germany, 2000. Fourteenth European Conference on Artificial Intelligence.