Semantic Web Mining: An Amalgamation for Knowledge Extraction

Karan Sukhija
Research Scholar,
Department of Computer Science and Application,
Panjab University, Chandigarh.

Abstract: *Semantic Web Mining is an emerging research area, aimed as amalgamation of two most rising arenas of research: the Web Mining and Semantic Web (SW). SW is an expansion of existing web where result knowledge is specified the distinct meaning. It enhances the web search. Web mining, as a mounting area of data mining, has three operations of interests in terms of data mining techniques– Clustering (i.e. find out the natural clustering between the pages of web, operators etc.),Association (i.e. the requested web addresses collectively inclined) and chronological scrutiny (i.e. the sort in which web address tendency to be salvaged). Semantic Web Mining purpose is to enhance the domino effect of Web Mining by exploring the novel-fangled semantic assemblies in the Web. It also makes usage of Web Mining for assembling up the Semantic Web.Both these arenas' distillate on the prevailing encounters of the World Wide Web: spinning amorphous data into machine-comprehensible data by means of Semantic Web tools.*

Keywords: *Web Mining, Semantic web, Ontology, Semantic web mining.*

I. INTRODUCTION

A. Web Mining

Web mining is an emerging trend of data mining that assists in extraction of valuable facts from web data (a range of web documents, hyperlinks among documents and usage logs etc). [1].Web mining is evolving rapidly with the help of novel techniques that are being developed by means of conventional and lenient computing tacticssimultaneously.Web mining has three operations of interests in terms of data mining techniques– clustering (e.g. discovering usualgrouping of web pages, users etc.), Association (e.g. the requested web addresses collectively inclined) and chronological analysis (e.g. the sort in which web address tendency to be salvaged). The basicprovinces of web mining are as follows:

- ***Content Mining:***It isconcerned with theanalysis of content [12] from various webresources. It retrievesthe findings of worthwhileinformation from the web contents (texts, images etc.), primarily based ontext mining techniques.

- ***Structure Mining:***Itconcentrates on theanalysisof the hyperlink structureamongweb pages. It [12] discovers the model that reveals thelink structures of the web and study thetopology of the hyperlinks.

- ***Usage Mining:*** It analysis the extraction of facts about web usage that the users clicks from Webserver logs. It also deals with reviewing [12] the data produced by the web surfer's session or behaviors.

Complete Web mining processes consist of four tasks [2] which are shown below:

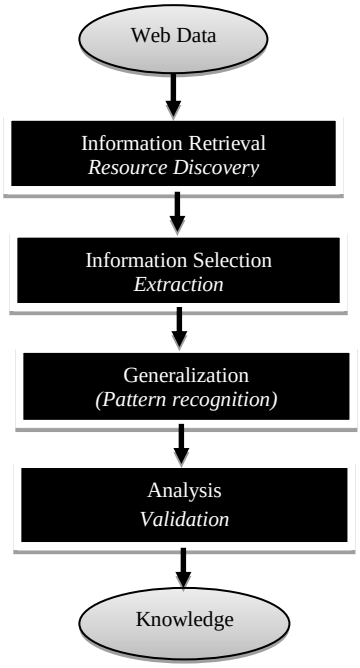


Figure 1: Web mining subtasks

1) Information Retrieval (IR)/ Resource Discovery: Resource discovery or IR deals with automatic retrieval of all relevant documents, while at the same time ensuring that the irrelevant ones are fetched as few as possible. The IR process mainly includes document representation, indexing, and searching for documents.

2) Information Selection/Extraction and Preprocessing: Once the documents have been retrieved the challenge is to automatically extract

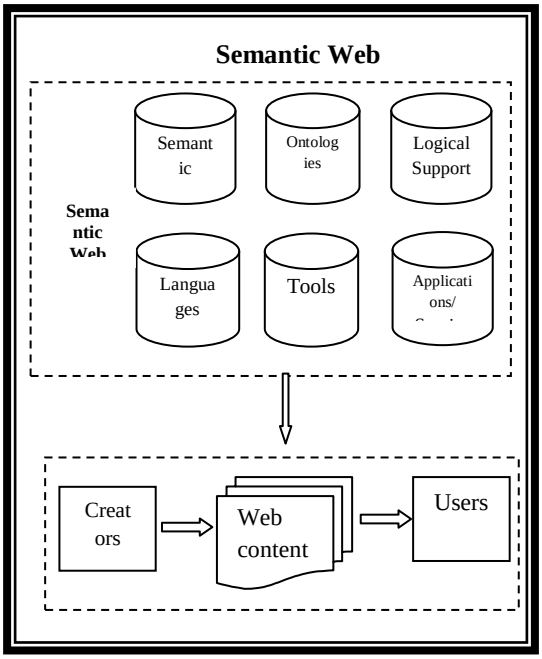
knowledge and other required information without human interaction. Information extraction (IE) is the task of identifying specific fragments of a single document that constitute its core semantic content.

3) Generalization: In this phase, pattern recognition and machine learning techniques are usually used on the extracted information.

4) Analysis: Analysis is a data-driven problem which presumes that there is sufficient data available so that potentially useful information can be extracted and analyzed. Humans play an important role in the information or knowledge discovery process on the web since the web is an interactive medium. This is especially important for validation and or interpretation of the mined patterns which take place in this phase.

B. Semantic Web

The current WWW has a huge amount of data that is often unstructured and usually only human understandable. The Semantic Web aims to address this problem by providing machine interpretable semantics to provide greater machine support for the user. Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in co-operation [3].



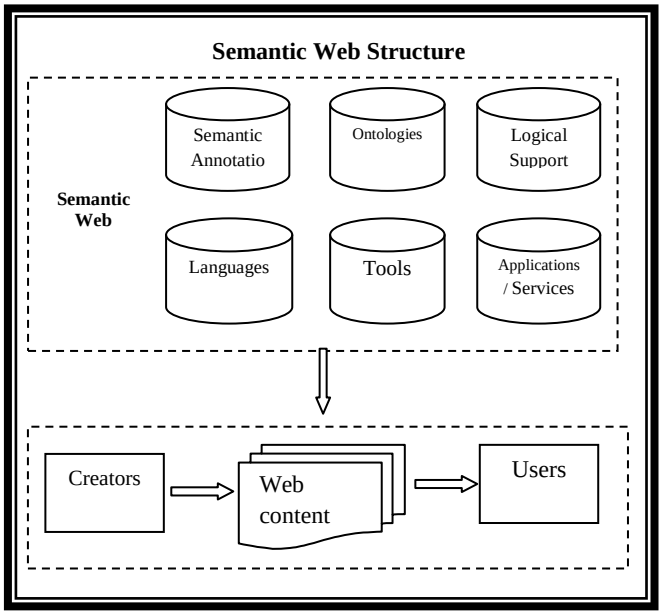


Figure 2: Semantic Web Future Trend [4]

The semantic web will provide intelligent access to heterogeneous, distributed information enabling software products to mediate between user needs and the information source available. Semantic Web:

- Provides a common syntax for machine understandable statements.
- Establishes common vocabularies.
- Agrees on a logical language.
- Uses the language for exchanging proofs.

The Semantic Web has a layer structure that defines the levels of abstraction applied to the web [5]. At the lowest level is the familiar World Wide Web, then progressing to XML, RDF [4], Ontology, Logic, Proof and Trust. The main tools that are currently being used in the Semantic Web are ontologies based on OWL (Web Ontology Language) and its associated reasons. Semantic Web is the advanced technique used by the web to fulfill the client's requirements. In Semantic Web, lot of information can be stored in RDF in an XML file format. This information can be extracted by the users depending upon their needs. This can be done by accepting the user request and providing response to the user by extracting the information from the RDF. Thus the information can be easily extracted by the Web.

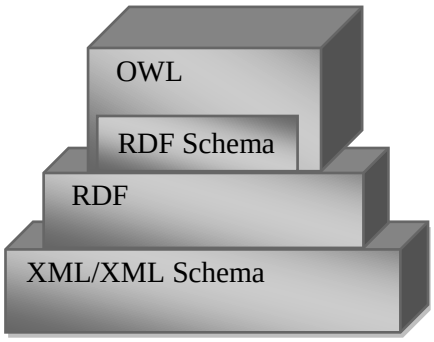


Figure 3: Semantic Web layered framework [6]

SW follows the layered framework, as shown in the figure 3, having the following layers as the basic ones [3]:

- an *XML layer* for representing the document structure and Web content;
- a *Resource Description Framework* (*RDF*) [4] layer for expressing the semantics/meaning of the content;
- an *Ontology layer* for describing the vocabulary of the domain;
- a *Logic layer* to enable intelligent reasoning with meaningful data

II SEMANTIC WEB MINING

The human ability for information processing is limited on the one hand, whilst otherwise the amount of available information of the Web increases exponentially, which leads to increasing information saturation[5]. In this context, it becomes more and more important to detect useful patterns in the Web, thus use it as a rich source for data mining [6]. The research area of Semantic Web Mining is aimed at combining two fast developing fields of research: the Semantic Web and Web Mining. The idea is to improve, on the one hand, the results of Web Mining by exploiting the new semantic structures in the Web; and to make use of Web Mining, on the other hand, for building up the Semantic Web. These two fields address the current challenges of the World Wide Web (WWW): turning unstructured data into machine-understandable data using Semantic Web tools. As the Semantic Web enhances the first generation of the WWW with formal semantics, it offers a good basis to enrich Web Mining: The types of (hyper)links are now described explicitly, allowing the knowledge engineer to gain deeper insights in Web structure mining; and the contents of the pages come along with a formal semantics, to apply mining techniques which require more structured input.

A) Semantic Web Content and Structure Mining

In the Semantic Web, content and structure are strongly intertwined. Therefore, the distinction between content and structure mining vanishes. However, the distribution of the semantic annotations may provide additional implicit knowledge. An important group of techniques which can easily be adapted to semantic Web content / structure mining are the approaches discussed as Relational Data Mining (formerly called Inductive Logic Programming (ILP)). Relational Data Mining looks for patterns that involve multiple relations in a relational database. It comprises techniques for Semantic Web Mining like classification, regression, clustering, and association analysis. It is quite straightforward to transform the algorithms so that they are able to deal with data described in RDF or by Ontologies [7]. There are two big scientific challenges in this attempt. The first is the size of the data to be processed (i.e. the scalability of the algorithms), and the second is the fact that the data are distributed over the Semantic Web, as there is no central database server. Scalability has always been a major concern for ILP algorithms. With the expected growth of the Semantic Web, this problem increases as well. Therefore, the performance of the mining algorithms has to be improved, e.g. by sampling. As for the problem of distributed data, it is a challenging research topic to develop algorithms which can perform the mining in a distributed manner, so that only (intermediate) results have to be transmitted and not whole datasets [8].

Semantic Web Usage Mining

Usage mining can also be enhanced further if the semantics are contained explicitly in the pages by referring to concepts of ontology. Semantic Web usage mining can for instance be performed on log files which register the user behavior in terms of ontology. A system for creating such semantic log files from a knowledge portal has been developed at the AIFB. These log files can then be mined, for instance to cluster users with similar interests in order to provide personalized views on the ontology [9].

III APPLICATION AREAS/ PROSPECTS

The future of Web Mining will to a large extent depend on developments of the SemanticWeb. The

role of Web technology still increases in industry, government, education,entertainment. Semantic web mining is applied in the E-Services areas of societal interest like E-Government, E-Politics and E-Democracy. Semantic web mining is also applied in genetics, molecules, social network analysis, as well as natural language processing [8].

- **E-commerce** The increased use of XML/RDF to describe products, services and businessprocesses increases the scope and power of Data Mining methods in e-commerce.Another direction is the use of text mining methods for modelling technical, socialand commercial developments. This requires advances in text mining and informationextraction.
- **E-learning** The Semantic Web provides a way of organizing teaching material, andusage mining can be applied to suggest teaching materials to a learner. This opensopportunities forWeb Mining. For example, a recommending approach [11] can be followed to find courses or teaching material for a learner. The material canthen be organized with clustering techniques, and ultimately be shared on the webagain, e. g., within a peer to peer network [12]. Web mining methods can be usedto construct a profile of user skills, competence or knowledge and of the effect ofinstruction.
- **E-government** Many activities in governments involve large collections of documents.Think of regulations, letters, announcements, reports. Managing access and availabilityof this amount of textual information can be greatly facilitated by a combinationof Semantic Web standardization and text mining tools. Many internalprocesses in government involve documents, both textual and structured.Web miningcreates the opportunity to analyze these governmental processes and to createmodels of the processes and the information involved.
- **E-science** In E-Science two main developments are visible. One is the use of text miningand Data Mining for information extraction to extract information from largecollections of textual documents.
- **Webmining for images and video and audio streams** So far, efforts in SemanticWebresearch have addressed mostly written documents. Recently this is broadened toinclude sound/voice and images. Images and parts of images are annotated withterms from ontologies.

IV CONCLUSION: Semantic Web is the emerging technology aiming at web-based information and services that would be understandable and reusable by both humans and machines. Semantic web uses ontology-based technologies and intelligent agents for semantic information processing. The combined area of Semantic Web Mining offers new techniquesto improve both areas. Semantics can improve the results of Web Miningby taking advantage of structures in the Web. Web Mining can improve theSemantic Web by finding new semantic structures to enrich the semantics. Theapplication of each area to the other creates a feedback loop, where the goal ofSemantic Web Mining is realized. This in turn will create a more usable Weband may help in transforming the Web into the Semantic Web.

V.REFERENCES

[1] Semantic Web Mining: State of the art and future directions, Stumme.G, Hotho. A, Berendt B, Web Semantics: Science, Services and Agents on the World Wide Web 4(2) (2006) 124 – 143 Semantic Grid – The Convergence of Technologies.

[2] Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, Sankar K. Pal, VarunTalwar, and PabitraMitra, IEEE transactions on neural networks, vol. 13, no. 5, september 2002.

[3] V. Kolovski ,J. Galletly,*"Towards E-Learning via the Semantic Web",*International Conference on Computer Systems and Technologies - CompSysTech'2003.

[4] H. W. Malik ,"Visual semantic web: ontology based E-learning management system", January 2009.

[5] Towards Knowledge Discovery in the Semantic Web, Krcmar H (2004), Informations management (German Edition). Springer, Berlin.

[6] Towards Semantic Web Mining, Berendt B, Hotho A, Stumme G (2002). ISWC 2002, First International Semantic Web Conference, Sardinia, Italy, June 9-12, 2002, Springer.

[7] Application based semantic web mining technique, Mahindra Pratap Singh Dohare*1 and Sanjaydeep Singh Lodhi, VinodMahor, Volume 2, No. 3, March 2011, JGRCS.

[8] A Roadmap for Web Mining: From Web to Semantic Web, Berendt, B., Hotho, A., Mladenic, D., van Someren, M., Spiliopoulou, M., Stumme, G. Web Mining: From Web to Semantic Web Volume 3209/2004 (2004) 1–22.

[9] Using Semantic Web Mining Technologies for Personalied ELearning Experiences, P. Markellou, I. Mousourouli, S. Spiros, and A. Tsakalidis (Greece),Proceeding (461) Web-based Education, 2005.

[10] Aguado, B., Merceron, A., Voisard, A.: Extracting information from structured exercises. In: Proceedings of the 4th International Conference on Information Technology Based Higher Education and Training ITHET03, Marrakech, Morocco. (2003).

[11] Tane, J., Schmitz, C., Stumme, G.: Semantic resource management for the web: An elearning application. In: Proc. 13th InternationalWorldWideWeb Conference (WWW 2004).

[12] K Sukhija, Web Content Mining equipped Natural Language Processing for handling web data, International Journal of Computer ApplicationsTechnology and Research, Volume 4–Issue3,209-213,2015.