Semantic Web Mining - A Review

Bhawandeep Kaur
Student of M.Tech
Department of Computer Science Engineering
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India

Simarjeet Kaur
Assistant Professor
Department of Computer Science Engineering
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India

ABSTRACT
An effective fetching of the most relevant documents from the web is difficult due to the vast amount of information in all types of formats. The vast amount of data is very difficult to understand by machines, but humans can easily understand. The semantic web is a web of data that make capable machines to understands the data on web pages and also known as Web 3.0. Semantic Web is a way to increase the accuracy of information retrieval systems. Web mining is the application of data mining to extract knowledge from web data using data mining techniques, including web documents, hyperlinks between documents, usage logs of web sites etc. The semantic web mining is aimed at combining both the semantic web and web mining. The main aim is turning unstructured data into machine understandable data using semantic web tools so machine can respond to human queries in less time and avoid tedious work and automatically extract knowledge hidden in the vast amounts of web data using web mining tools. This paper focuses on the various Semantic-web approaches and challenges.

General Terms
Semantic web, Web mining and semantic web approaches.

Keywords
Semantic Web Mining, Web 3.0.

1. INTRODUCTION
The two research areas Semantic Web and Web Mining both build on the success of the World Wide Web. They complement each other well because they each address one part of a new challenge posed by the great success of the current World wide web. Mostly data available on the Web pages are unstructured that they can only be understand by humans and difficult to understand by machines, but the amount of data is so vast that they can only be processed efficiently by machines [3]. The World Wide Web has become a new communication intermediate through with Web information entry. This consolidates with informational, cultural, social and evidential values to be specific. With the existence of various Search Engines like Google, Yahoo and many more, the users are tend to use them for retrieving their desired information from Web pages. But existing search engine cannot distinguish individual user's request. The semantic web addresses the first part of this challenge by trying to make the data machine-understandable, while web mining addresses the second part by automatically extracting the useful knowledge hidden in these data [3]. Web mining calls for creative use of data mining and/or text mining techniques and its distinctive approaches. Mining the web data is one of the most challenging tasks for the data mining. Semantic Web Mining aims at combining the two areas semantic web and web mining [3].

2. SEMANTIC WEB
The Semantic Web is based on a vision of Tim Berners-Lee, the inventor of the World Wide Web. In general, semantics is relating to pattern of sentence, what sentence actually means. If user change the pattern of sentence, then pattern changes but semantic remains the same. If a computer understand the semantics of a document, it does not just interpret the series of characters that make up that document, semantic web helps the machine to understand meanings behind the web page. The current WWW has a large amount of data that is often unstructured and usually only human understandable. For example, human brain uses logic:

1. Julie is a mother.
2. A mother is a parent.

Therefore, Julie must be a parent but machines cannot use these logic. The Semantic Web aims to address this problem by providing machine interpretable semantics to provide greater machine support for the user. If machine identify meaning behind the web page, then machine can easily help the users to access knowledge instead of unstructured material, allowing knowledge to be managed in an automatic way.

A simple search engine has not the capability to understand the relationships between keywords, phrases or parts of speech within a search phrase, but semantic search engine has therefore it allowing machine to understand the hidden meaning of the entire phrase by using ontologies [1]. Ontology is a precise explanation of terms and reasoning in a subject area. With use of ontologies computer can bring relevant outcomes with semantic things. By making the meaning so clear a machine can understand it or at least utilize it.

For example, a semantic search engine would be able to easily distinguish the differences between the following phrases made up of the same keywords but with different implications:

1. Semantic web mining by using java language.
2. By using java language semantic web mining.

In the example above, the sentences are made up of the same keywords, while the subject relationships are reversed. In traditional web search, which are based on ranking algorithms, since the relationships between the keywords or phrases are unknown, the engines would return identical or nearly identical results, even though it was being asked two completely different questions. Additional problems with traditional web search also arise when the keywords are too specific, producing few or no results or too general in which case the results are irrelevant [1].

A web of today is about documents but semantic web is about things on web pages means people, places, organizations and any concepts etc.

3. SEMANTIC WEB ARCHITECTURE

Berners-Lee suggested a layer structure for the Semantic Web. The development of the Semantic web proceeds in layers, one above another allowing for a more standardized way of developing.

3.1 Uniform resource identifier (URI)

URI is the foundation of web and holds the rest of the web together. The purpose of the URI is to unambiguously specify an identifier to represent a resource in a uniform way, identifying information representation constructs including classes, properties and individuals. As there is no ambiguity, it becomes possible to aggregate all data that refers to a given resource. URI is only a description not a location, only identifies. For example, one person could use abcd.com as his web page and it identifies person. So it associates the object abcd. Person can use his page but other users only see it. URIs provides users and software to know exactly what it is they are being referred to, they are globally unique and each occurrence of the same identifier means the same thing.

3.2 Unicode

Unicode is an encoding character sets and that allow all user languages can be used to read and write on the web by using standardized form.

3.3 Extensible markup language(XML)

XML is a language used to transport and store data on the web.XML is only to carry and describe data, not to display data. XML documents contain a user defined tags. XML schema is used to describe the structure of the XML document. XML namespace in semantic web is used to avoid conflict data or names. The aim of XML layer to provide the basic syntax and structure of the data on the web.

3.4 Resource description framework (RDF)

RDF is used for organizing information. RDF solves data linking problem**.** Resource is anything that has identity and these are identified by URI(uniform resource identifiers) . Description is really just a container holding several statements describing the resource. Framework is needed to make and understand statements. RDF allows the computer to process the question and come up with the answer.

3.5 Resource description framework schema (RDFS)

By using RDFs the classes and properties can be arranged in generalization/specialization hierarchies. The function of RDF and RDFS is to provide metadata to upper technologies placed on the layers on the top, in which that metadata can be exchanged and reused between these technologies or between these technologies and other applications.

3.6 Web ontology language (OWL)

Ontology is a exact explanation of terms and reasoning in a subject area. With ontology we can bring the semantic things. Description logics and web languages, the web ontology language was developed and can be used for defining ontologies. OWL satisfies the semantic web's requirements of providing minimal input from humans and supporting software requirements for a language with accurate meaning. Ontologies are helpful to clearly represent objects and also the relationship between them it may be direct or inverse relationship.

3.7 Rules, Proof & Trust layer

Proof layer is used to verify the results produced by the agents should be believed or authenticate the agent behavior.

Rules Layer is supposed to be used as a framework for making new conclusions how these conclusions should be expressed for the implementation of the semantic web.

Trust layer is to provide a mechanism for trust and confidence between Information sources and information users.

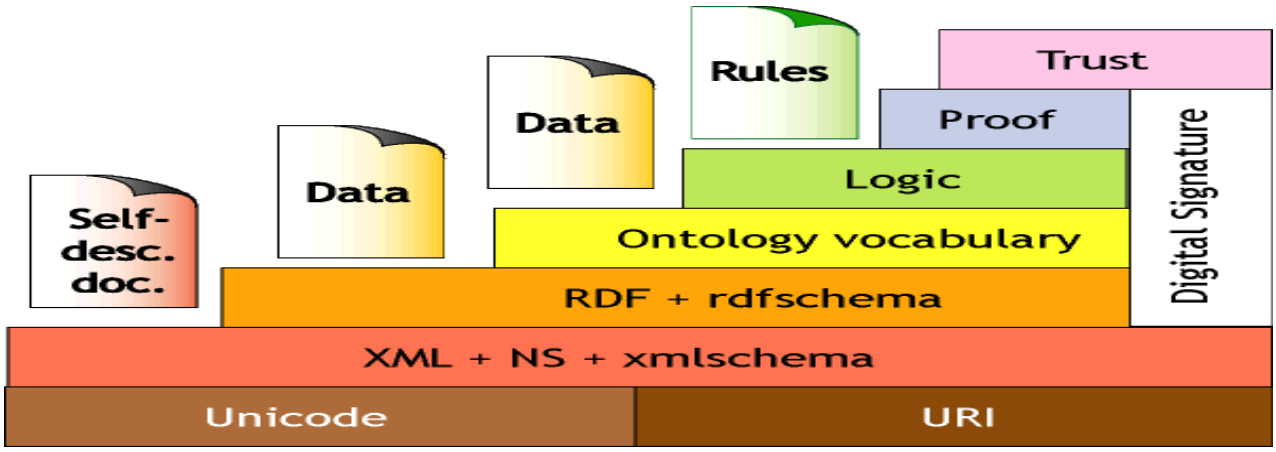

Figure 1 : Layer structure of semantic web [3]

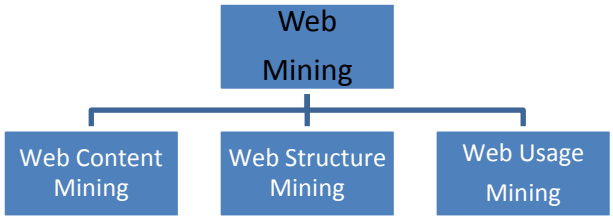

Figure 2 : Taxonomy of web mining [4]

4. WEB MINING

Web mining is the application of data mining techniques to fetch information from web data of web pages, including web documents, hyperlinks between documents, usage logs of web sites, etc [7]. It is thus the non trivial process of identifying valid, previously unknown and potentially useful patterns in the vast amount of these web 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, but it can also be applied to semi-structured or unstructured data. This means that web mining is helpful in the shift from human-understandable content to machine-understandable semantics.

4.1 Web content mining

Web content mining analyzes the content of web resources and process of extracting information from the contents of Web documents. It analyze content of the web pages as well and web searching. Mainly based on text mining techniques, but extensions to multimedia content is beginning to emerge in the research. The web content consists of several types of data such as textual, image, audio, video, metadata, as well as hyperlinks. Web content may be unstructured (plain text generally big data is closely associated with unstructured data), semi- structured (HTML documents), or structured (extracted from databases into dynamic Web pages)[7] .Big data is closely associated with unstructured data. Extremely large datasets that are difficult to analyze with traditional tools. So NLP techniques use for information retrieval.

4.2 Web structure mining

The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. Content is available everywhere what we need to insure that this content is relevance and quality is right then we use structure mining using page rank algorithm. This can be further divided into two kinds based on the kind of structure information used.

In addition, the content within a web page can also be organized in a tree structured format based on the various HTML and XML tags within the page.

4.2.1 Hyperlink

A hyperlink is a structural unit that links a location to a different location in web pages, either within the same web page or on a different web page. A hyperlink that links to a different part of the same page is called an intra-document hyperlink and a hyperlink that connects two different pages is called an inter-document hyperlink.

4.2.2 Document structure

Web mining can be broadly divided into three categories, according to the kinds of data to be mined. 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. The classification of web mining techniques represented in above Fig 2.

4.3 Web usage mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data or server log, in order to understand and better serve the needs of web-based applications and analyzes the users clicks from web server logs. Usage data captures the identity or origin of web users along with their browsing behavior at a web site.

Key feature of Web usage mining technique includes directly fixing the problems associated with web log mining. It is all about identifying user browsing patterns over the world wide web, with the help of knowledge acquired from web logs.

5. SEMANTIC WEB APPROACHES

5.1 Ontology approach

Ontology is the backbone of the semantic web. Ontology is a collection of URIs with a usually informally explain meaning. Ontologies are represented by a formal ontology language. Ontology is being represented as a set of concepts and their inter-relationships related to some knowledge domain. The knowledge provided by ontology is very useful in defining the structure and scope for mining web content [13]. Ontology is defined as a set of objects, concepts and other entities that are exist in some area and the relationships that occur them[13]. Fetching an ontology from the web is a challenging task. It builds on techniques from Web content mining and it combines machine learning techniques with methods from fields like information retrieval, applying them to discover the semantics in the data and to make them clear. The techniques produce intermediate results which must finally be integrated in a machine understandable format.

5.2 Semantic based web mining

Semantic-based web mining is a combination of two fast developing domains semantic web and web mining. It can be read as (Semantic Web) Mining and Semantic (Web Mining). Semantic web addresses the challenge by trying to make the data for both machine and user understandable. While web mining addresses the automatically extracting the useful knowledge or information from the huge amount of data on web pages. In semantic based web mining the web pages are mined by the machine can perform better understand the information on the web pages [13]. It is basically fetching XML and RDF documents along with ontologies and metadata. Semantic based web mining includes mining the data sources and information relating to the information management technologies on the web. Semantic web mining will develop from web mining. The

goal of semantic based web mining is to make easy use of the web. Semantic web requirement are considered in three major groups ontology, semantic web content and web service [13].

6. SEMANTIC WEB CHALLENGES

(1) Vastness of available data: The web contains huge amount of data on billions of web pages and existing technology has not yet been able to estimate all semantically duplicated terms.

(2) Unclearness: These are estimated concepts like young or tall. This arises from the unclearness of user queries, of matching queries with provider contents and of trying to combine different knowledge bases with overlapping but carefully different concepts.

(3) Changeableness of terms: These are precise concepts with changeable values. For example a teacher might present a set of rules for test which correspond to a number of different distinct capability of students each with a different probability.

(4) Inconsistency of ontologies: These are logical discrepancies which will surely arise during the development of large ontologies and when ontologies from separate sources are combined.

7. CONCLUSION

In this paper we have studied the two fast developing research areas are: web mining and semantic web. The combined area of Semantic Web Mining offers new techniques to improve both areas. Semantic based web mining can improve the results of Web Mining by using the new semantic structures in the Web. The Semantic Web can make mining of the Web much easier because of the availability of background knowledge and Web Mining can also construct new semantic structures in the Web. The resulting research benefits many areas of industry such as e activities, health care, privacy and security and search engines, knowledge management and information retrieval.

8. REFERENCES

[1] K. Sridevi and R. Umarani , An Ontology ranking algorithms on semantic web, International Journal of Advanced Research in Computer and Communication Engineering ,Vol. 2, Issue 9, September 2013.

[2] I. Rasekhet , Dynamic Search Optimization for Semantic Webs Using Imperialistic Competitive Algorithm., IEEE workshop, 2012 .

[3] G.Stumme, A.Hotho and B.Berendt, Semantic Web Mining State of the art and future directions, Web Semantics: Science, Services and Agents on the World Wide Web, February 2006.

[4] Sandhya and M. chaturvedi, A survey on web mining algorithms, International Journal of Engineering and Science, Vol. 2, Issue 3, 2013.

[5] S. Kaur and U. Kaur, An Optimizing Technique for Weighted Page Rank with K-Means Clustering ,International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7, July 2013 .

[6] T. Fischer and J. Ruhland ,Towards Knowledge Discovery in the Semantic Web, MKWI – Business Intelligence, 2010.

[7] H. Hassanzadeh and M. R. Keyvanpour , Semantic Web Requirements through Web Mining Techniques, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 7, August 2012 .

[8] M. Thakur and G. S. Pandey, Performance Based Novel Techniques for Semantic Web Mining, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 9, Issue 1, January 2012.

[9] M.P.S. Dohare and S.S. Lodhi and V. Mahor, Application based semantic web mining technique, Journal of Global Research in Computer Science, Vol. 2,No. 3, March 2011.

[10] S. F. Shazmeen and E. Ramyasree, Semantic Web Mining: Benefits, Challenges and Opportunities , International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 7, July 2013.

[11] R. Rani and V. Jain , An Weighted Page Rank using the Rank Improvement, International Journal of Advanced Research in Computer and Communication Engineering, July 2013.

[12] A. Shafeeq and Hareesha , Dynamic Clustering of Data with Modified K-means Algorithm, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 27, 2012.

[13] Sivakumar and Ravichandran , A Review on Semantic-Based Web Mining and its Applications, International Journal of Engineering and Technology, Vol. 5, Feb-Mar 2013.