ISSN: 2277-9655
Impact Factor: 4.116

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

SEMANTIC WEB MINING: ISSUES AND CHALLENGES

Karan Singh*, Anil kumar, Arun Kumar Yadav
* Research Scholar, Department of Computer Science & Applications K.U., Kurukshetra, Haryana, INDIA
Assistant Professor, Department of Comp. Sci., University College, K.U., Kurukshetra, Haryana, INDIA
Assistant Professor, Department of Comp. Sci., University College, K.U., Kurukshetra, Haryana, INDIA

DOI: 10.5281/zenodo.60109

ABSTRACT

The combination of the two fast evolving scientific research areas "Semantic Web" and "Web Mining" are well-known as "Semantic Web Mining" in computer science. These two areas cover way for the mining of related and meaningful information from the web, by this means giving growth to the term "Semantic Web Mining". The "Semantic Web" makes mining easy and "Web Mining" can construct new structure of Web. Web Mining applies Data Mining technique on web content, Structure and Usage. This paper gives an overview of where these two areas work together, the way how this integration of both (Web Mining & Semantic Web) gives maximum profitable outcomes on WWW (World Wide Web) and challenges in this area.

KEYWORDS: Challenges, RDF, Semantic Web, Web Mining.

INTRODUCTION

In a scattered informational system, documents and objects have been joined together for suitable collaborative access like on WWW, Where hyperlinks and URL addresses is utilized for finding the required information. Almost a million pages information are uploaded every day on WWW, a pre-existing page changes after some days, and some hundreds of gigabytes change every month.

But in the meantime adding more information on the WWW, gain access to this information is getting harder than before. The leading problem of this information access is because of unstructured or semi-structured web contents available on Web. So it is tough to construction, normalize and bring together them [1]. As the volume of data increases it is difficult to be managed this data on Web. In large volume data base it is almost impossible to extract relative solutions from WWW. To tackle this challenge "Web mining" area came into action. Web mining uses data mining techniques to discover and extract information automatically from documents and web services.

The "Semantic Web" is next version of the existing web with well-defined meaning, and with altering web contents into machine understandable form, would help in excellence and intelligence of the web [2] [3]. The level of these data must be upgraded and their embedded knowledge must be extracted before using in Semantic Web. In this article focus is on different applications of web mining methods in semantic web field and generally the mixture of these two important research fields.

WEB MINING

Web is a collection of hyperlinked documents on one or more Web servers [4]. Web mining term is defined as "quarrying (mining) of the World Wide Web to extract useful knowledge and data about user query, content and structure of the web". Web Mining helps users to find "models" or "patterns" within one page or among several pages over the WWW. These patterns finding queries (finding relative information over Web) can be applied on well-formed, semi-structured or unstructured data. Web Mining helps to transform human readable content to machine readable semantics over WWW.

The general development of Web Mining can be further divided into four subtasks (see in Figure 1): 1) Resource outcome: automatically extracting relevant documents on Web. 2) Information collection/Extraction and preprocessing: Select information and pre-process it into desired format. 3) Generalization/Machine Learning: automatically discovers general patterns at individual Web sites as well as across multiple sites. 4) Analysis: here Knowledge Discovery is proposed by human interaction. Validation and interpretation of the mined patterns are analyzed [5].

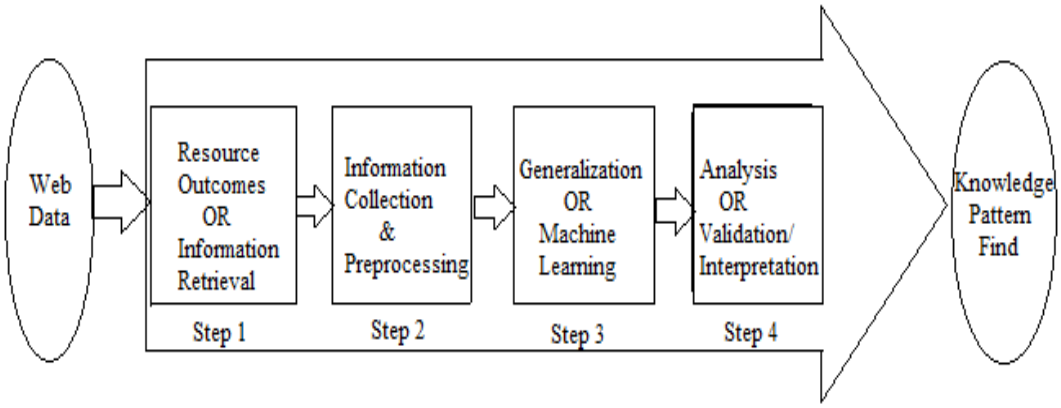

Figure 1: Web Mining Subtasks

Based on which part of Web is important for Mining Purpose, Web mining can be characterized to three parts (see in Figure 2): 1) Web-Content Mining, 2) Web-Structure Mining, and 3) Web-Usage Mining [6].

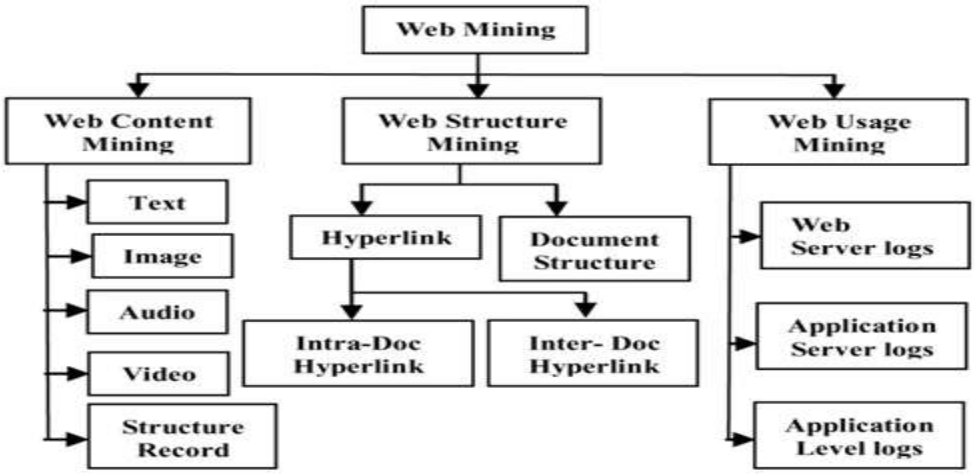

Figure 2: Three Areas of Web Mining

2.1) Web-Content Mining: Web content mining aims to extract useful information or knowledge from web page in the form of textual information. Web content mining is different from "data mining" because Web data are mainly semi-organized or unorganized, while data mining contracts mostly with organized data. The web content involves text data, image, audio, video, metadata and hyperlinks. Most stress is given on text and hypertext data in content mining. In web-content mining text documents are unstructured, HTML documents are semi-structured and Database tables

ISSN: 2277-9655
Impact Factor: 4.116

are more structured [7]. The goal of Web-content mining is primarily to support or to improve information discovery or cleaning the information.

2.2) Web-Structure Mining: This mining operates on the hyperlink structure of Web pages. Web-Structure Mining is the method of quarrying structure information on the hyperlink structure of Web and used to improve the structure of the web pages. This model can be used to classify web pages and is useful to produce information such as the similarity and association between dissimilar websites. The "Web-Content Mining" tries to discover the "intra-document structure" (structure within a document), "Web-Structure Mining" finds the "inter-document structure" (within the web itself) [8]. This Mining process is centered on the topology of the hyperlinks with or without an explanation of the links and can be implemented to classify Web pages and is useful for producing information such as the similarity relationship between Web sites.

2.3) Web-Usage Mining: A primary web resource in Web-usage mining is, "The collection of requests made by users to a particular websites, which is stored on Web server logs [9]". In Web-usage mining extracting relevant information is sensed from server log and is done on the behalf of user's history, proxy server logs, browser logs, user profiles, registration data, user sessions, cookies, web user behavior, user queries, bookmark data, mouse clicks and scrolls etc. The server logs can be scanned by client viewpoint or by server viewpoint. Web usage mining process can be divided into three independent tasks: Preprocessing, pattern discovery and pattern analysis.

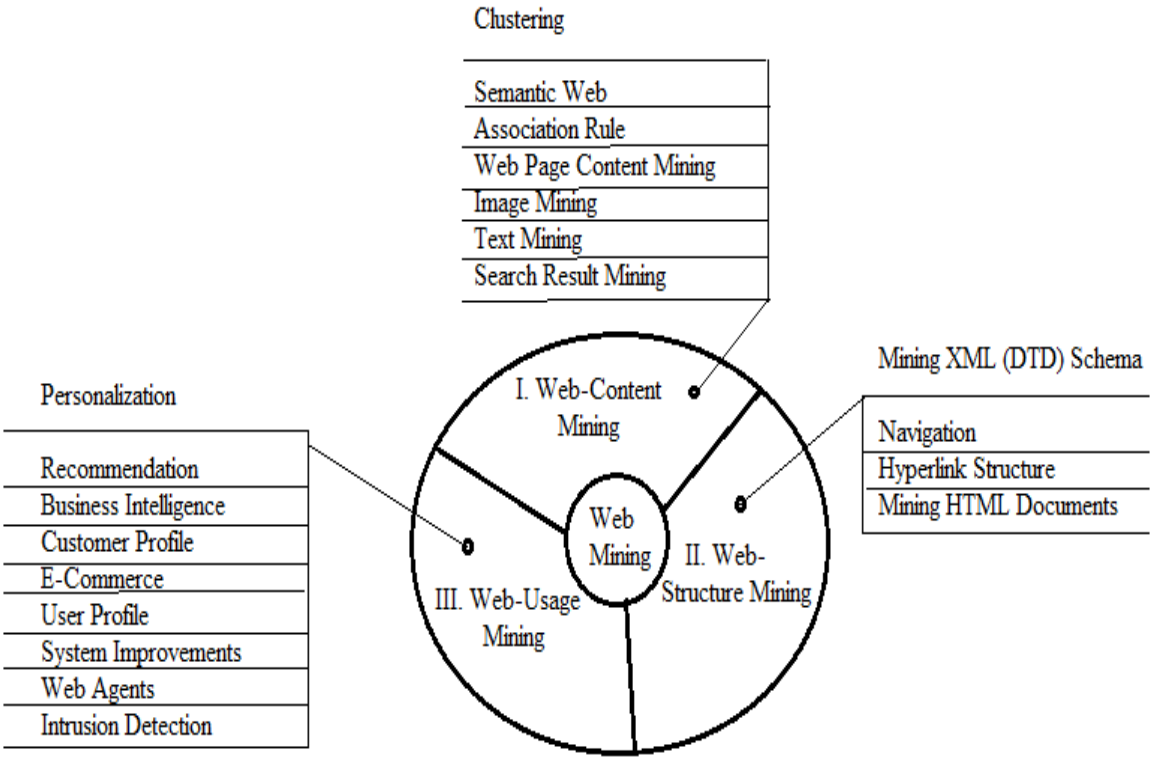

Figure 3: Web-Mining Techniques and Applications.

Web mining techniques and applications are shown in Figure 3. Web mining normally covers the techniques of refining search or customization by (i) learning interests of users based on access patterns, (ii) providing users with pages, sites, and announcements of concern, and (iii) using XML to expand search and information discovery on the Web [10].

ISSN: 2277-9655
Impact Factor: 4.116

SEMANTIC WEB

Semantic Web is a technique for satisfying the web users' requests. Semantic web is a way in which user query is sensed by machine and relative answer is replied back to users corresponding to their query [11] [12]. Machine processable information can point the search engine to the relevant pages and in this way can improve both correctness and memory. Semantic Web is a related to Web2.0 (second generation Web) and credit of its development goes to Sir Tim Berners Lee founder of WWW [12] [13]. Semantic Web symbolizes the extension of the WWW. World Wide Web gives ability to share their data outside all the hidden barriers and the limitation of programs and websites using the meaning of the web. These steps are followed in support of Semantic Web:

I) Common syntax for machine understandable format.
II) Launching common languages.
III) Agreeing on a logical language.
IV) Using the language for substituting proofs.

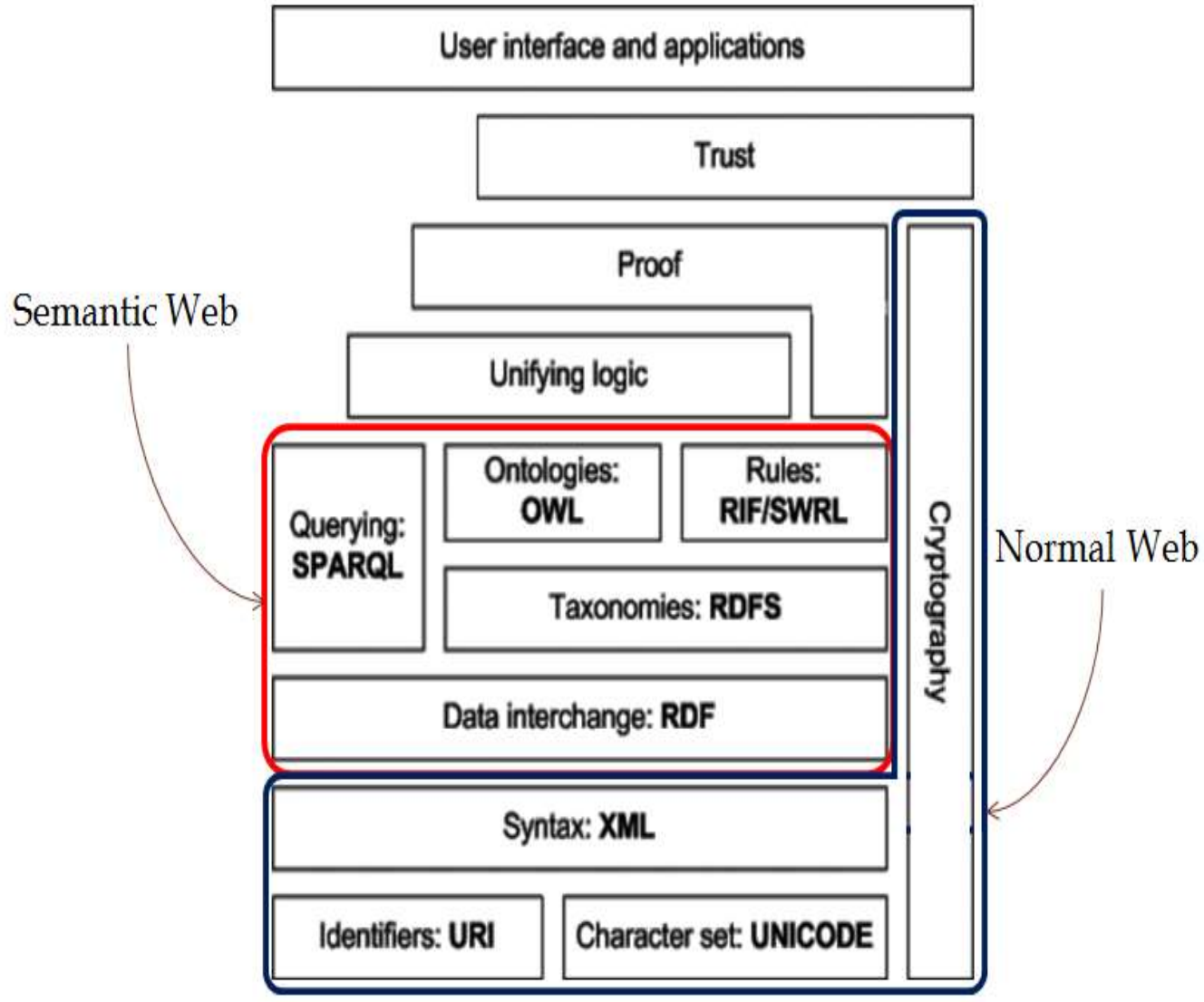

Figure 4: Semantic Web Layered Architecture [14].

ISSN: 2277-9655
Impact Factor: 4.116

A layered architecture is proposed by "Sir Tim Berners Lee" in support of steps involved in semantic web development (four steps mentioned above) and defines the levels of abstraction applied to the Web (see in Figure 4). Whole outlook of semantic web is shown on the behalf of applications, technology development and concepts (Figure 5).

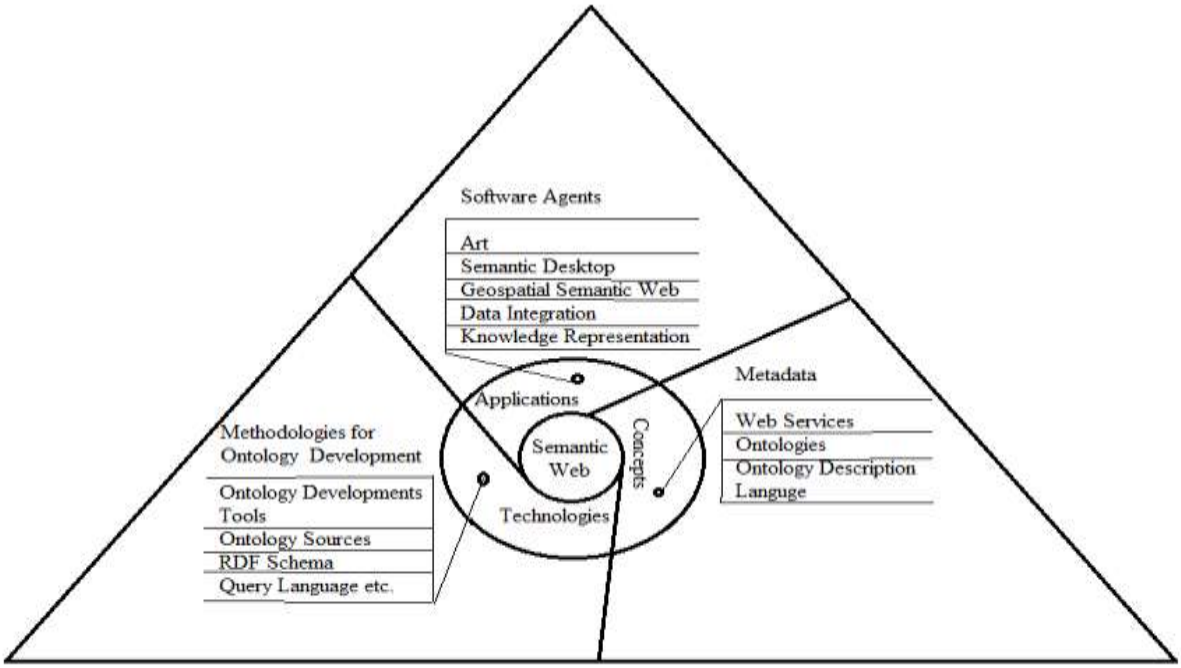

Figure 5: Semantic Web Perspectives.

3.1) Layered Architecture View of Semantic Web : The layered architecture of the Semantic Web was recommended by "Sir Tim Berners Lee" in support of four much needed steps to develop "Semantic Web" which are mentioned above . This layered design is discussed in detail in [15] and [16], which also talk recent research questions in Semantic Web development methods and challenges. Various elements like at the lowest level is the familiar World Wide Web (URI, UNICODE), then progressing to XML, RDF, Ontology, Logic, Proof and Trust are discussed here in support of proposed architecture (Figure 4).

A) URI and UNICODE: A URI (Uniform Resource Identifier) is a uniquely arranged string that identifies any entity resource. Every entity over WWW is referred via URI. A URI can be further categorized as a locator, a name, or combination of both. UNICODE is a standard for exchanging symbols.

B) XML (Extensible Markup Language): The XML technique is a well-known procedure to store, exchange, organize, and retrieve data on/from the web. XML enables its users to create their own tags and allows them to outline their content simply. In this way, with the help of XML semantic interactions towards data can be maintained [17] [18].

C) Resource Description Framework (RDF): RDF is a structure that empowers the encoding, exchange and reuse of organized metadata. RDF is an application of XML that provide unambiguous methods for expressing semantics. Information becomes machine recognizable format on RDF layer. This is RDF where foundation for processing metadata starts and RDF helps in exchanging machine-understandable information on web. The RDF data model is consists of three types of things (see in Figure 6): (I) Resource: Web pages or any other real-world entity which is not belongs directly to World Wide Web, these are always identified by URIs, (II) Properties: are precise features, characteristics, or relations describing resources, (III) Statement: A resource along with a property having a value for that resource form an RDF statement [19].

ISSN: 2277-9655
Impact Factor: 4.116

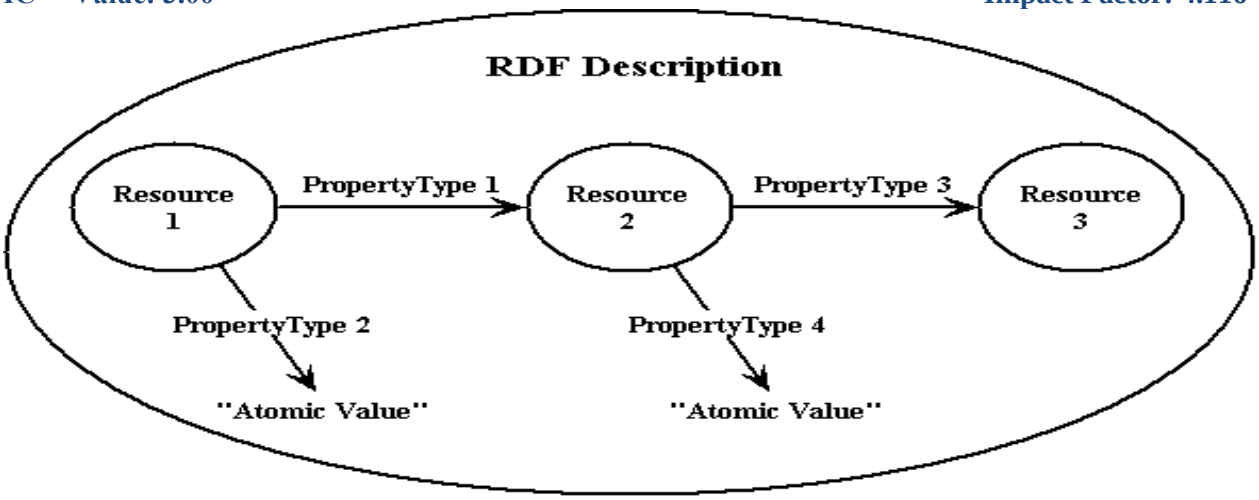

Figure 6: The RDF Data Model [20].

D) Ontology: Ontology is collection of Unicode Resource Identification. Ontology is vocabulary of Semantic web. The OWL (Web Ontology Language) is a more difficult language with enhanced machine interaction and understandable than RDF. Ontology is an agreed vocabulary that provides a set of well-founded constructs to build meaningful higher level knowledge for specifying the semantics of terminology systems in a well-defined and unambiguous manner.
E) Metadata: Metadata are data about data. Metadata help to index Web pages and Web sites in the Semantic Web, permitting other systems to recognize what the Web page is about.
F) Web Services: Semantic Web Service is server end of a client–server system for machine-to-machine interaction via WWW. Semantic services are a component of the semantic Web because they use markup which makes data machine-readable in a detailed and sophisticated way. This is also termed as a software system designed to support interoperable machine-to-machine interaction over a network.
G) Software Agents: Software agents intelligently running around the web and performing complex actions for their users. Agents communicate using an Agent Communication Language (ACL). An Agent is the important actor on an Agent Platform which combines one or more service capabilities, as published in a service description, into a unified and integrated execution model. S/W agent takes input from related environment and submits its related intellectual outputs to environment.

CHALLENGES IN SEMANTIC WEB MINING

Some most important challenges are identified in "Semantic Web Mining" and these are as follows:
(A) The availability of relevant content,
(B) Ontology availability, growth and advancement,
(C) Multiplicity of languages,
(D) Scalability,
(E) Visualization to reduce information overload,
(F) Stability of Semantic Web language,
(G) Ensuring user privacy, and
(H) Interaction among qualitative queries of humans with the machine generated quantitative responses.
For taking fruitful results from "Semantic Web" some challenging areas should be treatment very well which are listed above [21] [22]. Its success, failure and future scopes are also based on these challenges. In reference to "Semantic Web" some are discussed here for better understanding these challenges. After understanding these challenges one can find a way to achieve a better development path for Web Mining, analysis on web, ensure privacy and performance of web.
(A) The availability of relevant content: All the existing web content should be upgraded after some regular time intervals which included all databases, HTML static contents, XML contents , multimedia and dynamic contents for better availability of relevant information.

(B) Ontology availability, growth and advancement: Ontology is most important part for develop web services. Efforts should be made on worldwide common ontologies for the Semantic Web. An easy secure and widely common ontology should be preferred for Semantic Web development and evolution.
(C) Multiplicity of languages: Different languages apply to different situations. To avoid dependency over content provider Semantic Web methodology should offer services to access information in numerous languages. Figure 7 displays various languages and platforms that could be used in the semantic web. It is important that application designers are guided in the choice of language that is most appropriate to their task.

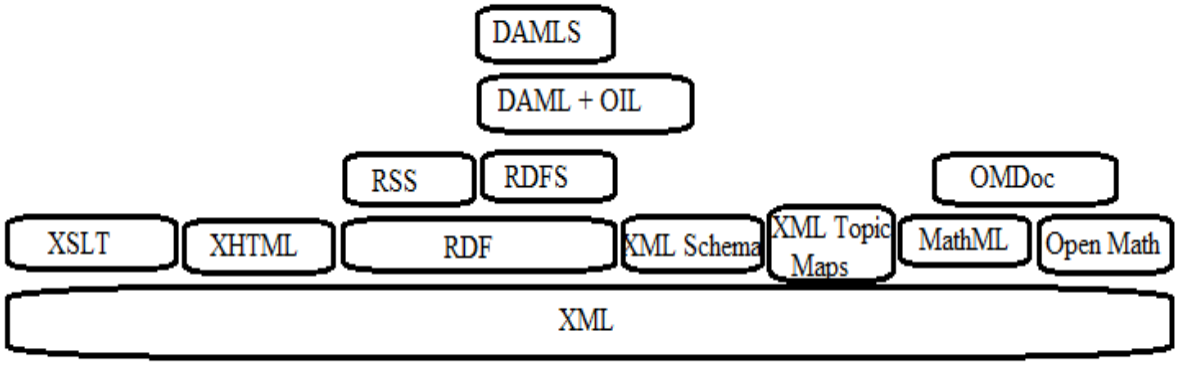

Figure 7: Available Set of Language for Implementing Semantic Web.

(D) Scalability: A significant effort must be made on how to organize Semantic Web content in scalable manner? Store it and provide the necessary mechanisms to find it.
(E) Visualization to reduce information overload: Whenever user demand the easy recognition of relevant content for their query then Semantic Web contents must be explored that differ from the usual hypertext structure visualization of the current web.
(F) Stability of Semantic Web language: In order to allow the creation of the necessary technology that supports the Semantic Web should be motivated.
(G) Ensuring user privacy: Semantic Web collects user's likes, dislikes, history of his/her query, his usage patterns and personal information about his nature over web. This information about user will help in sensing future searching and information related to him. Hence this is important to implement access rights mechanisms that can ensure the desired level of privacy for user data distributed across multiple products over Web.
(H) Interaction among qualitative queries of humans with the machine generated quantitative responses: This is important in semantic web that queries generated by users should map on machine generated responses. Interaction in this way may be said as "IQ of human (sentence generated for query) should be understandable by Semantic Web query accepter tool for relevant response to user query".

CONCLUSION

Under "Semantic Web Mining" two fast growing scientific research areas "Semantic Web" and "Web Mining" are discussed in this article. After brief introduction firstly "Web Mining", subtasks included in Web Mining and after that three classes of Web Mining (Web-Content Mining, Web-Structure Mining and Web-Usage Mining) are also discussed. Then basic discussions about "Semantic Web", its perspectives, essential components of layered architecture of "Semantic Web" are also elaborated. At last in the way of Semantic Web Mining some challenges are also addressed which should be followed for easy implementation and worldwide use. Note that, due to many challenges in area of Semantic Web Mining and its wide area of research domain, it is concluded that more researchers are needed specially related to web services.

REFERENCES

[1] T. Berners-Lee, N. Shadbolt and W. Hall, "The Semantic Web Revisited", IEEE Intelligent Systems, PP 96-101, 2006.
[2] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Second ed. The Morgan Kaufmann Series in Data Management Systems: Morgan Kaufmann Publishers, 2006.
[3] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web", Scientific American, 2001.
[4] V. Sitha Ramulu, Ch. N. Santhosh Kumar, K. Sudheer Reddy, "A Study of Semantic Web Mining: Integrating Domain Knowledge into Web Mining", International Journal of Soft Computing and Engineering (IJSCE), vol. 2, Issue 3.
[5] Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, Sankar K. Pal, Varun Talwar, and Pabitra Mitra, IEEE transactions on neural networks, vol. 13, no. 5, September 2002.
[6] Bettina Berendt, Andreas Hotho and Gerd Stumme, "Towards Semantic Web Mining".
[7] Faustina Johnson and Santosh kumar gupta, Web Content Mining Techniques: A survey, International Journal of Computer Applications, June 2012.
[8] Miguel Gomes da Costa Junior and Zhiguo Gong, "Web Structure Mining: An Introduction", Proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3, 2005, Hong Kong and Macau, China.
[9] J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan, "Web usage mining: discovery and application of usage patterns from web data", SIGKDD Explorations, 1(2):12(23, 2000).
[10] S. Mitra and T. Acharya, "Data Mining: Multimedia, Soft Computing and Biometric: John Welly and Sons", 2003.
[11] O. Mustapaşa, A. Karahoca, D. Karahoca and H. Uzunboylu, "Hello World, Web Mining for E-Learning", Procedia Computer Science, Vol. 3, No. 2, pp. 1381-1387, 2011.
[12] D. Jeon and W. Kim, "Development of Semantic Decision Tree," Proceedings of the 3rd International Conference on Data Mining and Intelligent Information Technology Applications, Macau, 24-26 October 2011, pp. 28-34.
[13] V. Sugumaran and J. A. Gulla, "Applied Semantic Web Technologies," Taylor & Francis Group, Boca Raton, 2012.
[14] http://semanticsage.blogspot.in/2013/03/the-semantic-web-architecture.html
[15] P. Patel-Schneider and D. Fensel, "Layering the semantic web: Problems and directions", in [82], pages 16–29, 2002.
[16] P. Patel-Schneider and J. Simeon, "Building the semantic web on XML", in [82], pages 147–161, 2002.
[17] J. Domingue, D. Fensel and J. A. Hendler, "Handbook of Semantic Web Technologies," Springer-Verlag, Heidelberg, 2011.
[18] A. Jain, I. Khan and B. Verma, "Secure and Intelligent Decision Making in Semantic Web Mining," International Journal of Computer Applications, Vol. 15, No. 7, pp. 14-18, 2011.
[19] V. A. Chakkarwar and Amruta A. Joshi, "Semantic Web Mining using RDF Data", International Journal of Computer Applications (0975 – 8887) Volume 133 – No.10, January 2016.
[20] http://www.dlib.org/dlib/may98/miller/05miller.html.
[21] V. Richard Benjamins, Jesús Contreras, Oscar Corcho and Asunción Gómez-Pérez, "Six Challenges for the Semantic Web".
[22] David Corsar, Peter Edwards, Nagendra Velaga, John Nelson and Jeff Pan, "Short Paper: Addressing the Challenges of Semantic Citizen-Sensing".