Semantic Web Mining and Semantic Search Engine: A Review

Nitika Uppal[1], Shivani Rana[2], Avni Sharma[3]

[1]Student, Department of Computer Science, HPTU University, Himachal Pradesh, India
[2,3]Assistant Professor, Department of Computer Science, HPTU University, Himachal Pradesh, India

---***---

Abstract – *Web 2.0 is the current web which we are using. This current web can only read the data simply but cannot understand its meaning. Semantic web is extended version of current web. Semantic web is the way to put the data on the web in a form that a machine can easily understand. It is the way to retrieve exact and accurate data from the web. Extracting exact and useful data from semantic web is called as semantic web mining. Main aim of semantic web mining is to combine both the semantic web and web mining. And for the retrieval of the data from the web search engines are required. Semantic search engines provide the facility to retrieve more meaningful data from the web. This paper provides a brief overview about the semantic web, semantic web mining and semantic search engines. In this paper we will also discuss the challenges for semantic search engines upon which further research can be done.*

***Key Words*: Semantic Web, Semantic Web mining, Semantic search engines, RDF, Ontology**.

1. INTRODUCTION

In computer science Mining is the process of extracting Useful knowledge or information from a huge amount of data. Mining can be of three types, Data mining, web mining and Text mining. The data for mining can be available either in structured or unstructured form. Data mining mainly deals with structured data organized in a database while text mining mainly handles unstructured data/text [1]. Web mining deals with both structured and unstructured data.

The Semantic Web was thought up by Tim Berners-Lee, inventor of the WWW, URIs, HTTP, and HTML [2]. Semantic web is the extended version of the current web which is also known as web 2.0 and its enhanced version is known as web 3.0(semantic web). Semantic web mining is the combination of two words as, semantic web and web mining i.e. the process of mining the semantic web. Semantic Web provides those features which are not included in web 2.0. In Web 2.0 machine can read the words or contents provided in user query but cannot understand their meaning. Our Web contains unstructured form of data, which can be understood by human but not by the machines. To make this data machine understandable we need the concept of semantic web mining, where semantic web makes data machine understandable and web mining is used to extract useful patterns or knowledge from this data. Semantic web information can support data integration, data discovery, navigation and automation of tasks [3].

To answer the user query there are so many search engines like Google, yahoo etc. But these search engines are not so efficient that they can provide exact answer of user query. They deliver the answer linguistically correct but in larger amount. In the present era, it takes too much time to search for the exact data from different links so new generations prefer semantic search engines because efficient searching is needed to get a well-defined and quality result [4].Semantic search engines provide information in precise manner with well define impression. A semantic Search engine uses various semantic technologies like RDF, Ontology etc.

The purpose of the paper is to provide an overview of semantic web mining and semantic search engines. The framework of semantic web mining and various semantic search engines with their different search approaches are discussed in this paper.

2. SEMANTIC WEB

The concept of semantic web was introduced by Tim Berne Lee in 2000. Semantic web is an extension of the current web in which information is given well defined meaning, enabling computers and people to work in cooperation efficiently [5].The word semantic means study of meaning in a language, and main goal of semantic web is to make this meaning understood by machine. When user provides a query, the machine interprets the words in the query without understanding their meaning. If user changes the pattern of his query then pattern will change but the semantic will remain same. It means by changing the structure of the sentence, its meaning is not changed.
For example, there are logical sentences as –

1. Arun is a father.
2. A father is a parent.

Therefore Arun is also a parent, but the given logic is grasping by human not by a machine. Semantic web can address this problem by using some rules and techniques due to which a machine can understand the semantics and it will be able to provide exact result of user query. The effort behind the Semantic Web is to add semantic annotation to Web documents in order to access knowledge instead of unstructured material, allowing knowledge to be managed in an automatic way [6].

2.1 Architecture of semantic web

The main objective of Semantic web is to make the current web more intelligent by altering its structure of data on the current web so that this data can be easily linked, processed and delivered to the users.

The data on the current web is in unstructured (images/ text /HTML) or in semi structured (XML) form. But the semantic web is used to convert this data into structured form. For this it make use of common data formats and exchange protocols on the web, and most commonly the Resource Description Framework (RDF), which is based upon Web ontology language (OWL).

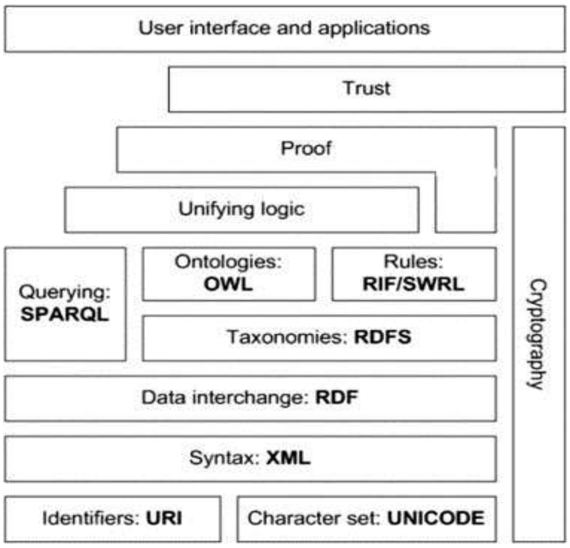

Fig-1: Architecture of Semantic Web; Source-W3C

The architecture of semantic web is composed of many layers which provide a more standardized way of development.

1) **Unicode**: It belongs to the bottom layer which is the World Wide Web layer of the architecture. It is an international standard of encoding, which is used to assign a unique numeric value to each and every character of a language or program that is independent of the underlying platform. Before this standard, different encoding standards were used by web which may cause conflicts during standardization. But now, Unicode solved this problem by providing a standard format of encoding.

2) **URI:** URI stands for Uniform Resource identifier or also called as Uniform Resource locator. In order to identify a resource uniformly on the web, a formatted string is used, which is known as URI. URI provide a standard way to access the resource by other machine across the network or over the World Wide Web.URI is similar to URL in order to specify the location of a file. Both are used interchangeably. The complete form of URL is referred as absolute URI.

3) **XML:** XML stands for extensive markup language, much like HTML. It is a W3C recommendation. It is a self descriptive language which is a public standard. It allows us to store the data rather than how it will be presented.

4) **RDF:** It is an acronym for Resource Description Framework. It is used to provide a standard framework for data interchange and describe the resources on the web. There is a need of some conventional tools for dealing with data and the relationship between data on the web, and RDF is one of the most powerful tools designed for this purpose. RDF defines a modeling technique in which data is defined in the form of triples. These triples express information in the graphical form.

5) **RDFS.** Resource Description Framework Schema is used to define a vocabulary to RDF data model. It is used to describe specific kind of resources along with their specific properties. RDFS vocabulary provided in the form of RDFS concept which makes uses of classes and properties to describe resources.

6) **Ontology:** Ontology is one of the basic building blocks of semantic web. It is a knowledge set which is related to a particular domain and it also defines the relationships that hold between them. This relationship may be direct or inverse. For expressing ontologies there are ontology languages such as OWL (Web Ontology Language). It provides a common model for data representation that allows the user to link one piece of information to other piece of information.

7) **Logic, Proof & Trust:** Logic layer is used to obtain the conclusions and describe how these conclusions should be used for the implementation of the semantic web. Proof layer is used to for the verification of the results produce by applying various logics. Finally the trust layer is used to enhance the trust level between users and source of information.

3. SEMANTIC SEARCH ENGINES

To retrieve information from web, search engines are required. With the development in the web, search engines are also developed. Under web 2.0, search engine retrieve those contents that contain keywords that are present in the given query. Due to the unstructured behavior of information in the web page, user still had to mine his required information from the documents which were retrieved by the search engine on the basis of keywords [7]. Due to which user get irrelevant results with low accuracy. In order to retrieve complex information, user has to provide separate queries and can manually merge the result. These search engines cannot figure out the state of affairs in which a word is being used.

Therefore, there is a need of intelligent search engine which a semantic web provides. These search engines are called as semantic search engines. Semantic search engine has the capability to understand the meaning of user query and on the basis of this meaning, they can retrieve the exact result or document from the web. Semantic search engines are based on context and concept of the searched phrase. They provide more meaningful result by understanding and evaluating the search query and finding the most relevant data in web. The next two sections will describe about the approaches used by semantic search engines and various types of semantic search engines.

3.1 Approaches of semantic search engine

Ability to understand the content of the web page is the main feature of semantic search. Semantic search engines provide more accurate and meaningful content in result. Working of each semantic search engine is different from another semantic search engine. The main goal of semantic search engine is to make data machine readable. It is needed to ensure that the semantics are not missing during the complete life cycle of information extraction [8]. In general are four approaches for semantic search. Each of the semantic search engines may use either one or more of the given approaches.

First approach: *A* Contextual approach to disambiguate and to make the queries single meaning [4]. For example the word "tear" refers to tear in eyes or rip or anything else.

Second approach: It mainly focuses on the reasoning. From the given library of facts it can generate new reasoning.

Third approach: In this approach the main focus is given on natural language understanding. With this approach the search engines try to identify the semantic of the query purposed by the user, by applying rules. Powerset makes wide use of the natural language understanding [8].

Fourth approach: This approach uses ontology for expansion of the query and knowledge representation. For example when user fire a query for a term "apple" the system add terms from its ontology (such as "fruit" because an apple is a kind of fruit) to make the search more focused as well as more accurate.

Semantic search engines can use combination of these approaches in order to become more efficient in searching.

3.2 Semantic search engines

To access the data present on the web in a well defined way, number of semantic search engines are introduced which can understand and evaluate the semantic of the data and can display the exact result as compared to the traditional search engines. There are various types of semantic search engines which are discussed here.

Swoogle

Swoogle was founded by Li Ding and Tim Finin. It was designed to automatically discover semantic web documents (SWD) on semantic web. It is a crawler based indexing and retrieval search engine for RDF and OWL documents. It mainly focuses on finding exact ontology and if it does not, it creates new ontology based on ontology ranking algorithm. Swoogle is also a content based search engine that analyses, discovers, and indexes knowledge in the web [9]. But swoogle has some limitations as it take a very long time for query response and it also lack in indexing. The key goal in building Swoogle is to design a system that will scale up to handle millions and even tens of millions of documents [10].

Hakia

Hakia was launched in March 2004. It is used to search the structured text. It does not use page popularity or keyboard match concept to produce the search result, rather it is based on the meaning of content. This search engine is specifically designed to produce output on its understanding power of web content. Hakia calls itself a "meaning based (semantic) search engine [11].

DuckDuckGO

DuckDuckGo is a privacy concern search engine. Other search engines track user's search even if user is using private browsing mode. But DuckDuckGo do not track user's search. It does not store IP addresses, does not log user information and user cookies only when needed to maintain the privacy of the user [9]. If mainly focus on retrieving data from accurate and most relevant sources rather than number of sources. DuckDuckGo represent relevant answer of a query called instance answer, which is further called as zeroclickinfo.

3.3 Challenges for semantic search engines

A. **User interface:** A friendly user interface is the first and foremost feature of any search engine. There are so many search engines like Google, Yahoo and Bing provide a best end-user interfaces. Even if some search engines do not provide best result for query response but still end users remain to these search engines because of their end user experience. Enhancements to the end-user interface of a semantic query search engine needs important development so that poor input representation of a query will automatically suggest corrections for spelling mistakes and poor grammar, and of course find the best matched results with a high accuracy[12].

B. **Efficiency:** The efficiency of a semantic search engine depends upon its performance such as the request time to web server and its associated response time. For an efficient semantic search engine query execution time must be as low as possible.

C. **Scalability:** Scalability is the ability of a search engine to efficiently handle rapid growth in data. Scalability for data in a semantic web presents additional challenges because of the openness of the RDF protocol [13]. Many of relational databases are very efficient, successful and scalable due to their structure of relational data.

D. **Cost Effectiveness:** Provide a cost effective solution is the main feature of a high quality search engines. As RDF is open source, due to which processing queries become very expensive. Some efforts have been made to introduce cost effective search algorithms such as the SPARQL as search technique [14].

E. **Page Rank:** The main goal of semantic search is to provide accurate and exact result in response to a user query. For this purpose, various kinds of page ranking algorithms are used for rating the web pages. There are billions of web pages in www which makes it very crucial for a search engine to sort and rank the retrieved documents effectively.

4. CONCLUSION

This paper gives a brief overview of semantic web mining and various semantic search engines. Semantic web is the intelligent version of current web that help us to retrieve meaningful data from the web. It is concluded that Semantic search engines provide a unique experience for user on web. But there are some challenges for semantic search engines. Future work includes developing an efficient technology that can deal with the challenges and also is compatible with the global standards of the web technology.

REFERENCES

[1] Brijendra Singh, Hemant Kuamr Singh “Web data mining research: a survey” 2010 IEEE.

[2] S. Brindha and M. Vasantha “Data Mining-Semantic Web mining” 2010 IJCCIS.

[3] Sivakumar J, Ravichandran K.S “A Review on Semantic-Based Web Mining and its Applications” IJET.

[4] Junaid Rashid and Muhammad Wasif Nisar “A study on semantic searching, semantic search engines and technologies used for semantic search engines” 2016, IJITCS.

[5] Rashmi Bakshi, Abhishek “Semantic web-an extensive literature review” IJMTER.

[6] K. Sridevi and Dr. R. Umarani “A survey of semantic based solutions to web mining” IJETTCS

[7] Ranjna Jain, Neelam Duhan, A.K. Sharma, “Comparative study on semantic search engines” Volume 131 – No.14, December 2015.

[8] Nikhil Chitre “Semantic Web Search Engine” Volume 4, Issue 7, July 2016.

[9] Vidhi Shah, Akshat Shah and Asst. Prof. Khushali Deulkar “Comparative study of semantic search engines” Volume 4 Issue 11, Nov 2015, IJECS.

[10] Li Ding , Tim Finin, Anupam Joshi, Yun Peng, R. Scott Cost, Joel Sachs, Rong Pan, Pavan Reddivari, Vishal Doshi “Swoogle: A Semantic Web Search and Metadata Engine”.

[11] G. anuradha, G. Sudeephi, Dr G. Lavany Devi, Prof M. surrendra Prasad babu “A comparative analysis of semantic web search engines”.

[12] Arooj Fatima, Cristina Luca, George Wilson “ New framework of semantic search engine” March 2014.

[13] M.M.El-gayar, N.Mekky , A. Atwan “Efficient proposed framework for semantic search engine using new semantic ranking algorithm” IJACSA,Vol. 6, No. 8, 2015.

[14] L. Chang, W. Haofen, Y. Yong and X. Linhao, “Towards Efficient SPARQL Query Processing on RDF Data” vol. 15, issue 6, Dec. 2010.