 ISBN:978-93-5254-640-4

Semantic Web Mining a knowledge digger

Surabhi D. Thorat

Asst. Prof., Dept. of MCA, MIT(E), Aurangabad(MS)
thorat.surabhi@gmail.com

***Abstract*— Semantic Web mining major goal is to integrate the areas of Semantic Web and Web Mining by using semantics to improve mining and find hidden knowledge. The integration of both these areas can result in making the web more 'semantic'. This paper high lighten the role of web semantic, layered architecture, challenges and its major building blocks.**

***Keywords*—Semantic Web, taxonomy, semantic web layered architecture.**

I. INTRODUCTION

With the evolution of Web 2.0 applications the web information is increasing tremendously. This increases the demand of finding useful patterns in the web that can be the rich source of data mining process. Semantic Web is an evolving extension of the Web 3.0 where information is tagged in relation to use and context .Similar information can be delivered more effectively to humans and machines. 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. The Semantic Web has a layer structure that defines the levels of abstraction applied to the Web. At the lowest level is the familiar World Wide Web, then progressing to XML, RDF, Ontology, Logic, Proof and Trust [1]. 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 Mining is considered as a combination of two areas Semantic Web and Data Mining. Semantic Web represents the extension of the World Wide Web that gives users of Web the ability to share their data beyond all the hidden barriers and the limitation of programs and websites using the meaning of the web [2] while Data Mining deals with large amount of data to find consistent patterns.

II. SEMANTIC WEB MINING TAXONOMY

- Web content mining focus on the discovery of knowledge from the content of web pages and therefore the target data consist of multivariate type of data contained in a web page as text, images, multimedia etc. It finds the relevance of the content with search query.
- Web usage mining consists of 3 main phase's namely pre-processing, pattern discovery and pattern analysis. It focuses on the discovery of knowledge from user navigation data when visiting a website. The target data are requests from users recorded in special files stored in the websites servers known as log files.
- Web structure mining deals with the connectivity of websites and the extraction of knowledge from hyperlinks of the web. It is the process of using graph theory to analyse the node and connection structure of a website.

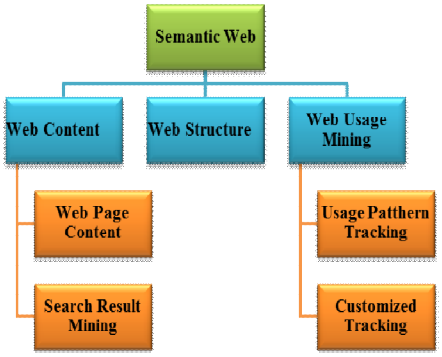

Fig 1. Semantic Web Mining Taxonomy

III. SEMANTIC WEB

The Semantic Web is based on a vision of Tim Berners-Lee, the inventor of the WWW.The Semantic Web initiative presents a prominent recent approach attempting to provide the web with a meaning not only people, but also machines can process. In a nutshell, meaning is usually understood as the process of giving sense to symbols of a language, or, in other words, associating the symbols with the real world objects and ideas they are supposed to refer to [3].

Semantic Web provides a common syntax for machine understandable statements, define common vocabularies.. Semantic Web not only gives support for information access on the Web by direct links or by search engines but also to support its use.

IV. SEMANTIC WEB LAYERED ARCHITECTURE

In the first layers, a common syntax is provided. Uniform resource Locator provides a standard way to refer to entities, while Unicode is a standard for exchanging symbols. In second layer the Extensible Mark-up Language (XML) fixes a notation for describing labelled trees, and XML Schema allows the definition of grammars for valid XML documents. XML documents can refer to different namespaces to make explicit the context of different tags. The Resource Description Framework (RDF) can be seen as the first layer where information becomes machine understandable. RDF

provides interoperability between applications that exchange machine understandable information on the Web. RDF documents consist of three types of entities: Resources, properties, and statements. Resources may be Web pages, parts or collections of Web pages, or any real-world objects which are not directly part of the WWW. In RDF, resources are always addressed by URIs. Properties are specific

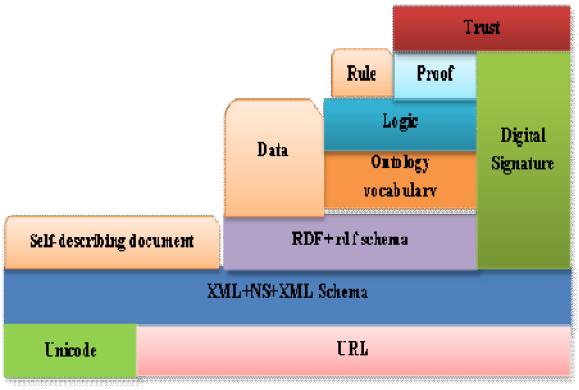

Fig 2. Semantic Web Layered Architecture

attributes, characteristics, or relations describing resources. A resource together with a property having a value for that resource forms an RDF statement. The next layer is the ontology vocabulary is an explicit formalization of a shared understanding of a conceptualization. Logic is the next layer according to Berners-Lee. Today, most research treats the ontology and the logic levels in an integrated fashion because most ontologies allow for logical axioms. By applying logical deduction, one can infer new knowledge from the information which is stated explicitly. Proof and trust are the remaining layers. They follow the understanding that it is important to be able to check the validity of statements made in the Semantic Web, and that trust in the Semantic Web and the way it processes information will increase in the presence of statements thus validated.

V. SEMANTIC WEB CHALLENGES

- Major search engines based on Structured-data mining that are not able to deal with public domain which specifically addresses the requirement of mining or searching unstructured data flavoured with semantic search.
- There is no standardized web form structure. We can search for and extract information available as HTML, but till date, we are not proficient to gain easy access to the hidden web. It is very difficult to get to the accurate web form, and even harder to find a suitable truthful web application and related service and therefore we cannot leverage the wealth of information residing behind web forms and services for the masses.
- We cannot apply the techniques for exploiting corpora of documents directly for searching unstructured data. The main reason is that searching unstructured data requires an understanding of its underlying semantics.
- Lack of global standards. There is any standard support available which can help the end-users to extract valuable and handy information from audio and video files available in huge quantity across the world.

VI. SEMANTIC WEB BUILDING BLOCKS

A. *Resource Description Framework (RDF)*

An XML-based standard knowledge representation format for exchanging arbitrary information. RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.

RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a “triple”). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications.

This linking structure forms a directed, labeled graph, where the edges represent the named link between two resources, represented by the graph nodes. This *graph view* is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations.

B. *Web Ontology Language (OWL)*

A standard for describing classes of objects and enabling inference. The Ontology Web Language (OWL) is a set of markup languages which are designed for use by applications that need to process the content of information instead of just presenting information to humans. The OWL ontology describe the hierarchical organization of ideas in a domain, in a way that can be parsed and understood by software. OWL has more facilities for expressing meaning and semantics than XML, RDF, and RDF-S, and thus OWL goes beyond these languages in its ability to represent machine interpretable content on the Web. OWL is part of the W3C recommendations related to the Semantic Web. RDF

Query, RDF Rules, Access, and more-Pre-standardization, software components

What is achieved?

- Integration of diverse data sources.
- Focus on information needs.

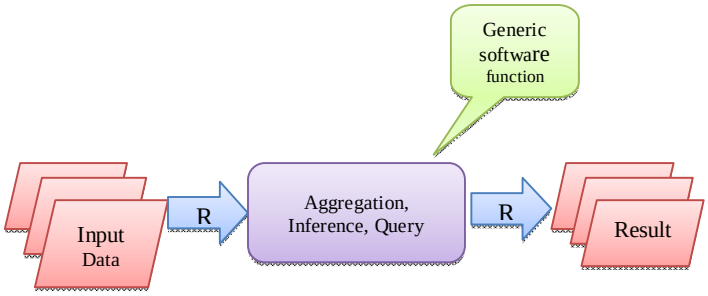

Fig 3. Key phases of Semantic Web

- Generate new knowledge
- Programs and sites can easily exchange information.
- Search Engines can display more relevant information in results.
- Data Mashers can combine data from different datasets to find new and outstanding things. Researchers can take large amount of data and try to make sense of it.
- Semantic web helps machines understand what the information on the web page is and the relationship between pieces of information.[4]

VII. CONCLUSION

Semantic Web Mining is a new and fast-developing research area combining Web Mining and Semantic Web. Semantic Web Mining can be used to discover interesting user navigation patterns, which then can be applied to real world problems such as website improvement, additional topic, product recommendations, customer behavior's study etc. In this paper a detailed review of Semantic Web Mining has been presented that explain how the semantic web is used for mining the World Wide Web.

VIII. REFERENCES

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

[2] Quboa, Qudamah K., and Mohamad Saraee. "A State-of-the-Art Survey on Semantic Web Mining." Intelligent Information Management 5.1 (2013).

[3] Ogden, C. K. and Richards, I. A. (1989). The Meaning of Meaning. Mariner Books.

[4] Syeda Farha Shazmeen , Etyala Ramyasree, Semantic Web Mining: Benefits, Challenges and Opportunities, International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-7 December-2012