A Review on Semantic-Based Web Mining and its Applications

Sivakumar J[#1], Ravichandran K.S[*2]
[#]School of Computing, SASTRA University,
Thirumalaisamudram, Thanjavur, Tamil nadu, India.
[*]School of Computing, SASTRA University,
Thirumalaisamudram, Thanjavur, Tamil nadu, India.
[1]jpsivas@gmail.com
[2]raviks@it.sastra.edu

Abstract: **In this paper we survey the Semantic-based Web mining is a combination of two fast developing domains Semantic Web and Web mining. These two fields address the current challenges of the World Wide Web (WWW). The idea is to improve the results of Web Mining by making use of the new semantic structure of the Web and to make use of Web Mining for creating the Semantic 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. This survey analyses the approach of both areas. This paper first introduces the knowledge of Semantic Web and Web mining techniques, and then discusses the semantic-based Web mining and its applications and finally discuss the survey on sematic based web mining tools.**

Keywords: Semantic web, Web mining, ontology, Semantic web mining.

I. INTRODUCTION

A. Semantic web:

The current World Wide Web (WWW) has a huge amount of data that is often unstructured and only human understandable. Web is rich with information; gathering and making sense of the data in the web is more difficult because the document of the Web is largely unorganized and unstructured. From the unorganized human readable web data semantic web is how to effectively and efficiently creating a machine-understandable, queriable, information and knowledge layer. If computer can understand the meaning behind the information, it can learn what we are interested in and it help us better find what we want.

Since the semantic Web mainly focuses on the data and information. Data in the Semantic Web is well defined and linked in a way that can be used for more effective discovery, automation. The nature of most data on the Web is unstructured that only understand by humans, the amount of data is very huge on the web that processed efficiently by machines.

The goal of the Semantic Web is to develop allowing standards and technologies designed for both user and machines understandable. Semantic web information can support data integration, data discovery, navigation, and automation of tasks. The Semantic Web Layered Architecture will describe in Figure 1.

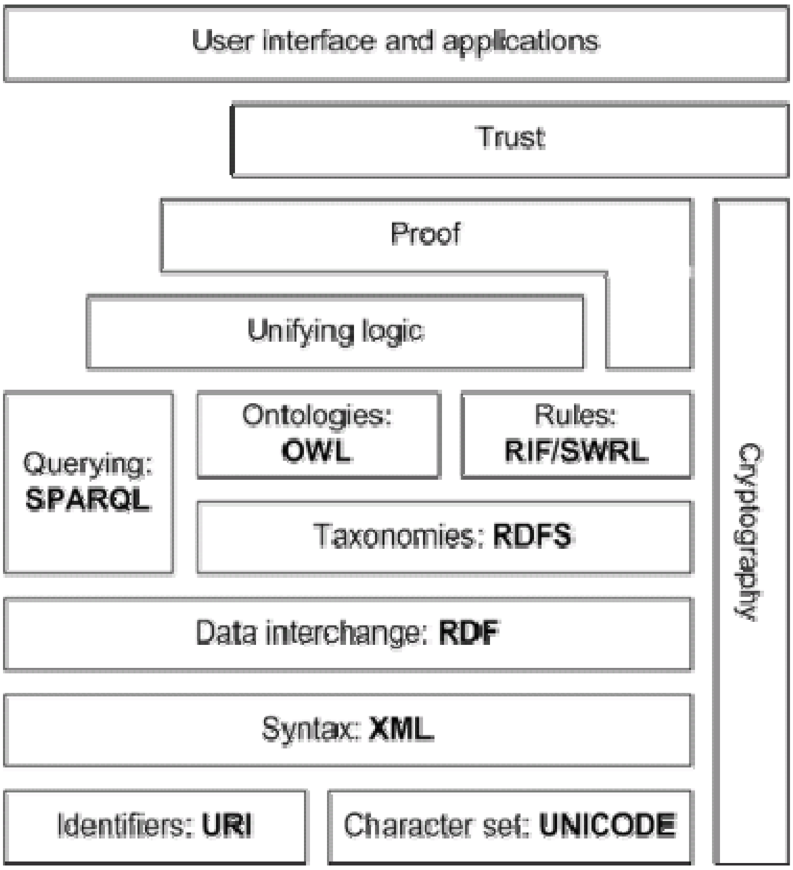

Fig.1. Layers of the Semantic web

i. Uniform Resource Identifiers (URI) and Unicode follow the important features of the existing WWW. URI is simply a Web identifier. URI identification allows interaction with representations of the resource over a network (usually the World Wide Web) using specific protocols like http or ftp. The purpose of an URI is to specify an identifier to represent a resource of a web in a uniform path. URI is used to identifying information representation and constructs including classes, properties and individuals. URIs is the fundamental benefit of semantic web technology. URIs provides users to know exactly what it is they are being referred.

ii. 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.

iii. Extensible Markup Language (XML): XML is a language used to transport and store data on the web.XML is only to carry 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 schema also called as XSD XML Schema Definition. XML Namespace in semantic web is used to avoid conflict data or names. XML layer aim to provide the basic syntax and structure of the data on the web.

iv. Resource Description Frameworks (RDF): RDF is a framework for semantic web based on XML. RDF is XML based language used to describe information and resource with classes, properties and values on the web. In web semantic RDF is used to describe the web resources. RDF representing metadata about World Wide Web resources, such as the author, title, and modification date of a Web page. RDF is used for storing any other semantic data. Semantic web uses RDF as the primary representation language and provide data inter change data on the web.

v. Resource Description Frameworks Schema (RDFS): RDFs extension of RDF. RDF Schema provides the framework to describe application specific classes and properties. RDFS is used to describe classifications of classes and properties. RDFs do not define the classes and properties. It is similar to OOP Object Oriented Programming.

vi. Web Ontology Language (OWL): OWL is based on the top of the RDF and XML based language. RDF is used to represent the rich and complex knowledge about things and their relationship. OWL provides processing information on the web. OWL is a part of web semantics. There are two types of OWL properties i.e. Object properties and Data type properties. OWL layer is uses to represent the ontologies of the semantic web.

vii. Rule Interchange Format (RIF) and Semantic Web Rule Language (SWRL): RDFS and OWL have defined semantics and it used to observation on the web. SWRL consist of OWL Lite and OWL DL. It is also based on XML. RIF and SWRL provide rules for the semantic web.

viii. Simple Protocol and RDF Query Language (SPARQL): SPARQL is query language like and protocol for RDF. SPARQL used to querying the RDF data, RDF Schema and OWL ontologies with knowledge. SPARQL based on RDF data model. The results of SPARQL queries in the form of XML.

Proof and Trust layer:

ix. *Proof layer* is used to verify the results produced by the agents should be believed or authenticate the agent behaviour. Trust layer is to provide a mechanism for trust and poise between information users (man or machine) and information sources.

On the top of the layer user interface and applications are built. Table 1 summarization of the semantic web layers.

Layers	Name	Description
Layer 1	URI and Unicode	Unicode Processing resources to encoding, URI: Used for identification of resources.
Layer 2	XML	Used to represent the data content and structure
Layer 3	RDF and RDF schema	Used to describe resources on the Web and types
Layer 4	OWL	Describe the various types of resources and the relationship between resources
Layer 5	RIF	In the following four layers operate on the basis of logical reasoning
Layer 6	SPARQL	Query language and protocol for RDF.
Layer 7	TRUST and PROOF	According to logic, to verify statements in order to draw conclusions and The establishment of a trust relationship between users

Table 1: Layers of Semantic web

Web mining:

Web is a collection of hyperlinked documents on one or more Web servers. Web mining is data mining techniques used to extract knowledge from Web. Web mining is a helpful tool in the process of transforming human understandable content in to machine understandable semantics. The classification of web mining techniques represented in below Figure 2.

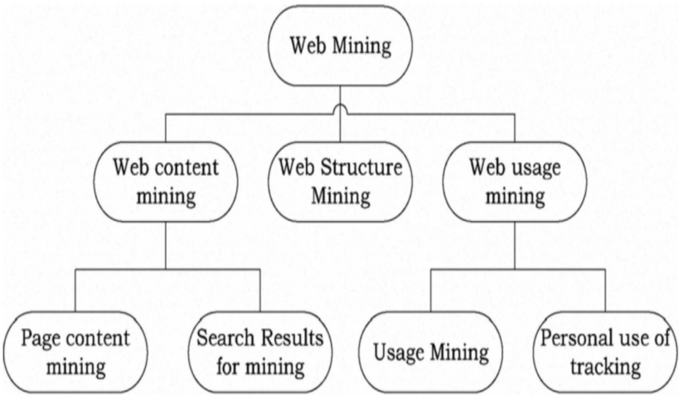

Fig 2. Classification of Web Mining Techniques

i. Web Content mining: Web Content Mining is the process of extracting information from the contents of Web documents. It examines content of the web pages as well and web searching. Content data corresponds to the collection of facts a Web page was designed to convey to the users. Web content may be unstructured (plain text), semi- structured (HTML documents), or structured (extracted from databases into dynamic Web pages).

Such dynamic data cannot be indexed and consist what is called “the hidden Web”. A research area closely related to content mining is text mining.

ii. Web structure mining: Web structure mining is mostly interested in the hyperlinks of the web pages. Web Structure Mining can be is the process of mining structure information from the Web. Web structure mining is used to improve the structure of the web pages. Depending upon the hyperlink, the web pages categorize the Web pages and the related information and inter domain level.

iii. Web usage mining: Web usage mining is the process of extracting information from server logs i.e. user’s history and web user behaviour. The logs can be examined by client perspective or server perspective. This information takes as input the usage data, i.e. the data exist in in the Web server logs showing the visits of the users to the Web site. Web usage mining is the process of identifying browsing patterns by analysing the user’s navigational behaviour.

To attain the concept, Web data (usage, content, structure) are represented by using developing model of representation, ontologies. This representation had the gap between Semantic Web and Web Mining areas, to create a research area, which of Semantic based Web Mining [1].

B. 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) a. 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, hidden data, and making available of web data. It is essentially mining the information pertaining the semantic web. In semantic based web mining the web pages are mined by the machine can perform better understand the information on the web pages. It also mining the data source to develop an effective semantic web. Figure 3 illustrates the semantic based web mining technology.

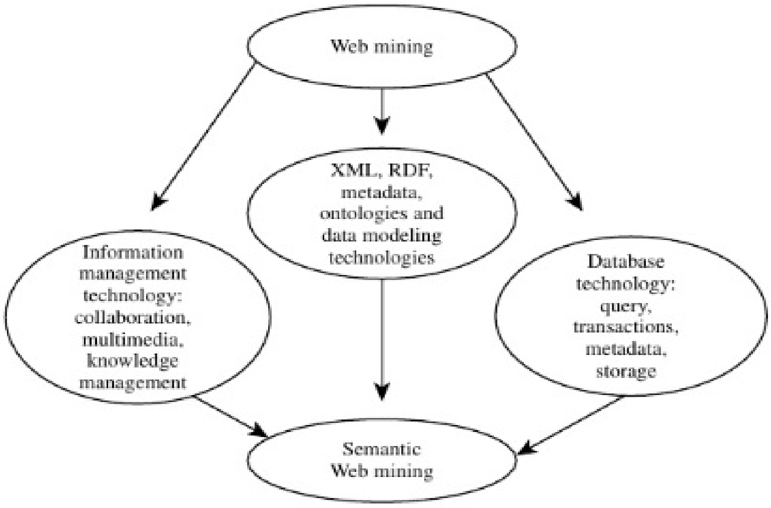
Web mining
XML, RDF, metadata, ontologies and data modeling technologies
Information management technology: collaboration, multimedia, knowledge management
Database technology: query, transactions, metadata, storage
Semantic Web mining

Fig.3. Semantic based web mining Technologies

It is basically mining 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. It also used provide the human and machine can better perform their task. Semantic web requirement are considered in three major groups ontology, semantic web content and web service.

C. *An Ontology approach:*

Ontology is the backbone of the semantic web. A Semantic Web vocabulary can be considered as a special form of usually light-weight. Ontology is a collection of URIs with a usually informally described meaning. Ontologies are represented by a formal ontology language. In [2], [3] ontology plays a major role. 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. Ontology is defined as a set of objects, concepts, and other entities that are exist in some area and the relationships that occur them. Figure 4 illustrates the ontology and semantic based web mining representation.

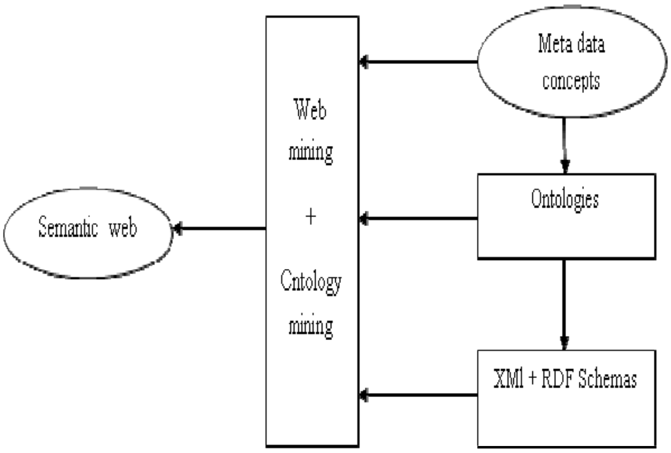

Fig.4. Ontology and semantic based web mining representation

In Semantic based web mining also mine the ontologies on the web, using ontologies on the web provide that it more intelligent. Ontologies are developed from metadata. For example RDF schemas mine the metadata and ontologies. RDF schema used to developing the semantic web. I.e. information or data mined from ontologies and RDF schemas can be used to perform better understandable of web pages. Semantic web technologies represent meaning using ontologies and provide reasoning over the rules, logic, relationships and conditions represented in the ontologies.

II. Applications of Semantic based Web Mining

Mahindra Pratap et al., [4] The Internet has developed from a collection of static HTML pages. Internet consists of static web pages and to providing dynamic, interactive content. Semantic based web mining application plays a major role in e-commerce to managing business processes. Semantic Web applications are decentralized; open to operate on distributed data. Semantic web applications follows semi structured schemas.

Berendt,B et al., [5] Semantic based web mining application includes many areas such as e-activities, health care, bioinformatics, privacy and security, and knowledge management and information retrieval . It proposes great chances in finance, business, marketing, commerce, finance, education, research and development.

P. Markellou et al., [6] Web mining research focused on E-learning is focused on web usage mining; based on how the student performs and their activities. Now-a-days digital libraries are also accessible from the web. Many commercial institutions are transforming their businesses and services electronically. The challenge of the Semantic based Web Mining technologies in the E-Learning domain to provide the personalized experiences for the users. These applications can take into the individual needs and requirements of user or learners.

Lappas, G [7] Semantic based web mining is applied in the E-Services areas like E-Government, E-Politics and E-Democracy. Only web mining applications have been related to these areas. Most of the government information is placed on the web. Current web mining research focused on E-Politics is based on web structure mining to identify political groups. It appears that the fields of E-services and web mining have recently had leading benefits to society.

Dzeroski S et al., [8]. Semantic based web mining is also applied in genetics, social network analysis, molecules and natural language processing .Semantic based web mining is also being applied in the search engine.

Semantic based web mining application of ontology is grouped into two classes improved search web data and better web browsing capability.

Naing et al., [9] Improved search of web data with semantic ontology of web data can be indexed by concept and relationship. By using concept and relationship of semantic ontology provides a better search.

Jean Vincent et al., [10] better web browsing capabilities in web searching web pages browsed by using ontology concept and relationship. If web pages concept and relationship instances can be created a web page virtual link between web pages belongs to the concept of interests. It provides the better web browsing capabilities.

Xiaohui Tao et al., [11] Semantic based web mining application with ontology based on the semantic based personalized web search engine is used to achieved by user recommended web pages only display. The displayed web pages are personalized web pages with user interests. Personalized web pages take advantages of semantic web and web mining, it may provide to improve the web search. Table 3 shows the summarization of web mining applications.

Authors	Semantic based web mining applications
Mahindra Pratap et al., [4]	E-Commerce for business process
Berendt,B et al., [5]	Health care, bioinformatics , privacy and security and Information Retrieval
P. Markellou et al., [6]	E-learning
Lappas, G [7]	E-Services like E-Politics , E-Government and E-Democracy
Dzeroski S et al., [8].	Genetics ,social network analysis and Natural language processing
Naing et al., [9]	Improved search engine of web data
Jean Vincent et al., [10]	For better browsing capabilities
Xiaohui Tao et al., [11]	Semantic based personalized search engine

Table 3: Semantic based web mining Applications

III. Literacy survey on semantic based web mining tools:

Diana Cerbu et al., [12] proposed two developing area Semantic Web and Web Mining. The author's proposed how these two areas can be combined with three different approaches to semantic based web mining an approach to pattern mining; a text classification algorithm is called AdaBoost and a framework for creating well customized content on the web by using web mining and semantic ontologies.

Thomas Fischer et al., [13]. Motivated the application of relational data mining algorithms in semantic web. They have outlined important differences to the knowledge discovery process. The modelling, selection and transformation are different to the standard approach. The knowledge discovery process recommended choosing parts of semantic data to fully use information and background information derived from a web sources.

Nizar R. Mabroukeh et al., [14] proposed a generic framework called SemAware that integrates semantic information into web usage mining. Semantic information can be combined into the pattern discovery. A semantic distance matrix is used in the agreed sequential pattern mining algorithm to trim the search space and partially relieves the algorithm from support counting. A 1*st*-order Markov model is used to build the mining process and enriched with semantic information.

Yao et al., [15] proposed a framework designed for an intelligent agent that dynamically gives the recommended to the web site's users by learning from web usage data and users' behaviour is called PagePrompter,. Like a guide, an agent supports a user in navigating the web site. PagePrompter can also be used as a tool for understanding user behaviour, the design of web sites, system performance analysing, web site designer for improving web sites and generating an adaptive web site.

Engels et al., [16], exposed a technical solution for improving the semantics. The CORPORUM tool set is used to develop the task exists for a set of programs. Tasks like either as standalone or augmenting. The aim of the semantic web is also to enable the use of logical reasoning on web contents.

Yuefeng Li et al., [17] developed an ontology mining technique for retrieving related information from the web. The ontology consists of two parts: the top backbone and the base backbone. The top backbone ontology explains the linkage between compound classes of the ontology. The base backbone ontology explains the linkage between primitive classes and compound classes. The mathematical model is used to represent discovered knowledge on the ontology.

P. Markellou et al., [18] proposed a framework for personalized E-Learning based on collective usage profiles and domain ontology. The authors distinguished two stages online and offline tasks. An Offline task includes ontology creation, data preparation and usage mining and online tasks that include the production of recommendations.

Baoyao Zhou1 et al., [19] proposed a web usage mining approach for semantic web personalization. Provide a semantic web personalization needs to challenge the technical issues on web access activities, convert them into ontology automatically discover hierarchical relationships from web access activities, and presume personalized usage knowledge from the ontology. This approach combined fuzzy logic into Formal Concept Analysis to mine client side web usage data for automatic ontology generation, and then applied fuzzy approximate reasoning to deduce personalized usage knowledge from the ontology.

IV. CONCLUSION

In this survey we have studied the two fast developing research areas in World Wide Web 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; and to make use of Web Mining for building up the Semantic 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, bioinformatics, privacy and security, and search engines, knowledge management and information retrieval.

V. REFERENCES:

[1] Berendt, B., Hotho, A. and Stumme, G 2002. Towards Semantic Web Mining Proceedings of the First International Semantic Web Conference on the Semantic Web Springer-Verlag, pp. 264-27.

[2] Yuefeng Li; Ning Zhong; 2006. Mining ontology for automatically acquiring Web user information needs, Knowledge and Data Engineering, IEEE Transactions on , vol.18, no.4, pp. 554- 568.

[3] Maedche, A., Motik, B. and Stojanovic, L.Managing 2003. Multiple and Distributed Ontologies on the Semantic Web VLDB Journal volume 1pages: 286--302 .

[4] Mahindra Pratap Singh Dohare*1 and Sanjaydeep Singh Lodhi, Vinod Mahor, 2011. Application based semantic web mining technique, JGRCS journal Volume 2, pp No. 3.

[5] Berendt, B., Hotho, A., Mladenic, D., Someren, M., Spiliopoulou, M. and Stumme, 2004. G.A Roadmap for Web Mining: From Web to Semantic Web Berendt, B., Hotho, A., MladeniÄ?, D., Someren, M., Spiliopoulou, M. & Stumme, G. (ed.)Web Mining: From Web to Semantic Web Springer Berlin Heidelberg, ,Vol. 3209, pp. 1-22.

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

[7] Lappas, G 2007. An Overview of Web Mining in Societal Benefit Areas E-Commerce Technology and the 4th IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services, , vol., no., pp.683-690, 23-26.

[8] Dzeroski S Relational Data Mining Maimon, O. & Rokach, L. (ed.) 2010. Data Mining and Knowledge Discovery Handbook Springer US, pp. 887-911[books].

[9] Naing, M.M, Lim, E.P., and Chiang, R. H.L. Core 2005. A Search and Browsing Tool for Semantic Instances of Web Sites", Asia Pacific Web Conference (APWeb'05).

[10] Jean Vincent Fonou-Dombeu1, and Magda Huisman, 2011. Combining Ontology Development Methodologies and Semantic Web Platforms for Egovernment Domain Ontology Development International Journal of Web & Semantic Technology (IJWesT) Vol.2

[11] Xiaohui Tao, Yuefeng Li, and Ning Zhong, 2011. Senior Member, IEEE, "A Personalized Ontology Model for Web Information Gathering, ieee transactions on knowledge and data engineering, vol. 23, no. 4.

[12] Diana Cerbu, Romania Konstanz 2008. Semantic Web Mining journal of web service and Semantic web.

[13] Thomas Fischer,Johannes Ruhland 2010. Towards Knowledge Discovery in the Semantic Web, MKWI – Business Intelligence, vol 2 pp 151 -166

[14] Nizar R. Mabroukeh and Christie I 2009 .Using Domain Ontology for Semantic Web Usage Mining and Next Page Prediction,. Ezeife CIKM'09, 2–6.

[15] Y.Y. Yao, H.J. Hamilton, and Xuewei Wang PagePrompter 2008. An Intelligent Agent for Web Navigation Created Using Data Mining, Volume 27 Issue 3, Pages 59 - 74.

[16] R.H.P. Engels, B.A.Bremdal and R. Jones, Halden 2001. Norway CORPORUM: a workbench for the Semantic Web Conference of EXML/PKDD workshop.

[17] Yuefeng Li; Ning Zhong; 2006. Mining ontology for automatically acquiring Web user information needs Knowledge and Data Engineering, IEEE Transactions on , vol.18, no.4, pp. 554- 568.

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

[19] Baoyao Zhou1, Siu Cheung Hui, and Alvis C. M. Fong 2006 Web Usage Mining for Semantic Web Personalization.

[20] Wang Yong-gui; Jia Zhen; 2010. "Research on semantic Web mining," Computer Design and Applications (ICCDA), 2010 International Conference on, vol.1, no., pp.V1-67-V1-70, 25-27.

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

[22] T.Berners-lee, N.Shadbolt and W.Hall 2006. The Semantic web revisted, IEEE intelligent systems pp :96-101