ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801
International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 2, April 2013

A Survey on Semantic Web and Knowledge Processing

M.Venu Gopalachari[1], Dr. P. Sammulal[2]

Dept. of CSE, Chaitanya Bharathi Institute of Technology, Hyderabad, India[1]

Dept. of CSE, JNTU College of Engineering, Jawaharlal Nehru Technological University, Hyderabad, India[2]

ABSTRACT: The biggest challenge in the next several years is how to effectively and efficiently find what has been requested. A normal user generally spends hours to find the exact requested information. Semantic Web Mining contributes responses to address this problem. It aims to integrate the areas of Semantic Web and Web Mining by using semantics to improve mining and mining to generate semantics. The integration of both these areas can result in making the web more 'semantic'. This paper provides an overview of the state of the art in the research on semantic web and knowledge processing and presents some recent research initiatives.

KEYWORDS: Semantic web, Artificial Intelligence, Data Mining, Information overload, Web Mining.

I. INTRODUCTION

To filter the results provided by various search engines like Google and Yahoo, the personalized access to the information available on the Web is required (Svatopluk et al., 2005). As of 2008, the estimated size of the web's portion accessible by search engines was already one trillion pages [21]. Individual users are hardly able to make any sense out of even tiny fractions of the available content, being overwhelmed by numerous resources that may or may not be related to what they are looking for. The sheer scale of the web, together with its decentralised, highly redundant and largely inaccurate nature, makes using the knowledge within rather cumbersome. Moreover, the relevant knowledge can be scattered across many resources, which renders the attempts to make use of all the available content even more complicated.

This problem is often referred to as "information overload". To some extent, the problem has been tackled by advanced technologies based on the field of information retrieval, which power the nowadays web search engines and make finding of resources relatively easy. The resulting information overload [22] problem is being faced by many state of the art technologies drawing inspirations from various branches of computer science. Probably the most influential field in this context (at least regarding industry-strength Introduction applications) is information retrieval [3], most visibly encountered in the form of web search engines like Google, Yahoo or Bing (cf. http://www.google.com, http://www.yahoo.com, http://www.bing.com, respectively). The information retrieval methods cover substantial portion of the web content, but they merely oat on the surface of the actual meaning of the data they index due to their reliance on mere string-base The Semantic Web attempts to complement the rather shallow information retrieval approach by adding meaning to the strings of the web content with the statistics and heuristic ranking [4]. In the next sections, we start with a brief overview of the areas Semantic Web and Web Mining. After that section, an overview of challenges and future trends in the implementation of semantic web is described.

II. SEMANTIC WEB

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 [5]. Before one can proceed with proposing a knowledge representation and processing framework, a computational formalization of the rather abstract notion of meaning has to be adopted. In a very broad sense, the Semantic Web initiative is about giving a machine-readable meaning to the content on the World Wide Web. For the actual meaning representation, primarily the methods researched in the field of artificial intelligence (and namely in its knowledge representation and reasoning sub-discipline) have been adopted as essential design blocks of the forthcoming Semantic

Web. The traditional knowledge representation methods are not applicable to the web data in an out-of-the-box manner, though. They have had to be adapted so that they can face the challenges of the vast and chaotic world out there. Examples of the most critical challenges are the distributed nature and sheer scale of the knowledge on the web. The distributed nature has been addressed by basing the core Semantic Web standard (RDF) [6]--on model-theoretic semantics mixed with certain fundamental principles of the web (mainly naming of entities by unique resource identifiers, organized in distributed, yet interlinked name spaces). The large scale of the web is being tackled by adapting state of the art (distributed) database technologies to storage and querying of RDF data [7], and also by incorporating sophisticated optimization or parallel computing techniques into the reasoning algorithms. Overviews of various emerging semantic web technologies are given in table1.

Table1: A List of Emerging Semantic Web Technologies

Technology	**Definition**
OWL-S	OWL-S (formerly DAML-S) is a services ontology that enables software agents to discover, invoke, compose, and monitor Web resources.
OWL 2	OWL 2 extends the Web Ontology Language (OWL) with a small but useful set of features (EL, QL, and RL) that enable effective reasoning.
WSMO	The Web Service Modeling Ontology (WSMO) provides a conceptual framework and a formal language for semantically describing all relevant aspects of Web Services to facilitate the automation of discovering, combining, and invoking electronic services over the Web.
WSML	The Web Services Markup Language (WSML) provides a formal syntax and semantics for the WSMO and consists of several variants, such as WSML-Core, WSML-DL, WSML-Flight, WSML-Rule, and WSML-Full.
SWRL	The Semantic Web Rule Language aims to be the Semantic Webs standard rule language and is based on a combination of the OWL DL, OWL Lite, RuleML and so on.
RuleML	RuleML constitutes a modular family of Web sublanguages including derivation rules, queries and integrity constraints as well as production and reaction rules.
RIF	The Rule Interchange Format (RIF) aims to be the standard rule language of the Semantic Web for Rule Interchange.

Several other challenging features the Semantic Web applications have to tackle in order to become truly applicable have also been addressed recently. This includes changing knowledge [8], inconsistencies [9] or uncertainty (tackled from the fuzzy logic [10] or from the probabilistic [11] perspective).

Most approaches handling these features seek for a solution that is compatible with or an extension of the core Semantic Web standards (mainly RDF and OWL). This is a reasonable approach motivated by the well understood theoretical foundations of the standards, as well as by the pragmatic necessity of a gradual development. However, the benefits might easily be outweighed by problems that are not entirely dissimilar from the knowledge acquisition bottleneck in AI. However, the current Semantic Web solutions are often too deeply rooted in the classical paradigms of artificial intelligence (e.g., logical knowledge representation). Still, even the Semantic Web suffers from certain imperfections as of now. The prevailing logics-based conception of the expressive knowledge representation on the Semantic Web makes manual acquisition of such knowledge expensive and impractical, while the results of automatic and/or community based (and thus cheap) knowledge acquisition methods are often too noisy and sparse to be meaningfully processed.

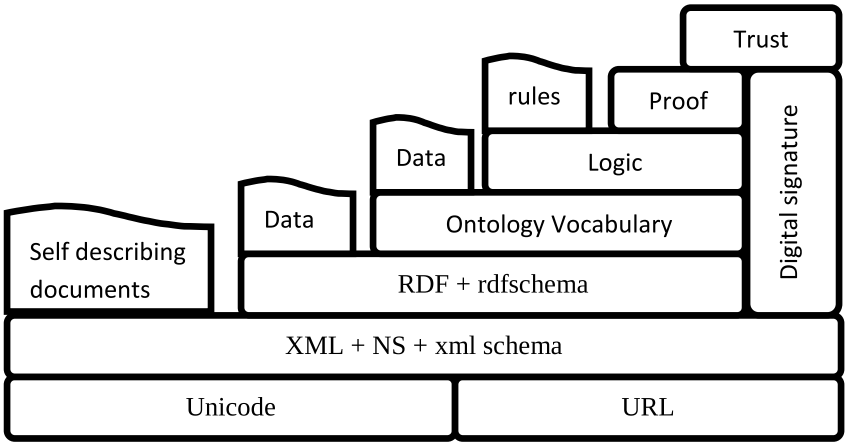

Figure 1: Semantic Web Architecture

The standards for complex knowledge representation on the web like OWL[12] or OWL2[13] are suitable for rather small-scale resources that have been well-crafted by experts, but are far from sufficient when facing the vast and noisy jungle of loosely structured knowledge on the web. The RDF [14] standard is much more simple and universal, which makes it applicable to modeling in broad range of practical scenarios. However, it still lacks formal support for some important features of the web knowledge, such as uncertainty, negation or multi-faceted contexts. Moreover, the RDF interpretation is still based on model-theoretical i.e., logical semantics, which makes it as brittle and cumbersome as the more complex Semantic Web standards from the theoretical point of view.

Last but not least, all the current Semantic Web standards have rather machines and developers as users in mind, which makes them rather inaccessible by the huge amounts of lay people daily interacting with the web. Despite of the primary focus of the Semantic Web on machines, the low level of understanding of the standards by people is an issue if we suppose the people to provide some annotations of the web content, which is often the case in practice, indeed. The two major disciplines that deal with theoretical underpinnings of the notion of meaning are philosophy and linguistics.

LINGUISTICS

The meaning (of natural language) from the viewpoint of linguistics is studied by its specific sub-discipline, semantics. The meaning is analyzed at the level of words, phrases, sentences and larger units of discourse [15]. The basic subjects of study in semantics are signs [16], which may be understood as discrete units of meaning (words, images, gestures, scents, tastes, textures, sounds, etc., essentially all forms of a message in which information can be transferred by the participants in a communication process). Two major distinct conceptions of signs have been proposed by two key figures involved in the birth of the modern linguistics:

Dualistic signs--- According to Saussure, a sign is composed of the signifier and the signified [16]. The former is conceived as a language representation of a conceivable and/or existing entity or idea, while the latter is the mental representation or a concept of the entity or idea that is being signified (i.e., the meaning of the signifier). The binding between the signifier and signified in a sign is purely arbitrary (not dependent in any way whatsoever on the actual meaning or the form of the language representation).

Signs as triadic relations--- Peirce rejected the idea of a stable relationship between a signifier and its signified. Departing from language-based motivations, he introduced a notion of sign motivated largely by philosophical logic [17]. His main focus was on proposing a theory of production of meaning instead of a theory of language per se. The result is the notion of sign that establishes meaning by recursive relationships between three sets, corresponding to three basic semiotic elements:

- representamen-- the symbolic representation of the denoted object or idea (essentially the Saussure's signifier);
- object --- the thing being represented by the sign;
- interpretant --- the meaning of the sign, represented by yet another sign determined by the process of interpretation.

The relations between the three sets of semiotic elements present the ways how the meaning of a sign is associated with its actual representation in the language and in the world. The main tools employed in the investigations, which aims for lexical semantics, are lexical relations like synonymy, antonymy, hyponymy or hyperonymy ("sameness", "oppositeness", "being a subtype of" or "being a super-type of" relations, respectively). The meaning of lexical units is usually determined in a top-down way by human experts (lexicographers) after studying relevant language resources (e.g., corpora). The meaning itself is constructed by empirical analysis of various general patterns appearing between words in the large scale data sets. The approach of statistical, or distributional semantics is essentially a bottom-up and can be automates to large extent.

Analysis of the meaning of single words or phrases is only the first step towards studying the semantics of more complex natural language structures like sentences. The meaning of a sentence is analyzed by parsing it into its syntactic tree first. The components of the parse tree are then transformed into a logical form, which is in turn used for the sentence's logical analysis by means of associated truth conditions (i.e., the interpretations that render the logical form of the sentence true). The particular formalisms applied to the analysis of parsed natural language sentences usually stem either from first order predicate logic (as described in [18]), or from typed lambda calculus (elaborated in the Montague grammar [19] or in its extension, transparent intentional logic [20]).

In computer science, meaning of semantics is studied mainly from two distinct perspectives – firstly regarding formal semantics of programming languages and secondly with respect to meaning in computational knowledge representation systems. The former allows for studying the meaning of programs or functions (in terms of executed procedures or computed results) regardless of their syntactical representations. The latter branch of semantics in computer science, which is directly related to the focus of our thesis, is concerned with machine-readable representations of knowledge about a real-world domain of interest. This is related to the assignment of an actionable (i.e., comprehensible by computers) meaning to the representations. This can consequently be utilized to infer new implicit facts from the explicitly stored knowledge.

III. WEB MINING

Web mining is a very interesting research topic which combines two of the activated research areas: Data Mining and World Wide Web. With the huge amount of information available online, the World Wide Web is a fertile area for data mining research. The Web mining research relates to several research communities, such as database, information retrieval, and AI. The World Wide Web (Web) is a popular and interactive medium to disseminate information today. The Web is huge, diverse, and dynamic and thus raises the scalability, multimedia data, and temporal issues respectively. It was Oren Etzioni who first coined the term Web mining in his paper in 1996. Etzioni starts by making a hypothesis that the information on the Web is sufficiently structured and outlines the subtasks of Web mining [1] and describes the Web mining processes. Web data mining can be defined as the discovery and analysis of useful information from the WWW data.

Since then, there have been several works around the survey of data mining on the Web. Although Web mining puts down the roots deeply in data mining, it is not equivalent to data mining. The unstructured feature of Web data triggers more complexity in the process of Web mining. An exponential growth in on-line information combined with the almost unstructured web data necessitates the development of powerful yet computationally efficient web data mining tools [2]. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services [1]. This area of research is so huge today partly due to the interests of various research communities, the tremendous growth of information sources available on the Web and the recent interest in ecommerce.

IV. CHALLENGES AND FUTURE TRENDS

The Web presents new challenges to the traditional data mining algorithms that work on flat Data. We have seen that some of the traditional data mining algorithms have been extended or new algorithms have been used to work on the Web data. With explosive growth of the information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tool in order to find the required information resources, and to track and analyze their usage patterns. These factors give rise to the necessity of creating server side and client side intelligent systems that can effectively mine for knowledge. The analysis of large web log files is a complex task not fully addressed by existing web access analyzers. However, it is hard to find appropriate tools for analyzing raw web log data to retrieve significant and useful information. There are several commercially available web log analysis tools, but most of them are disliked by their users and considered too slow, inflexible, expensive, difficult to maintain or very limited in the results they can provide.

While some tools using data mining techniques to help web log analyses are being developed, the research is still in its infancy. The existing techniques for analyzing web usage have different drawbacks, i.e., either huge storage requirements, excessive I/O cost, or scalability problems when additional information is introduced into the analysis.

Most of the currently available Web server analysis tools provide only explicitly and statistical information without real useful knowledge for Web managers. The task of mining useful information becomes more challenging when the Web traffic volume is enormous and keeps on growing. The potential of using a website as a data collection tool for web based information systems is enormous. This is because of its interactive nature, simplicity and unobtrusiveness. The results of the data mining would ideally be integrated into the dynamic website to provide an automated, end-to-end functional system for target marketing and customer relationship management. Most of the web mining tools are evolving and the present web mining techniques still have rooms for improvement to make them prevail in the web based information systems. Some problems like the need for greater integration, scalability issue, and the need for better mining tools are frequently mentioned by many researchers.

The sharpening on the mining tools in many different aspects is important for the future development in this area:

- Web usage mining must handle the integration of offline data with e-business analytic tools, RDBMS, catalogs of products and services and other applications.
- Some new variables or logs should be sought that can be used for finding more natural, meaningful and useful patterns.
- New tools are needed which will not use up too much resources or process time during the web mining process.
- There will always be a need to have benchmark tests to improve the performance of mining algorithms, as the efficiency and effectiveness of a mining algorithm can be measured and a better tool for web data mining can be derived.
- It is important to improve visualization, as much of the data is unorganized and difficult for the user to understand.

V. **CONCLUSION**

Designing and maintaining web based information systems, such as Web sites, is a real challenge. On the Web, it is much easier to find inconsistent pieces of information than a well structured site. There is a strong relation between structured documents (such as Web sites) and a program; the Web is a good candidate to experiment with some of the technologies that have been developed in software engineering.

Web mining is a new and rapidly developing research and application area. With more collaborative research across different disciplines like database, artificial intelligence, statistics and marketing, we will be able to development web mining applications that are very useful to the web based information systems. Web Mining has been an important topic in data mining research in recent years from the standpoint of supporting human-centered discovery of knowledge. The present day model of web mining suffers from a number of shortcomings as listed earlier. As services over the web continue to grow, there will be a continuing need to make them robust, scalable and efficient.

REFERENCES

[1] O. Etzioni, The World-Wide Web: quagmire or gold mine?, *Communications of the ACM*, 39(11), 1996, 65-68 .
[2] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E. Lim, Research Issues in Web Data Mining, *Data Warehousing and Knowledge Discovery*, 1999, 303-312.
[3] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison Wesley.
[4] Berners-Lee, T., Hendler, J., and Lassila,O. (2001).The semantic web. Scientific American,5.
[5] Ogden, C. K. and Richards, I. A. (1989). The Meaning of Meaning. Mariner Books.
[6] Manola, F. and Miller, E. (2004). RDF Primer. Available at (November 2008): http://www.w3.org/TR/rdf-primer/.

[7] Meersman, R. (2001). Ontologies and databases: More than a fleeting resemblance.
[8] Hein, J. and Hendler, J. (2000). Dynamic ontologies on the web. In Proceedings of AAAI 2000, AAAI Press.

[9] Haase, P., van Harmelen, F., Huang, Z., Stuckenschmidt, H., and Sure, Y. (2005). A framework for handling inconsistency in changing ontologies. In Proceedings of ISWC'05, volume 3792 of LNCS, pages 353{367. Springer.
[10] Bobillo, F. and Straccia, U. (2008). fuzzyDL: An expressive fuzzy description logic reasoner. In In Proceedings of FUZZ-08.
[11] Peng, Y., Ding, Z., and Pan, R. (2005). BayesOWL: A probabilistic framework for uncertainty in semantic web. In Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI05).
[12] Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., and Stein, L. A. (2004). OWL Web Ontology Language Reference. Available at (February 2006): http://www.w3.org/TR/ owl-ref/.
[13] Motik, B., Grau, B. C., Horrocks, I., Wu, Z., Fokoue, A., and Lutz, C. (2008). OWL 2 Web Ontology Language: Profiles. Working draft, available at http://www.w3.org/TR/owl2-profiles as of Dec 11, 2008.
[14] Manola, F. and Miller, E. (2004). RDF Primer. Available at (November 2008): http://www.w3.org/TR/rdf-primer/.
[15] Cruse, A. (2004). Meaning in Language: An Introduction to Semantics and Pragmatics, Oxford University Press.
[16] de Saussure, F. (1983). Course in General Linguistics, Open Court, La Salle, Illinois.
[17] Peirce, C. S. (1960). Collected Papers of Charles Sanders Peirce, Harvard University Press.
[18] Kamp, H. and Reyle, U. (1993). From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory, Springer.
[19] Dowty, D., Wall, R., and Peters, S. (1981). Introduction to Montague Semantics, Kluwer Academic Publishers.
[20] Tichy, P. (1988). The Foundations of Frege's Logic, de Gruyter, Berlin, New York.
[21] Alpert, J. and Hajaj, N. (2008). We knew the web was big. Available at http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html (April, 2009).
[22] Flew, T. (2008). New Media: An Introduction. Oxford University Press: Australia.

BIOGRAPHY

M. Venu Gopalachari received B.Tech in CSE from Jawaharlal Nehru Technological University, Hyderabad, India in the year 2005 and M.Tech in CSE from JNTU, Kakinada, India in the year 2008,He is currently pursuing Phd from JNTU, Hyderabad, India. His main areas of interest are web mining, cluster computing.

Dr. P. Sammulal received his Doctorate from Osmania University, Hyderabad in the year 2010. He completed his M.Tech from JNTU University, He received his B.E. degree from Osmania University, Hyderabad. He published more than 20 papers in various topics. His research interests includes cluster computing, distributed computing and web mining and image processing.