International Journal of Research in Advent Technology (IJRAT)
Vol. 1, No. 1, August 2013, ISSN: 2321–9637

A REVIEW ON INFORMATION REPRESENTAION AND RETRIVAL IN SEMANTIC WEB

Parag D. Thakare
Department of Computer Engineering,
JCOET, Yavatmal (445001),
paragthakare2003@yahoo.co.in

ABSTRACT:
Nowadays, large quantity of data is being accumulated in the data repository. For thousands of years people have realized the importance of archiving and finding information. With the advent of computers, it became possible to store large amounts of information; and finding useful information from such collections became a necessity. The huge increase in the amount and complexity of reachable information in the World Wide Web caused an excessive demand for tools and techniques that can handle data semantically. The current practice in information retrieval mostly relies on keyword-based search over full text data, which is modeled with bag-of-words. However, such a model misses the actual semantic information in text. The purpose of Web mining is to develop methods and systems for discovering models of objects and processes on the World Wide Web and for web-based systems that show adaptive performance. The integration of the two fast-developing scientific research areas Semantic Web and Web Mining is known as Semantic Web Mining. The huge increase in the amount of Semantic Web data became a perfect target for many researchers to apply Data Mining techniques on it. In this paper we study the techniques to represent the semantic web information and semantic web mining..

Keywords: *Semantic Web, RDF, Semantic Indexing, Ontology, semantic web mining.*

1. INTRODUCTION

The huge increase in the amount and complexity of reachable information in the World Wide Web caused an excessive demand for tools and techniques that can handle data semantically. Most people believe they can easily find the information they're looking for on the Web[1]. They simply browse from the prelisted entry points in hierarchical directories (like yahoo.com) or start with a list of keywords in a search engine. However, many Web information services deliver inconsistent, inaccurate, incomplete, and often irrelevant results. For many reasons, existing Web search techniques have significant deficiencies with respect to robustness, flexibility, and precision. The disadvantage of the traditional search can be overcome with the proposal of semantic web. Semantic web also called the intelligent web or next generation web. Semantic web is approach towards understand the meaning of the contents. Semantic information is stored in the form of ontologies.To deal with this issue; ontologies are proposed [5] for knowledge representation, which are nowadays the backbone of semantic web applications. Both the information extraction and retrieval processes can benefit from such metadata, which gives semantics to plain text. 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. There are so many techniques to represent the semantic web information and the data mining techniques to retrieve the information from the semantic web. The semantic indexing and the SPARQL used to retrieve the data from the semantic web. These techniques improve the traditional search results.
The rest of the paper organized as follows, in section 2, brief discussion about the related work. In section 3, about the Semantic Web Representation Techniques. In section 4, semantic web mining. In section 5, conclude the paper and remark for future work.

2. LITERATURE REVIEW

Semantic Web is about providing meaning to the data from different kinds of web resources to allow the machine to interpret and understand these enriched data to precisely answer and satisfy the web users' requests [1]. the Semantic Web is introduced to crack two specific problems: the limitations of data access in the web (for example retrieving documents according to a given request using ambiguous terms, and the current retrieving systems problem of acquiring only a single "best fit" documents for a query), and the delegation tasks' problems (such as integrating information) by supporting access to data at web-scale and enabling the delegation of certain classes of tasks. The fundamental problem of understanding intelligence is not the identification of a few powerful techniques, but rather the question of how to represent large amounts of knowledge in a fashion that permits their effective use.

The Semantic Web is changing the way how scientific data are collected, deposited, and analyzed [4]. In this paper a detailed state-of-the-art survey of ongoing research in Semantic Web Mining has been presented. This study analyzes the merging of trends from both areas including a) using semantic structures in the Web to enrich the results of Web Mining and b) to build the Semantic Web by employing the Web Mining techniques.

A new generation of intelligent search engines incorporates Web semantics and uses more advanced search techniques based on concepts such as machine learning. These approaches enable intelligent Web information services, personalized Web sites, and semantically empowered search engines.

The idea behind the Semantic Web is to augment Web pages with mark-up that captures some of the meaning of the content on those pages[3]. Automatic tools can collect and "understand" the knowledge annotated on a page, and ontologies help make such mark-up compatible across various information sources and queries. An *ontology* is an explicit specification of a vocabulary for a domain, and it includes definitions of classes, relations, functions, and constraints. Because the range and diversity of data on the Web is too extensive, most ontologies are domain-specific or personalized to express the specific interests of individuals or communities.

The semantic web is based on the visualization of Tim Berners-Lee [2], the inventor of the World-Wide-Web (WWW). According to him, "The semantic web is not at all visualized as a separate web but it is an expansion of the existing one, in which information is given well-defined sense and significance, better enabling PCs and people to work in cooperation." Figure 1 shows architecture for semantic web mining.

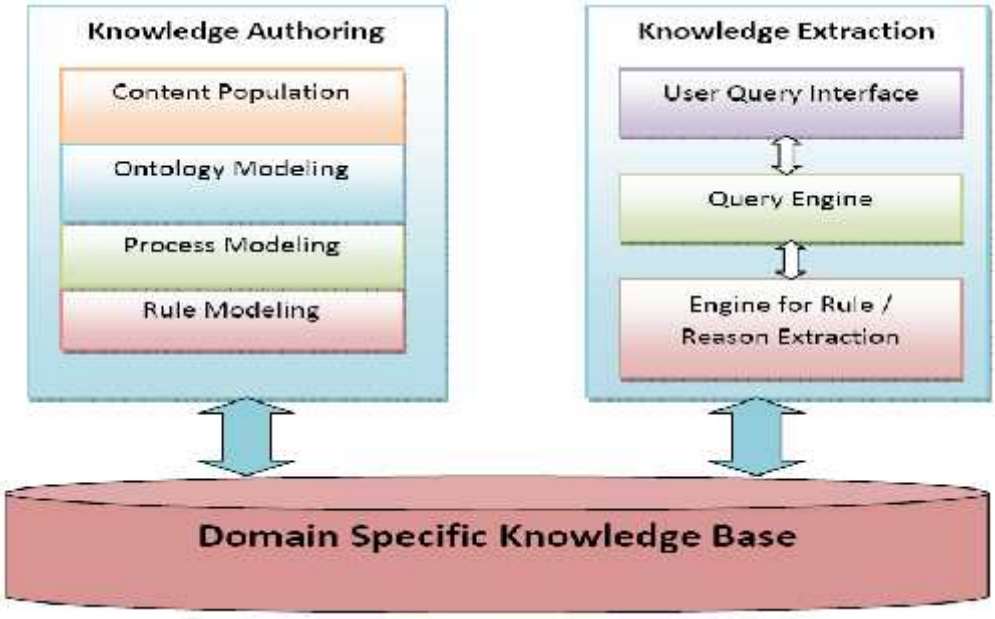

Figure 1: Semantic Web Solution Architecture

The architecture is primarily divided into three logical modules, namely knowledge extraction block, authoring block and domain specific knowledge base. The availability of current search engines has to a great extent improved our ability to carry-out a meaningful data search on the web. But, such search option is still primarily restricted to structured data. In semantic technology, the focus is generally to formulate flexible data model (called Triples) from the user friendly domain query. There are many challenges when it comes to mining of the

unstructured data in the context of semantic web: No standardized web form structure, Non-availability of standard Semantics, Lack of global Standards, Lack of proven frameworks, Non-standard implementation and Less-explored Ontology framework.

Semantic indexing is one of the techniques to improve precision and recall values[8]. Semantic index is created on ontological data and keyword search is performed. Semantic indexing approach is based on the lucene indexing. SPARQL can be used to query ontological data. Problem with SPARQL is that, it is not made for end users as we require the knowledge of domain ontology and syntax of the language. Therefore, semantic web community works on simplifying the process of query formulating for the end user.

3. SEMANTIC WEB REPRESENTATION TECHNIQUES

3.1 *Extensible Markup Language*

The Extensible Markup Language (XML) technique has been established as a generic technique to store, organize, and retrieve data on/from the web. By enabling users to create their own tags, it allows them to define their content easily. Therefore, the data and its semantic relationships can be represented

3.2 *Resource Description Framework*

The Resource Description Framework (RDF) is a common language that enables the facility to store resources' information that are available in the World Wide Web using their own domain vocabularies. Three types of elements contented in the RDF: resources (entities identified by Uniform Resource Identifiers URIs), literals (atomics values like strings and numbers), and properties (binary relationships identified by URIs) [3]. This is a very effective way to represent any kind of data that could be defined on the web.

RDF representation of the metadata of the article might include the date when it was published which is not described in the article. Assuming the article is put at http://www.ei.sanken.osaka-u.ac.jp/pub/WI2001-Miz.pdf, the RDF description would be:

<rdf:Description rdf:about="http://www.ei.sanken.osaka-u.ac.jp/pub/WI2001-Miz.pdf">
<author>Riichiro Mizoguchi</author>
<pub-date>2001-10-23</pub-date>
</rdf:Description>

Although RDF has been designed for metadata representation model, it can be used as a general-purpose knowledge representation, which might be apparent from the fact that it is a kind of semantic network model.

3.3 *Web Ontology Language*

The Web Ontology Language (OWL) is considered a more complex language with better machine-interpret- ability than RDF. It precisely identifies the resources' nature and their relationships [8]. To represent the Semantic Web information, this language uses ontology, a shared machine-readable representation of formal explicit description of common conceptualization and the fundamental key of Semantic Web Mining. Ontology creators are expressing the interest domain which is based on classes, and properties (represent atomic distinct concepts and rules in other semantic languages respectively) [9]. Here we propose a preliminary set of design criteria for ontologies whose purpose is knowledge sharing and interoperation among programs based on a shared conceptualization.

1. Clarity: An ontology should effectively communicate the intended meaning of defined terms. Definitions should be *objective.* While the motivation for defining a concept might arise from social situations or computational requirements, the definition should be independent of social or computational context. *Formalism* is a means to this end. When a definition can be stated in logical axioms, it should be. Where possible, a *complete* definition (a predicate defined by necessary and sufficient conditions) is preferred over a partial definition (defined by only necessary or sufficient conditions). All definitions should be documented with natural language.

2. Coherence: An ontology should be coherent: that is, it should sanction inferences that are consistent with the definitions. At the least, the defining axioms should be logically consistent. Coherence should also apply to the concepts that are defined informally, such as those described in natural language documentation and examples. If a sentence that can be inferred from the axioms contradicts a definition or example given informally, then the ontology is incoherent.

3. Extendibility: Ontology should be designed to anticipate the uses of the shared vocabulary. It should offer a conceptual foundation for a range of anticipated tasks, and the representation should be crafted so that one can extend and specialize the ontology *monotonically*. In other words, one should be able to define new terms for

special uses based on the existing vocabulary, in a way that does not require the revision of the existing definitions.

4. Minimal encoding bias: The conceptualization should be specified at the knowledge level without depending on a particular symbol-level encoding. An *encoding bias* results when representation choices are made purely for the convenience of notation or implementation. Encoding bias should be minimized, because knowledge-sharing agents may be implemented in different representation systems and styles of representation.

5. Minimal ontological commitment: An ontology should require the minimal ontological commitment sufficient to support the intended knowledge sharing activities. An ontology should make as few claims as possible about the world being modeled, allowing the parties committed to the ontology freedom to specialize and instantiate the ontology as needed. Since ontological commitment is based on *consistent* use of *vocabulary*, ontological commitment can be minimized by specifying the weakest theory (allowing the most models) and defining only those terms that are essential to the communication of knowledge consistent with that theory.

The main reasons for developing an ontology are to share a common understanding of the structure of information among people or software agents, to enable reuse of domain knowledge[12]. Following figure shows the Ontology development process. The general stages in the design and development of an ontology are as follows.

- The first step involves determining the domain and source and also purpose and scope of the ontology. Questions that should be addressed at this stage include: what domain will the ontology cover?, what is the purpose of the ontology? and for what sorts of questions should the information in the ontology be able to provide answers?
- The second step is to ascertain if an ontology has been developed previously in the same subject area. If such an ontology exists, it is easier to modify the existing ontology to suit ones needs than to create a new ontology.Reusing existing ontologies may also be a requirement if the system needs to interact with other applications that have already committed to particular ontologies.
- The third step is to enumerate important terms in the ontology.

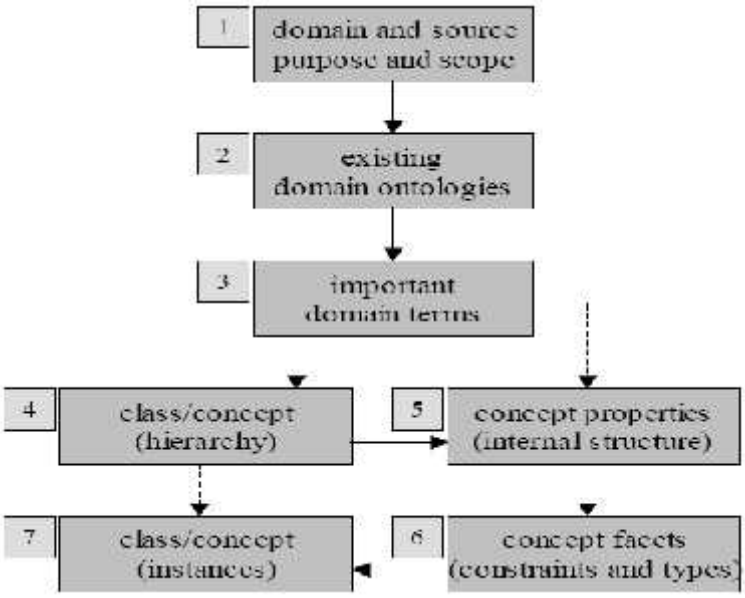

Figure 2: Ontology development process.

- Steps 4 and 5 are closely intertwined. They entail defining the classes (concepts) and the class hierarchy (Step 4), and defining the properties of classes (Step 5).
- Step 4. A number of different approaches can be taken when determining the hierarchy of classes. One could use a top-down approach, which starts with the definition of the most general concepts in a domain and continues with more specialized concepts. Another approach is the bottom-up approach, which starts with the definition of the most specific classes (the leaves of the hierarchy), with subsequent grouping of these classes into more general concepts. From the list of terms drawn up in Step 3, those terms that describe objects that have an independent existence should be extracted as these

will form the classes (concepts) of the ontology. Todetermine the hierarchical organization of the ontology, for each class one should ask if the instances of thatclass could also be instances of a more general class. If the answer is yes, then this class constitutes a subclass ofthe other class and, hence, is further from the root concept in the ontology.

- Step 5. Once the classes have been defined, the next step is to describe the internal structures (properties) of the concepts. Again, these should be readily available from the list produced as a result of Step 3.
- Step 6 involves attaching facets to the properties, that is, describing the value type, allowed values, the number of allowed values (cardinality) and other features that are deemed to be necessary. In this way, constraints are placed on the types of data that are allowed.
- The final step 7 in the procedure is to create instances of the classes, that is to provide examples of each of the classes.

For example, if one wants to add some constraints, he/she has to use OWL. The following OWL code shows a constraint stating Conference Paper and Journal Papers are mutually exclusive.

```
<owl:Class rdf:ID="JournalPaper">
<rdfs:subClassOf rdf:resource="#Paper"/>
<owl:disjointWith rdf:resource="#ConferencePaper"/>
</owl:Class>
```

4. TECHNIQUES FOR SEMANTIC WEB MINING

Various techniques for Semantic Web mining are web content mining, web usage mining and web structure mining.

4.1 *Web Content Mining*

A well -known problem, related to web content mining, is experienced by any web user trying to find all and only web pages that interests him from the huge amount of available pages. Current search tools suffer from low precision due to irrelevant results. Search engines are not able to index all pages resulting in imprecise and incomplete searches due to information overload. The overload problem is very difficult to cope as information on the web is immensely and grows dynamically raising scalability issues. Agents may be used for intelligent search, for classification of web pages, and for personalized search by learning user preferences and discovering web sources meeting these preferences. Web content mining is more than selecting relevant documents on the web. Web content mining is related to information extraction and knowledge discovery from analyzing a collection of web documents. Related to web content mining is the effort for organizing the semi-structured web data into structured collection of resources leading to more efficient querying mechanisms and more efficient information collection or extraction. This effort is the main characteristic of the “Semantic Web” [6], which is considered as the next web generation. Semantic Web is based on “ontologies”, which are meta-data related to the web page content that makes the site meaningful to search engines.

4.2 *Web Usage Mining*

Web usage mining research focuses on finding patterns of navigational behavior from users visiting a website. These Patterns of navigational behavior can be valuable when searching answers to questions like: How efficient is our website in delivering information? How the users perceive the structure of the website? Can we predict user’s next visit? Can we make our site meeting user needs? Can we increase user satisfaction? Can we targeting specific groups of users and make web content personalized to them? Answer to these questions may come from the analysis of the data from log files stored in web servers. Web usage mining has become a necessity task in order to provide web administrators with meaningful information about users and usage patterns for improving quality of web information and service performance [7]. Successful websites may be those that are customized to meet user preferences both in the presentation of information and in relevance of the content that best fits the user.

4.3 *Web Structure Mining*

Web structure mining is closely related to analyzing hyperlinks and link structure on the web for information retrieval and knowledge discovery. Web structure mining can be used by search engines to rank the relevancy between websites classifying them according to their similarity and relationship between them [8]. Google search engine, for instance, is based on Page Rank algorithm [9], which states that the relevance of a page increases with the number of hyperlinks to it from other pages, and in particular of other relevant pages. Personalization and recommendation systems based on hyperlinks are also studied in web structure mining. Web

structure mining is used for identifying “authorities”, which are web pages that are pointed to by a large set of other web pages that make them candidates of good sources of information. Web structure mining is also used for discovering community networks by extracting knowledge from similarity links. The term is closely related to “link analysis” research, which has been developed in various fields over the last decade such as computer science and mathematics for graph-theory, and social and communication sciences for social network analysis.

5. CONCLUSION

Semantic web is a next generation of web search. Semantic web information can be represent in the form of ontologies. Ontology represents the knowledge in term of classes and subclasses. Knowledge represented in ontology can be interpreted by machine. Machine can add more data and relation on behalf of users. In this paper we studied the different techniques for semantic web representation and semantic web mining. The semantic indexing and SPARQL used for the information retrieval on domain specific information and try to improve the traditional search performance. The concept can be used to build ontology for different domain which collects the data from different language repository and also find the best fit data mining techniques that satisfy the user requirement.

REFERENCES

[1] Soner Kara,Ozgur Alan,Orkunt Sabuncu(2010).“An Ontology-Based Retrieval System Using Semantic Indexing“,ICDE Workshops 2010,978-1-4244-6523-1/10.

[2] Manoj Manuja & Deepak Garg.“ Semantic Web Mining of Un-structured Data: Challenges and Opportunities“, International Journal of Engineering (IJE), Volume (5) : Issue (3) : 2011

[3] Thomas R. Gruber. “Toward principles for the design of ontologies used for knowledge sharing”, Int. J. Hum.-Comput. Stud., 43(5-6):907–928, 1995

[4] Yuangui Lei, Victoria S. Uren, and Enrico Motta.” Semsearch: A search engine for the semantic web”, In EKAW, pages 238–245, 2006.

[5] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. “Introduction to Information Retrieval”, Cambridge University Press, New York, NY, USA, 2008.

[6] Doruk Tunaoglu, Ozgur Alan, Orkunt Sabuncu, Samet Akpinar, Nihan K. Cicekli, and Ferda N. Alpaslan. “Event extraction from turkish football web-casting texts using handcrafted templates”, In In Proc. of Third IEEE Inter. Conf. on Semantic Computing (ICSC)(in press), 2009.

[7] Haofen Wang, Kang Zhang, Qiaoling Liu, Thanh Tran, and Yong Yu. “Q2semantic: A lightweight keyword interface to semantic search”, In ESWC, pages 584–598, 2008.

[8] S.M.Patil,D.M.Jadhav (2012).”semantic information retrieval using ontology and SPARQL for cricket”,International Journal of Advances in Engineering & Technology,ISSN:2231-1963 vol.4,issue 2,pp.354-363

[9] Gerard Salton, A. Wong, and C. S. Yang. “A vector space model for automatic indexing”, Commun. ACM, 18(11):613–620, 1975.

[10] The semantic web wiki, http://semanticweb.org, 2008.

[11] Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarrin.” Indexing with wordnet synsets can improve text retrieval”, pages 38–44, 1998.

[12] Soner Kara, Ozg ¨ ur Alan, Orkunt Sabuncu, Samet Akpınar, Nihan K. C¸ ic¸ekli, and ¨ Ferda N. Alpaslan.” An ontology-based retrieval system using semantic indexing”, In 1st International Workshop on Data Engineering meets the Semantic Web (DESWeb’2010) (co-located with ICDE’2010), November 2010.

[13] Boyce, S., & Pahl, C. (2007).” Developing Domain Ontologies for Course Content“, Educational Technology & Society, 10 (3),275-288.