Web to Semantic Web and Role of Ontology in its Development

Zeeshan Ahmed
University of Wuerzburg, Am Hubland, D 97074, Germany
Email: zeeshan.ahmed@uni-wuerzburg.de

Abstract. This research is mainly about to discuss World Wide Web leading to the need of the proposition of a new concept i.e. Semantic Web. Concerning Semantic Web, the conducted research provides information for the implementation of semantic based web applications by providing the concept of structuring of data over the web to take advantage in extracting semantic based information. Going into the details, this paper presents Ontology as the main building block of Semantic Web of present time, its supporting technologies i.e. XML, RDF and OWL, and some existing limitations with respect to its use in real time web application development. Furthermore, before concluding the discussion some Semantic Web based applications are presented which are developed with the use of Ontology and providing lots of values in the implementation of semantic based applications by providing structured data over the web to take advantage in implementing efficient web based information retrieval search mechanisms.

Keywords: Web, Semantic Web, Ontology

1. Introduction

Targeting the challenge of implementing a web based system capable of performing semantic based search to extract desired information from attached repositories over the web, the field of Web and Semantic Web is explored, as it promotes the implementation of semantic based web applications by providing the concept of structuring of data over the web to take advantage in extracting semantic based information. World Wide Web is a global information sharing and communication system made up of three standards Uniform Resource Identifier (URL), Hypertext Transfer Protocol (HTTP) and Hypertext Mark-up Language (HTML) by Tim Berners-Lee to effectively store, communicate and share different forms of information. The Information is provided over the web in text, image, audio and video formats using HTML, considered unconventional in defining and formalizing the meaning of the context.

Most of the data is structured only inside the available databases over the web and due to this it is quite easy to go for scattered extensive information by looking into bookmarked web pages but quite difficult to extract a piece of needed information. Although some search engines and screen scrapers are invented, search engine uses full text query to search information but can only return unstructured contents not the actual structured information stored in database on web where as screen scrapers extracts and repurpose fragments from web pages but insufficient in creating a rich multi domain information environment [10]. Most of the search engines are not satisfactory because they require excessive manual pre-processing e.g. designing a schema, cleaning raw data, manually classifying documents into taxonomy and manual post processing e.g. browsing through large result lists with too many irrelevant items [11]. To increase the integration and interoperability over the web the concept of “Web Service” was introduced. Due to the dynamic nature web services became very famous in industry in short time but with the passage of time due to the heavily increase in number of web services end-to-end service authentication, authorization, data integrity and confidentiality problems were identified which are still alive and not handled by existing web technologies [12].

HTML documents are formatted such that these cannot be processed semantically because these are only available in a readable format. This deficiency
leads to the problems of searching, extracting, maintaining, uncovering and viewing the knowledge based information over the web. More over this format deficiency becomes the major cause of some semantic problems and the need of some other approach which will publish data over the web in not only the readable but also in a processable format. Because, if data will be available in Meta data (readable and processable data format) then it will improve the process of search, extraction and maintenance of data over the web. To take advantage of interactive information sharing, interoperability and user centered design web application development Web 2 was introduced, which then improved to the concept of Web 3 to include transformation of the Web into a database to provide accessibility of the contents by multiple non browser applications. Then continuing the streak of advancement in existing web and to cope with the currently existing web problems i.e., Information filtration, security, confidentiality and augmentation of meaningful contents in mark-up presentation, the concept of “Semantic Web” was proposed by Tim Berners Lee [4]. Semantic Web is also renowned as the modified version of Web3.

2. Web to Web 3

To take advantage of interactive information sharing, interoperability and user centered design web application development Web 2 was introduced [23], which then improved to the concept of Web 3 to include transformation of the Web into a database [24] to provide accessibility of the contents by multiple non browser applications as shown in Figure 1.

![Figure 1: Web to Semantic Web](image)

Then continuing the streak of advancement in existing web and to cope with the currently existing web problems i.e., Information filtration, security, confidentiality and augmentation of meaningful contents in mark-up presentation, the concept of “Semantic Web” was proposed by Tim Berners Lee [4]. Semantic Web is also renowned as the modified version of Web3. To conclude with one major current web based technology a feature based comparison is performed between Web, Web2 and Web3 and presented in Table 1 [25], [26]. Comparison shows the feature dominancy of web 3 over Web and Web 2.

<table>
<thead>
<tr>
<th>Features</th>
<th>Web</th>
<th>Web2</th>
<th>Web3</th>
</tr>
</thead>
<tbody>
<tr>
<td>With readable contents</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Client Server based architecture</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Top down data presentation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Text Based Interface</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Peer to Peer</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Writable Interface</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Bottom Up data presentation</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Video Based</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Use Web Services</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Interact and communicate with other webmasters and sites</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Exchange of information</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Portable and Personal</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Widgets with flexible GUI options like drag and drop facilities etc.</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>With Dynamic Contents</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

3. Semantic Web

Semantic Web is a mechanism of presenting information over the web in a format so that human being as well as machines can understand the semantic of context. Semantic web is a mesh of information which can be linked up in a way, so that it can easily be processed by machines [1] and aim to produce technologies capable of reasoning on semi structured information [5]. The semantic web is an intelligent incarnation and advancement in World Wide Web to collect, manipulate and annotate information independently by providing effective access to the information. Semantic web provides categorization and uniform access to resources, promoting the transformation of World Wide Web into semantically modeled knowledge representation systems and common framework which allows data to be shared and reused [13]. Semantic web also gives the concept of semantic based web services to provide solutions to the problems of dynamically composed service based applications.

Currently, semantic web is standing on a very important building block Ontology [2]. Moreover semantic web aims at providing information in machine processable semantic models which assigns informa-
tion resources to classes whose meaning is defined in ontologies [14], a collection of interrelated semantic concepts. Ontology is the explicit representation and description of already available finite sets of terms and concepts used to make the abstract model of a particular domain, described. Moreover, along with the processing ability semantic web agent is capable of communicating, receiving and transferring information to different sources (agent or human).

The main and currently not achieved goal of semantic web is to structure the meaningful contents of unstructured published data over web to take advantage in improving the search process [4] and to involve knowledge management in making some more advanced knowledge modeled management systems. No doubt semantic web using ontology has contributed in the progress of web but still there are some limitations and due to them semantic web is currently not successful in attaining the actual goal of completely structuring the information over the web which can be processed by machines and making advanced knowledge modeled system. The need is to enhance the concept of ontology with respect to development point of view because all the theories can be fruitful if the implementation is possible.

4. Ontology

Ontology is a main building block of Semantic Web to provide the information in machine processable semantic models and produce semantically modeled knowledge representation systems. It is playing a vital role in solving the existing web problems by producing semantic aware solutions. Ontology makes machines capable of understanding the semantic of languages that humans use and understand by producing the abstract modeled representation of already defined finite sets of terms and concepts involved in intelligent information integration and knowledge management [15]. Ontology is basically categorized in three different categories i.e., Natural Language Ontology (NLO), Domain Ontology (DO) and Ontology Instance (OI) to provide relationships between generated lexical tokens of statements based on natural language, knowledge of a particular domain and to generate automatic object based web pages [8]. Ontologies are constructed and connected to each other in a decentralized manner to clearly express semantic contents and arrange semantic boundaries to find out required needed information [16].

Natural language based information is treated as the input to the ontology construction process, which parses the text in nouns and verbs. Nouns are represented as “Classes” and verbs as “Properties” containing values, relationships with other properties and some constraints. Classes are further divided in main and sub class categories maintained in taxonomical hierarchy. The size of ontology varies due to the increase in number of classes and instances.

Ontologies can be made manually from scratch, by extracting information from web and by merging already existing ontologies into new ontologies. But this manual process sometimes becomes very complex and time consuming especially when dealing with the large amount of data. Moreover, to support the process of semantic enrichment reengineering for the building of web consisting of meta data depends on the proliferation of ontologies and relational meta data. This requires high production of meta data at high speed and low cost. So in these cases machine learning approaches can be very helpful in generating ontologies automatically because they provide real time schemes like classification rules, instance based learning, numeric predictions, clustering, Bayesian networks and decision trees which can be very helpful in the generation of ontologies.

Ontology development is an iterative process based on six main activities i.e., Determine Scope, Enumerate Terms, Classify Ontology, Define Classes, Define Properties and Create Instances as shown in Figure 2. In the beginning of an ontology development process it is very important to determine the scope otherwise it will be very time and effort consuming. Then enumerated terms need to be identified to classify ontologies within their respective types. Classes and their respective properties along with their relationships and constraints are defined using identified enumerated terms. In the end only the instances are created and used. To implement ontology development process some experience, a powerful user friendly ontology supporting tool and communication between domain experts and developers is required e.g. example LibraryWine etc [9].
5. Ontology Supporting Languages

First step in building ontologies is to create the nodes and edges. Once the concepts and relationships of graph based ontology are constructed then next step is to quantify the strengths of semantic relationships [11]. Ontologies can be constructed manually and automatically by using some ontology supporting languages i.e., XML (eXtensible Mark-up Language), RDF (Resource Description Framework) [1] and OWL (Web Ontology Language) offering ways of more explicitly structuring and richly annotating Web pages.

5.1. XML

XML is one of the fundamental contributions towards middleware technologies [7]. It is a markup Meta language which allows sharing of information between different applications through markup, structure and transformation. As the major contribution towards semantic web, XML uses Data Type Definitions (DTD) and depends on data types, attributes, both internal and external elements structure documents and provides syntax serialization and abbreviation for data modeling [3]. The XML schema restricts the syntax to be only used for the structured documents, because of this XML has two main problems in process of information extraction; first it is without semantic and second is the arbitrary naming and structuring of elements.

5.2. RDF

RDF a URL based syntax data representation provides a secure and reliable mechanism for the exchange of metadata between web applications. RDF processes Meta data by making an abstract data model based on three object types attributes Resource, Property and Statement. Resource is an expression, Property is an attribute to describe a resource where as the statement is a resource having some property and value. RDF uses three containers, Object bag, Sequence, and Alternative, to keep multiple available and alternative values arranged in an order in resources and properties. Bag contains resources, Sequence contain resources along with their properties having single or multiple values arranged in order and Alternative contains resources having alternate value(s) of a property [6]. RDF provides syntax serialization and abbreviation for RDF data modeling. Serialized syntax expresses the full capabilities of data modeling in a very regular fashion and abbreviated syntax includes additional constructs to provide a more compact form in representing a subset of the data model. RDF is more useful than XML in ontology construction because it provides semantic based features for data including domain independency, vocabulary and privileges in defining terminologies used in schema language. Furthermore it also provides syntax based on reification (statements about statements), data types, attributes, nesting, elements, element types, element container and no restrictions in structuring document like XML. RDF has its own grammar but not complete, relies on the support of XML to fulfill the need. Moreover RDF modeling mechanism is insufficient in expressing various logical statements [3].

5.3. OWL

OWL is derived from American DARPA Agent Markup Language (DAML). OWL is based on ontology, inference and European Ontology Interchange Language (OIL), claims to be an extension in RDF in expressing logical statements. It is rich in vocabulary because it not only describes classes and properties but also provides the concept of namespace, import, cardinality relationship between the classes and enumerated classes. OWL has some limitations like only one “Namespace” per project is allowed, “Import” is not currently supported, no database backend and Multi-User support and a few OWL Language features are missing [9].

To conclude with one most beneficial available Ontology supporting language of all earlier above discussed, a comparison is performed amongst them and presented in Table 2. The comparison concludes with the provision of more features in OWL than RDF and XML.

<table>
<thead>
<tr>
<th>Features</th>
<th>XML</th>
<th>RDF</th>
<th>OWL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middleware</td>
<td>Yes</td>
<td>Yes</td>
<td>yes</td>
</tr>
<tr>
<td>Markup Meta language</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Allows sharing of information between different applications</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Depends on data types</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Restricted syntax schema</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>With Semantic</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Secure and reliable mechanism for the exchange of metadata between web applications

<table>
<thead>
<tr>
<th>Feature</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processes meta data</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Model data</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Domain independent</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Own grammar</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>OIL based</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Express logical Statements</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Rich vocabulary</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

6. Limitations of Ontology

The development of ontology driven applications is difficult because of some limitations and principal problems which are as follows;

1. Natural language parsers used to parse the information to construct the ontologies are limited because they can only work over a single statement at a time [7].
2. Existing methodologies of structuring ontologies are insufficient and need to be improved because now it is quite impossible to define the boundaries of ontology based particular domain’s abstract model and automatically handle the increase in size of ontology due to the increase in number of classes and instances.
3. Creating ontologies manually is a time consuming process which becomes very complex when there is a large amount of data to create large number of ontologies from. To take advantage in creating large number of ontologies by reducing the complexity and time, an automatic ontology creation mechanism is required. Some mechanisms are already proposed and implemented to create ontologies automatically but they are insufficient and less qualitative. While creating nouns based classes using existing automatic ontology creation mechanism, it is quite impossible to identify the possible existing relationships between classes to draw the taxonomical hierarchy [17]. Furthermore it is also quite impossible to perform automatic emergence of ontologies to create new ontologies [19].
4. Currently available ontology validators are restricted and not capable of validating all kind of ontologies e.g. based on complex inheritance relationship [12].
5. Domain specific ontologies are highly dependent on the domain of the application and because of this dependency domain specific ontologies contain specific senses which are not possible to find in general purpose ontology [18].
6. The process of semantic enrichment reengineering for web development consists of relational meta data required to be developed at high speed and in low cost depending on proliferation of ontologies, which is currently also not possible.
7. Handling the dynamically raised calculations caused by the comparison of big complexities of similar ontologies is also not possible [19].

7. Semantic Web Applications

Residing in the domain of Semantic Web many products are available and several approaches have been introduced by many researchers which are providing lots of values in the implementation of semantic based applications with use of Ontology and providing structured data over the web to take advantage in implementing efficient web based information retrieval search mechanism e.g. Semantic Dektop[20], Reisewissen [21] and Meta Data Search Layer[22]. etc In this section, without going in to much product detail, we will only present some semantic web based approaches, to take advantage in having an idea about semantic based system development using Ontology.

7.1. Semantic desktop: Personal Information Model (PIM)

Authors promoted the idea of stepping into user’s mental model by implementing Personal Information Model (PIM). PIM is designed to improve the process for the identification of documents and retrieval of no unnecessary document. The design is based on ontologies and classes, the relationships of classes and ontologies are predefined and the information can be accessed using RDF graphs. Four rules based on forward changing principle are defined to retrieve the information, this information is divided in three parts Author (single or team), Relevant
project, and Relevant solution. The system works in
the following way. First query runs aiming to find
out the project and if project is found then it moves to
find out the related documents of the project.

The proposed architecture mainly consists of
three main components: Receiver, Interpreter and
Analyzer. Receiver is used to provide index services
and obtain the information about the structure of in-
dexed files with the help of so called brainFiller .
Interpreter first retrieves Information (structure / un-
structured) using full text search and then uses so
called LiveLink to structure the contents of obtained
information with the help of manual annotation and
meta data, to store and retrieve contents based on
their properties and preferences. At the last step the
Analyzer queries using Jena inference engine on
created RDF models to infer runs and also uses F-
Logic to integrate rules. To take the advantage of
proposed approach by sharing the information from
search, four case scenarios are designed: Local search,
group search, closed community and open community.
Local search scenario only deals with the search
mechanism and can only be applied to a personal
desktop, group search can be applied with in a particular
network domain, closed community consists a
number of users having different roles but same topic
where as open search consists of users with different
roles and different topics.

7.2. Reisewissen

Reisewissen is proposed to provide quality services
by semantically connecting, organizing and sharing
the isolated pieces of information by transpiercing to
data sources, caching & fetching of data, transform-
ing data from heterogeneous to RDF models, mapping
of ontologies between database and triples, matching RDF and non RDF based information.
Moreover Reisewissen is implemented using seman-
tic web technology based Ontology (RDF), by pro-
ducing RDF models and manually mapping ontolo-
gies during the implementation of search mechanism.

Reisewissen identifies potential relevant knowl-
edge sources and provide quality services by seman-
tically connecting, organizing and sharing the cur-
rently isolated pieces of information in an online por-
tal to anticipating customer behavior. Proposed ap-
proach is implemented using semantic web technolo-
gies in a project Reisewissen, a hotel recommenda-
tion engine and travel information system. The de-
sign of Reisewissen is composed of three main com-
ponents Data Connectors (DC), Evaluation Frame-
work (EF) and Evaluation Engine (EE). Data connec-
tors are used to provide transparency to data sources
and transformation of data from heterogeneous to
common data format (RDF and Java objects), moreover it also provides the caching and fetching of data.
Evaluation Framework is a workbench to test the
quality of data and rules by providing functions and
filters to map resources and return result in decisive
format (Boolean or float value) and Evaluation En-
gine combines individual filters to rank and filter
information by weighting and yielding the overall
score.

Information is obtained using Simple Object
Access Protocol (SOAP) based web services and
stored in both RDF and non RDF formats, which are
then matched to find out the desired result. Data
stored in RDF format is based on developed ontolo-
gies mapped between database and RDF triples.
Moreover Reisewissen uses Prolog to capture ex-
pert’s knowledge which can be formalized and can
generate new data by implementing the customer
request in evaluator encapsulated rules. Data is
matched semantically by combining data properties
to ontology and similarities between two concepts are
determined by distance reflecting their respective
positions in hierarchy. As a result list of selected re-
sults are generated to customer.

7.3. Meta Data Search Layer

Authors have discussed a successful process for meta
data search based on three questions for information
eextraction: What user needs, Where it lies, and How
it can be retrieved. The targeted objective is to identi-
fy the location from set of locations contained by a
document and avoid looking into non-specific docu-
ment. Scalability and efficiency of this approach is
determined using simulation of documented meta
data keywords, location pointers, node connections
and node knowledge. The whole process of identify-
ing target location and search consists of nine proce-
dural steps.

- Select target document from network having
  at least one keyword.
- Use keywords contained in document for the
  construction of a search query.
- Start with free node (not already containing
  any target document).
- Make a record of start node.
- Start node’s knowledge treated as base
  knowledge for the selection of other sub
  nodes.
• If number of nodes is equal to forwarding degree select those nodes.
• If number of nodes is less than forwarding degree select additional nodes.
• If number of nodes is more than equal to forwarding degree selects subset of nodes.
• For each selected node, if node contains target document then update connectivity and if it doesn’t then continue search using nodes.

Authors have explicitly mentioned that this search mechanism is good but still there is a room for improvement in examining the path length of searches for different and same users characterized by their different query distributions. Moreover time to converge to a stable network can be an ambiguity and need to have more realistic simulation using parameters and distributions.

8. Conclusion

In this research paper World Wide Web is briefly discussed. Going into the details of Web and for the implementation of a web based application with the provision of extraction of structured information over the web Semantic Web is discussed. A vital semantic web technology Ontology is also discussed in this chapter in detail including the brief information about its supporting technologies, approaches and limitations. Furthermore a comparison between current ontology supporting technologies is performed to conclude with one i.e., OWL.

References

[21] Magnus Niemann, Malgorzata Machol and Robert Tolksdorf, “Improving Online Hotel Search – What Do We Need Seman-
tic For”, In Proceedings of Semantic Systems from Visions to Applications 2006, Vienna Austria 2006


