Chapter 2

SEMANTIC WEB

2.1 Introduction

The term ‘Semantic’ refers to a sequence of symbols that can be used to communicate meaning and this communication can then affect behavior in different situations. Semantics has been driving the next generation of the web termed as the semantic web, where the focus is on the role of semantics for automated approaches to exploit web resources [11]. Semantic web is being developed to overcome the following main limitations of the current web:

- Ambiguity of information resulting from poor interconnection of information.
- Automatic information transfer is lacking.
- Unable to deal with enormous number of users and content ensuring trust at all levels.
- Incapability of machines to understand the provided information due to lack of a universal format.

Semantic web implies a web that can process information both for humans as well as machines in such a way that a machine can interpret and exchange the information on web, without human interruption, producing more relevant data. The concept of semantic network model [74] was conceived first time in early sixties by cognitive scientist Allan M. Collins, linguist M. Ross Quillian and Elizabeth F [54]. They discussed the concept in the context of how human brain uses long term memory to relate things to ascertain the truth of a sentence. They categorized objects on the basis of their attributes and draw
inference from this categorization to decide whether a statement is true or false [54]. Later on Tim Berners-Lee [11], inventor of the World Wide Web and director of the World Wide Web Consortium ("W3C"), coined the term semantic web in 2001. He emphasized that semantic web will bring structure to the content of web pages, creating an environment where software agents roaming from page to page will readily carry out sophisticated tasks for users [21].

The term semantic web encompasses efforts to build a new WWW architecture that enhances content with formal semantics. This will enable automated agents to reason about web content, and produce an intelligent response to unforeseen situations [14]. Semantic web aims to change web development in such a way that machine can make the sense of words displayed on the web pages, can easily relate them and produce information more relevant to the subject at the ease of surfing. Human are possessed with vocabulary and the contextual information of different words. Our brain can relate incomplete and irrelevant words and still draw conclusions on the basis of knowledge and experience, where as providing same capability to the computers is not an easy task and semantic web aims to achieve such a challenge. It aims to describe things in a way that various applications and web services can understand them. It is not about more links between web pages rather it describes the relationships between entities and their properties. The upcoming sections provide a brief description of semantic web architecture.

2.2 Architecture of Semantic Web

The main objective of semantic web is to express the meaning. In order to achieve this objective semantic web must meet certain standards and tools. Implementation of
semantic web demands for a tool through which information spread across various sources can be accessed and processed. Tim Berners-Lee [14] indicated that software agents will serve as the core working component in semantic web, since they are able to draw information from distributed heterogeneous sources, relate them and draw conclusion. Implementation of semantic web will lead to machine-readable web content and automated searching.

Figure. 2.1 Four Versions of the Semantic Web Reference Architecture (V1-V4).

Architecture for semantic web will assist in the development of specifications and applications. The most well-known versions of the layered architecture that exist in the literature are provided by Berners-Lee. Berners-Lee proposed four versions of semantic web architecture in 2001 [12], 2003 [13] & 2006 [11] respectively (see figure 2.1). However these versions were not consistent with the principles of layered architectures.

As shown in figure 2.2, Gerber et al [2] in 2008 proposed an orthogonal layered architecture called Comprehensive Functional Layered (CFL) architecture for the semantic web. All the above architectures (refer fig 2.1) suffered from inconsistency in layer arrangement as well as confusion in the purpose of layers. Gerber et al evaluated all
the four architectures on the principles of layered architectures and highlighted that the architectures proposed by Tim Berners-Lee can not be termed as layered architectures at all. Their work throws light on inconsistent layer arrangement and also the doubt in the functionality of the layers.

Figure 2.2: CFL Architecture of Semantic Web

CFL architecture overcomes the drawbacks of Lee’s architecture and provides a sound base for semantic web development. The four versions of semantic web architecture given by Berners-Lee, however vague they proved as milestones in semantic web implementation and served as reference architecture. All layers of semantic web are explained in detail in upcoming subsections.

2.2.1 Unicode and Uniform Resource Identifier (URI)

This layer is responsible for the encoding of any symbol of any language or character set and at the same time responsible for uniquely identifying different resource or entities. Unicode and URI follow the important features of the existing world wide
web. Unicode is a standard of encoding international character sets and it emphasizes that all human languages should be encoded using a standardized form. URI is a string in standard form that allows unique identification of resources. URL is a subset of URI, which contains access mechanism and a location of the web contents. The usage of URI is important in distributed world wide web as it provides understandable identification of all resources. An international variant of URI is Internationalized Resource Identifier (IRI) [56] that allows usage of Unicode characters in identifiers, for which a mapping to URI is defined.

**2.2.2 Extensible Markup Language (XML)**

XML layer along with XML namespace and XML schema definitions make sure that a common syntax be used in the semantic web. XML is a general purpose markup language for documents containing structured information. An XML document contains elements that can be nested and may have attributes and contents. XML namespace allow specifying different markup vocabularies in one XML document. XML schema provides schema for a particular set of XML documents. Currently XML is not a necessary component of semantic web technologies as alternative syntaxes such as Terse RDF Triple Language (Turtle) [2] which is a serialization format for Resource Description Framework (RDF) graph are available.

**2.2.3 Resource Description Framework (RDF)**

Resource Description Framework provides the core data representation format for semantic web. RDF is recommended standard by W3C for publishing data on semantic web. RDF is a framework for representing information about a resource in a graphical form. It was primarily intended for representing metadata about WWW resources, but it
can be used for storing any other data as well. It makes use of subject-predicate-object which results in graphical data. It provides a simple language for expressing data models i.e. objects and their relationships. An RDF based model can be represented in a variety of syntaxes such as RDF/XML etc. RDF is a fundamental standard of the semantic web. RDF itself serves as a description of graph formed by triples.

Anyone can define vocabulary of terms used for more detailed description of taxonomy and other ontological constructs. RDF schema extends RDF and is a vocabulary for describing properties and classes of RDF based resources with semantics for generalized-hierarchies of such properties and classes & thus can be used to describe taxonomies of classes and properties and use them to create lightweight ontology. Figure 2.3 explains the above mentioned concept.

![RDF Triples](image)

Figure 2.3: RDF Triples
2.2.4 Ontology

Ontology refers to the vocabulary of a domain. To make computers understand meaning of various terms these must be supported by some files containing description of terms along with their relationship with each other.

Ontology can be defined as a collection of terms used to describe a specific domain. It provides a mechanism for describing objects their properties and the relation between different resources or objects. Ontology should have the ability to support inference (see figure-2.4). Some applications need simple ontology while others may need ontology with broad capabilities. Thus the ontologies can be developed either using RDF schema for describing objects properties and their relationships for a simple ontology, or Ontology Web Language (OWL) can be used for developing more descriptive ontology. This depends on the needs and the aim of the ontology under consideration.

![Figure 2.4: Role of Ontology in Inference Drawing](image)

Figure 2.4: Role of Ontology in Inference Drawing
There are two main classes of ontology. The first would be the one that is employed to explicitly capture "static knowledge" of a domain in contrast the 2nd ontology provides reasoning about the domain knowledge (problem solving knowledge).

Table 2.1: Types of Ontology

<table>
<thead>
<tr>
<th>TYPE</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain Ontology</td>
<td>Designed to represent knowledge relevant to a certain domain. For example medical, mechanical etc.</td>
</tr>
<tr>
<td>Generic Ontology</td>
<td>Represents knowledge or concepts of general nature that can be applied to a variety of domain types Also called ‘super theory’ [22]. For example Human being, animals etc.</td>
</tr>
</tbody>
</table>

2.2.5 Rules

This layer aims to support inference drawing to allow query and filtering. Up till now, there is no recommended language for this layer but Rule Interchange Format (RIF) is the most commonly used language. Rules are those used by the production systems presented in the corresponding knowledge representation subsection. They capture dynamic knowledge as a set of conditions that must be fulfilled in order to achieve the set of consequences of the rule. The technology behind this layer is the Semantic Web Rule Language (SWRL) [4]. It is based on Rule Modeling Language (RuleML). Similar to RuleML, SWRL covers the entire rule spectrum, from derivation and transformation rules to reaction rules. Thus it supports queries and inferences drawing using web ontologies, mappings between various ontologies, and dynamic web behaviours of workflows, services and agents.
2.2.6 *Logic & Proof Layer*

This layer provides facility of writing logic into documents thus providing rules for deduction of one type of document into another type. This layer includes predicate logic and quantifiers so as to facilitate deductions. Knowledge Interchange Format (KIF) is the language used to specify logic in this layer. The proof is a chain of assertions and reasoning rules with pointers to all the supporting material. This layer helps in checking the authenticity of the users and thus helps in trust establishment. Figure 2.5 represents the structure of proof layer.

![Figure 2.5: Structure of Proof Layer](image-url)

Figure 2.5: Structure of Proof Layer [84]
2.2.7 Trust Layer

Trust layer ensures that the source of information is genuine and authentic at the same time ensuring that only authorized application agents and users can have access to the information. It requires that the reasoning system must include signature verification system. This will result into a system which can express and reason about relationships across the whole range of public key based security and trust systems.

2.3 Development of Semantic Web

Development of semantic web involves publishing information on web as documents along with semantic markups. Semantic web development focuses on the way through which machines can understand the structure, behavior and even meaning of the published information thereby making search of information and its integration more efficient. Although semantic publication has the potential to change the whole scenario on which current web is working but its acceptance depends on the emergence of applications facilitating this publication. Broadly following are the two approaches that can be adopted for publishing meaningful contents on the web.

- As data objects, using semantic web language like RDF and OWL [83].
- As embeded formal metadata in documents using new markup language like Microformats [83].

2.3.1 Publishing information as data objects using Semantic Web Languages

RDF is a language for describing information and resources on web. In RDF files, information is kept in such a way that makes it possible for computer programs to search, discover, collect, analyze and process information from web. The RDF data model is
similar to that of classic conceptual modeling approach such as entity-relationship or class diagrams, as it is based on the idea of making statements about resources in the form of subject-predicate-object expressions. These expressions are known as triples in RDF terminology. The subject denotes the resource and predicate and expresses the relationship between the subject and the object. For example “The sky has the color blue” in RDF triple, “the sky” is the subject, “has the color” is predicate and blue is an object. RDF is an abstract model with several file formats and thus the particular way in which a resource can be encoded vary depending upon the format used.

2.3.2 Embedded formal metadata in documents using Microformats

A microformat is a web based approach based on the reusability of existing HTML/XML tags to convey metadata and other attributes in web pages. This approach allows software to provide information intended by the end user automatically. Although the content of web pages facilitate keyword based automatic processing but they can’t support the concept of semantic web as the traditional markup tags displays information on the web and do not describe the meaning of information. Microformats can bridge this gap by attaching metadata or simply semantic which will support more complicated method of automated processing such as natural language processing. The use, adoption and processing of microformats enables data items to be indexed, searched, saved or cross-referenced so that the information can be combined and reused.

2.4 Semantic Web Applications

Currently, computers are changing from single, isolated devices to an entry points to a world wide network of information exchange and business transactions. However, the success of the WWW has made it increasingly difficult to find, access, present and
maintain the information required by a wide variety of users. In response to this problem, many new research initiatives and commercial enterprises have been set up to enrich the available information with machine understandable semantics. Thus semantic web is providing intelligent access to heterogeneous, distributed information, enabling software products (agents) to mediate between user needs and the information sources available.

Semantic web application areas are experiencing intensified interest due to the rapid growth in the use of the web, together with the innovation and renovation of information content technologies. The semantic web is regarded as an integrator across different content and information applications, systems and provides mechanisms for the realization of Enterprise Information Systems (EIS) [62]. This section explores some areas of semantic web applicability.

2.4.1 Semantic Search Engine

Search engine are the most common tools to extract desired information from the web. A semantic search engine stores semantic information about web resources and is able to solve complex queries, considering the context where the web resource is targeted; it also allows clients to obtain information about commercial products and services as well as about sellers and service providers which can be hierarchically organized. Semantic search engines may can significantly contribute to the development of electronic business applications since it is based on strong theory and widely accepted standards [5]. Thus in order to provide efficient search results search engines are also developed using semantics and have emerged as semantic search engines. Currently many of semantic search engines are developed and implemented in different working environments. Some existing semantic web search engine are:
(A) Hakia

Hakia [34] is an innovative search engine that uses semantics to present the relevant search results to the users. It works like any other search engine where a user has to type the keywords to initiate a search query. However, Hakia provides the results based on laws of ontological semantics, mathematics and fuzzy logic. As a result, the user gets the most relevant answers to the query asked for. A user can give a combination of search words that are pertinent to their query and expect results based on the relevance rather than popularity of the webpage. If the search results are very common, instead of scattering the information along many pages, Hakia presents the user with a gallery on the subject. These galleries are placed on the first page of the search that tab the different aspects of the search query; for example if a user searches for Clint Eastwood [5], it will find his biography, images, photograph sites, interviews, blogs, quotes, contact information, filmography etc presented under different tabs on the first page.

From the second page onwards the search results are simply presented in the order of their relevance. Most search queries are accompanied by a blurb on top of the page that explains the content of the page. In short Hakia tries to provide the users with information that is really needed; while trying to make it easy for the user to search through the results. However, there are times when the keyword typed in by the user restricts the coverage of the search engine, hence the key words must be given very carefully [34].

(B) SenseBot

SenseBot [21] is a web search engine that summarizes search results into one concise digest on the topic being searched. The search engine attempts to understand
what the result pages are about. For this purpose, it uses text mining to analyze web pages and identify their key semantic concepts. This way SenseBot helps the user to get a better grasp of what the relevant term is about. Thus, the user does not have to go through a large number of web pages and comb through results with incomprehensible expert [21].

2.4.2 Ontology Search Engines

Maedche et al [6] designed an integrated approach for ontology searching, reuse and update. In its architecture, an ontology registry is designed to store the metadata about ontologies and ontology server stores the ontology. The ontologies in distributed ontology servers can be created, replicated and evolved. Ontology metadata in ontology registry can be queried and registered when a new ontology is created. Search in ontology registry is executed under two conditions: Query-by-example i.e to restrict search fields and search terms and Query-by-term i.e. to restrict the hyponyms of terms for search.

2.4.3 Knowledge Management

Knowledge management [49] concerns itself with acquiring accessing and maintaining knowledge within an organization. It has emerged as a key activity of large businesses because they view internal knowledge as an intellectual asset from which they can draw greater productivity, create new value, and increase their competitiveness. Knowledge management is particularly important for international organizations with geographically dispersed departments. Most information level of web increases day by day. In traditional day’s uses of the web were limited but today web is used in every field. There are different categories of web, surface web, deep web and semantic web. Normal web user spends maximum time on the surface we currently available in a weakly
structured form. From the knowledge management perspective, the current technology suffers from limitations in the following areas:

(i) Searching Information: Companies usually depend on keyword-based search engines, the limitations of which had already outlined.

(ii) Extracting Information: Human time and effort are required to browse the retrieved documents for relevant information. Current intelligent agents are unable to carry out this task in a satisfactory fashion.

(iii) Maintaining Information: Currently there are problems, such as inconsistencies in terminology and failure to remove outdated information.

(iv) Uncovered Information: New knowledge implicitly existing in corporate databases is extracted using data mining. However, this task is still difficult for distributed, weakly structure collections of documents.

(v) Viewing Information: Often it is desirable to restrict access of certain information to certain groups of employees. “Views,” which hide certain information, are known in context of databases but are hard to realize over an internet (Web).

*The aim of the semantic web is to allow much more advanced knowledge management systems*. Where:

(i) Knowledge will be organized in conceptual spaces according to its meaning.

(ii) Automated tools will support maintenance by checking for inconsistencies and extracting new knowledge.

(iii) Keyword-based search will be replaced by query answering: requested knowledge will be retrieved, extracted, and presented in a human friendly way.

(iv) Queries involving reference to several documents will be supported.
Defining who may view certain parts of information (even parts of documents) will be possible.

### 2.4.4 Semantic Network

A semantic network [62] is a graphical notation for representing knowledge in patterns of interconnected nodes and arcs. Computer implementations of semantic networks were first developed for artificial intelligence and machine translation but earlier versions have long been used in philosophy, psychology, and linguistics.

As shown in figure 2.6, common to all semantic networks is a declarative graphical representation that can be used either to represent knowledge or to support automated systems for reasoning about knowledge.

![Semantic Network Diagram](image)

Figure 2.6: Semantic Network

The semantic network represents knowledge as a directed graph. The nodes of the graph are different things, and the edges of the graph express relationships between those things.
This is most easily illustrated with a diagram. Here, the things are "Animal", "Mammal", "Fish", and so on. The relationships are "is a", "has" and "lives in".

Some versions are highly informal, but other versions are formally defined systems of logic. Six most common kinds of semantic networks are Definitional Network, Assertional Networks, Implicational networks, Executable networks, Learning networks, Hybrid Networks and are defined as follows.

(i) A definitional network [62] emphasizes the subtype or is-a relation between a concept type and a newly defined subtype. The resulting network, also called a generalization hierarchy, supports the rule of inheritance for copying properties defined for a super type to all of its subtypes. Since definitions are true by definition, the information in these networks is often assumed to be necessarily true.

(ii) Assertional networks are designed to assert propositions. Unlike definitional networks, the information in an assertional network is assumed to be contingently true, unless it is explicitly marked with a modal operator. Some assertional networks have been proposed as models of the conceptual structures underlying natural language semantics.

(iii) Implicational networks use implication as the primary relation for connecting nodes. They may be used to represent patterns of beliefs, causality, or inferences.

(iv) Executable networks include some mechanism, such as marker passing or attached procedures, which can perform inferences, pass messages, or search for patterns and associations.

(v) Learning networks build or extend their representations by acquiring knowledge from examples. The new knowledge may change the old network by adding and
deleting nodes and arcs or by modifying numerical values, called weights, associated with the nodes and arcs.

(vi) Hybrid networks combine two or more of the previous techniques, either in a single network or in separate, but closely interacting networks.

2.4.5 E-Commerce

Semantic web technologies enable machines to interpret data published in a machine-interpretable form on the web. At the present time, only human beings are able to understand the product information published online. The emerging semantic web technologies have the potential to deeply influence the further development of the internet economy [36].

Over 300 million searches are conducted everyday on the internet by people trying to find what they need. A majority of these searches are in the domain of consumer e-commerce, where a web user is looking for something to buy. This represents a huge cost in terms of people hours and an enormous drain of resources. Agent enabled semantic search will have a dramatic impact on the precision of these searches. It will reduce and possibly eliminate information asymmetry where a better informed buyer gets the best value. By impacting this key determinant of market prices semantic web will foster the evolution of different business and economic models.

2.4.6 Domain Understandability

Ontologies provide formal semantics, thereby making information understandable not only to humans but also machines understandable. In addition ontology mappings or translation could foster the understandings between domains so as to enable the dialogue
across multiple domains. Ontology translation could bridge them. Internet commerce requires automatic negotiation and contracting for all searched results. This feature could significantly help machines process a large amount of business partner information that humans cannot handle, and thus save time and money.

2.5 Semantic Web Services

Web services are software modules which provide some kind of service and which can be accessed and invoked via a network. Semantic web services [45] are considered to be the next step in the evolution of web services. In addition to web services, semantic web services provide formal and machine understandable descriptions of their capabilities which allow them to be understood and processed by automatic algorithms. Semantic web services are envisioned to enable sophisticated tasks such as automated discovery of services by matching specified service requests with a large pool of service providers, automated composition of services to instantiate high-level descriptions of complex tasks by a sequence of calls to simpler services, and service monitoring to evaluate the quality of work provided by services [45]. Some existing semantic web services are for communication, description & discovery. There are briefly discussed as follows:

2.5.1 Semantic Web Services for Communication

SOAP (Simple Object Access Protocol) [45] is a simple XML-based messaging protocol which has been proposed to exchange messages among web services. A SOAP message is a document which follows a pre-defined format of encoding message content, specifying sender and recipient, content and control fields of the message.
2.5.2 **Semantic Web Services for Description (WSDL)**

The Web Service Description Language is a language proposed by the W3C to describe web services. A service is described as a set of endpoints operating on messages. A Web Service in the definition of WSDL consists of ports, each of which can be accessed by a number of specified operations. Operations in turn are invoked by sending messages whose exact format is also determined by the WSDL document.

2.5.3 **Semantic Web Services for Discovery**

Universal Description, Discovery and Integration (UDDI) is a standard proposed by companies such as Sun, IBM and HP. The standard specifies a web service description format and architecture for a web services description repository. UDDI addresses the issue of discovering web services.

2.6 **Conclusions**

This chapter provides a brief description of semantic web with its architecture. Architecture for semantic web will assist in the development of specifications and applications. Semantic web applications are experiencing intensified interest due to the rapid growth in the use of the web, together with the innovation and renovation of information content technologies and development focuses on the way through which machines can understand the structure, behavior and even meaning of the published information thereby making search of information and its integration more efficient. The services of semantic web are also explained in this chapter. Next chapter provides details of intelligent agents.