SEMANTIC WEB : THE FUTURE OF WWW

PROTITI MAJUMDAR

Abstract

The World Wide Web has a great impact on the people for communication and transactions. It has helped transform communities towards knowledge economy, broadly speaking into a knowledge society. Now the web is facing a significant change since its inception, and Sir Tim Berners-Lee is again at the forefront of the change introducing the concept of Semantic Web. In this paper I will be trying to emphasize on some of the following points. At first a general idea about www then search engines general features and advanced features after that its drawbacks and semantic web and its different aspects. The main moto behind this paper is to show how this Semantic Web technology helps in information retrieval which are not possible by the search engines.

Keywords : Semantic Web/ Interoperability

1. Introduction

The World Wide Web (WWW) has drastically changed the availability of electronically Accessible information. The essential property of the Worlds Wide Web is its universality. Typical uses of the web today involve people's seeking and making use of information, searching for and getting in touch with other people, reviewing catalogs of online stores and ordering products by filling out forms etc. The WWW currently contains some 3 billion static documents, which are accessed by over 300 million users internationally. However, this enormous amount of data has made it increasingly difficult to find, access, present and maintain the information required by a wide categories of users. This is because information content is presented primarily in natural language, and the only tools for finding information of the net are search engines such as Google, Yahoo and others. But often patrons are not satisfied with the results, which retrieve more noise than relevance. The main reason for this the way information is represented on the web.

Thus, a wide gap has emerged between the information available for tools aimed at addressing the problems above and the information maintained in human-readable form. In response to this problem, many new research initiatives and commercial enterprises have been set up to enrich available information with machine-processable semantics. Such support is essential for "bringing the web to its full potential". Tim Berners-Lee, Director of the World Wide Web Consortium, referred to the future of the

current WWW as the "semantic web" - an extended web of machine-readable information and automated services that extends far beyond current capabilities.

The semantic web approach aims to develop languages for expressing information in a machine processable way. Tim Berners-Lee, who is the inventor of the World Wide Web, first envisioned a semantic web that provides automated information access based on machine-processable semantics of data. The explicit representation of the semantics of data, accompanied with domain theories (i.e. ontologies), will enable a web that provides a qualitatively new level of service.

2. What is WWW ?

(WWW, W3, The Web) A distributed information retrieval system, which originated from the High-Energy Physics laboratories in Geneva, Switzerland. An extensive user community has developed on the Web since its public introduction in 1991. [15]

3. Search Engine

A search engine is a program designed to find information stored on a computer system such as the World Wide Web, or a personal computer. The search engine allows one to ask for content meeting specific criteria (typically those containing a given word or phrase) and retrieving a list of references that match those criteria. Search engines use regularly updated indexes to operate quickly and efficiently.

A program that searches documents for specified keywords and returns a list of the documents where the keywords were found. Although search engine is really a general class of programs, the term is often used to specifically describe systems like Alta Vista and Excite that enable users to search for documents on the World Wide Web and USENET newsgroups.

3.1 How it work ?

Typically, a search engine works by sending out a spider to fetch as many documents as possible. Another program, called an indexer, then reads these documents and creates an index based on the words contained in each document. Each search engine uses a proprietary algorithm to create its indices such that, ideally, only meaningful results are returned for each query.

3.2 Types of Search Engines

We can categorize this search engine in this way: -

- Automatic Search engine
- Directories
- Pay per click and
- Meta Search engine

3.2.1 Automatic Search Engines : These are based on information that is collected, sorted, and analyzed automatically by indexing spiders.

Examples:

- Excite
- AltaVista
- Google

3.2.2 Directories : Directories are created by people who actually create the website's listing on the search page, as opposed to a 'robot' or 'spider' to do this automatically

Examples:

- Yahoo
- AOL
- AltaVista

3.2.3 Pay Per CLICK : Pay per click search engines and pay per click advertising provides web site owners the opportunity to buy their way to better positions on search results pages.

Examples:

- Overture
- Google Adwords

3.2.4 Meta Search Engines : These search engines display results, which are actually a combination of the results of many search engines

Examples:

- Mamma
- Dogpile

Categorisation of Meta Search Engine

In terms of MSE classification, there are two basic groups:

- "pseudo" Meta Search Engines
- "real" Meta Search Engines

Pseudo Meta Search Engine : They exclusively arrange search results by single Search Engine and provide separate retrieval results from a single Search Engine.

REAL Meta Search Engine : They provide collated retrieval results and offer some control mechanisms such as filter and timeout

3.3 features of search engine :

I listed certain features of search engines which found in more or less all of the search engines, so we can say these are the general features of search engines.

3.3.1 General Features

- Images: Google, Altavista, MSN etc.
- News: -MSN, Yahoo, Altavista etc.
- Local search: - Google, MSN, Ask jeeves etc.
- Movies: -Ask jeeves, Google etc.
- Music /Audio: - Alta vista, Lycos. Yahoo etc.
- Spell check: - Ask jeeves, Google etc.
- Case Sensitivity:-Teoma, Yahoo etc.
- Stop Words: - Google, etc
- Mail: - Yahoo, Google, etc
- Chat: - Yahoo, Google, etc

3.3.2 Advanced Features : This advance features are found to certain restricted search engines.

- Filtration
- Boolean Search
- Proximity searching
- Truncation
- Limits

3.4 Examples of some popular search engines

Google, Yahoo, MSN, Alta Vista, AQL Search, Ask Jeeves, HotBot, Lycos, Teoma, Clustry, Kartoo,

Google : It provides the following features, which are the unique for it. These are

Book search

- Cached links
- Language
- Calculator
- MSN

It includes a search Builder that includes an option to retrieve results based on recent updates, popularity and exact or approximate match.

AltaVista : It searches web sites and Usenet news groups with advanced Boolean and field options. It gives also translation service.

Ask Jeeves : Submit question in plain English and view suggested relevant sites. It also offers open directory.

Hot Bot : It clusters results by presenting one hit per site.

Clusty : It clusters results from a variety of surface and deep web sources and organizes them into clusters by topic, site or URL.

Vivisimo: Searches multiple engines and directories and organizes results into topical categories.

3.5 Sort Comes of Search Engines

1. Regardless of the growing sophistication, many well thought-out search phrases produce list after list of irrelevant web pages. The typical search still requires sifting through dirt to find the gems.
2. Using search engines does involve a learning curve. Many beginning Internet users, because of these disadvantages, become discouraged and frustrated. [2]

To solve this problem semantic web come into the scenario.

4. Semantic Web

4.1 Definition from dictionary

Created by Tim Berners-Lee, the "semantic Web" allows machines to interpret data and instinctively transmit that data in a way that is very perceptive to the user. This information exchange takes documents with computer-comprehensible meaning (semantics) and puts them on the WWW. [3]

So, in a simple term we can say semantic web as — The word semantic implies meaning or, as WordNet defines it, "of or relating to study of meaning and changes of meaning." In the term "semantic web", 'semantic' also indicates that the meaning of data on the web can be discovered- not just by people, but also by computers.

It is a vision in which computers-software as well as people can find, read, understand, and use data over the World Wide Web to accomplish useful goals for users.

The semantic web is a vision of the next generation web, which enables web applications to automatically collect web documents from diverse sources, integrate and process

information and interoperate with other applications in order to execute sophisticated tasks for humans.

4.2 Motivation for Semantic Web

The semantic web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is based on the idea of having data on the web defined and linked such that it can be used for more effective discovery, automation, integration, and reuse across various applications.

Keyword-based search engines, such as Google, Yahoo and Alta Vista, are the main tools for using today's web. It is clear that the web would not have been the huge success were it not for search engines. However there are serious problems associated with their use, like,

- High recall/ low precision: Even if the main relevant pages are retrieved, they are of little use if another 28,758 mildly relevant or irrelevant documents were also retrieved. Too much can easily become as bad as too little.
- Low or no recall: Often it happens that we don't get any answer for our request, or that important and relevant pages are not retrieved. Although low recall is a less frequent problem with current search engines, it does occur.
- Results are highly sensitive to vocabulary: Often our initial keywords do not get the results we want; in these cases the relevant documents use different terminology from the original query. This is unsatisfactory because semantically similar queries shuld teturn similar results
- Results are single web pages: If we need information that is spread over various documents, we must initiate several queries to collect the relevant documents, and then we must manually extract the partial information and put it together.

To realize the above-mentioned vision the following are necessary.

Automation - it would be nice if computers could do more (on the web). Solution: make information on the web more "machine-friendly".

Interoperability- combining information from multiple sources (so big organizations can avoid duplication)

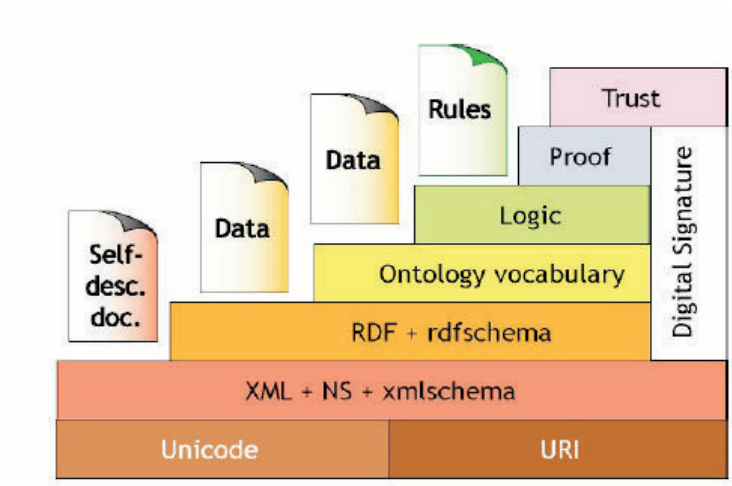
- Web services - discovery, composition

Departure from the tool paradigm- instead of using computers like tools, make them work on our behalf

- Removing humans from the loop to the extent possible.

4.3 Semantic Web Layer Approach

Semantic web technology is built in a layered manner, i.e. it is processed in steps, each step built on top of another. The pragmatic justification of it is that it is easier to achieve consensus on small steps, whereas it is much harder to get everyone on board if too much is attempted.



4.3.1 Layered Approach to Semantic Web : In building the semantic web in a layered manner, two principles should be followed:

1. Downward Compatibility: Agents (Agents are pieces of software that work autonomously and proactively.) fully aware of one layer should also be able to interpret and use information written at lower levels. e.g. agents aware of the semantics of OWL can take full advantage of information written in RDF and RDF Schema.
2. Upward Partial Understanding: agents fully aware of one layer should also be able to take at least partial advantage of information at higher levels. e.g. an agent aware of only RDF and RDF Schema semantics can interpret partial knowledge written in OWL, by disregarding those elements that go beyond RDF and RDF Schema.

The "layer cake" of the semantic web technology as shown in the above figure, describes the main layers of the semantic web design and vision.

4.3.2 XML : At the bottom of the Semantic Web layer is XML (Extensible Markup Language) and XML Schema. XML is a subset of SGML (Standard Generalised Markup Language). Today HTML (Hypertext Markup Language) is the most popular language in which Web pages are written was developed from SGML, because SGML was considered far too complex for Internet-related purposes. And XML was driven by shortcomings of HTML.

XML lets everyone create their own tags-hidden labels such as or that annotate web pages of sections of text on a page, but it says nothing about what the structures mean. XML is particularly suitable for sending documents across the web.

Important Features of XML:[5]

- Extensible: tags can be defined; can be extended to lots of different applications.
- Markup language: which allow one to write some content and provide information about what role that content plays.
- Allows the representation of information that is also machine-accessible:
- XML document is, more easily accessible to machines because every piece of information is described. Moreover, their relations are also defined through the nesting structure. For example, the <author> tags appear within the <book> tags, so they describe properties of the particular book. A machine processing the XML document would be able to deduce that the author element refers to the enclosing books element, rather than having to infer this fact from proximity considerations, as in HTML.
- Separates content from formatting: that means, the same information can be displayed in different ways, without requiring multiple copies of the same content; moreover, the content may be used for purposes other than display.
- A metalanguage for markup: it does not have a fixed set of tags but allow users to define tags of their own.

Limitations of XML

- XML is a universal metalanguage for defining markup. It provides a uniform framework, and a set of tools like parsers, for interchange of data and metadata between applications. But it has also some limitations like,
- XML is not equal to machine accessible meaning, only to people. For example, one can use an element as 'Author'; another can use it as 'Writer'. Here, human can make out that both are same, but how system can? This creates confusion when machines try to share data with each other.

  The nesting of tags does not have standard meaning; it is up to each application to interpret the nesting. *For example,*

  Nabonita Guha is a lecturer of Information Technology.

  There are various ways of representing this sentence in XML. Two possibilities Are:

  <course name=" Information Technology ">
  <lecturer> Nabonita Guha </lecturer>
  </course>
  <lecturer name=" Nabonita Guha ">
  <teaches> Information Technology </teaches>
  </lecturer>

The above two formalizations include essentially an opposite nesting although they represent the same information. So there is no standard way of assigning meaning to tag nesting.

- Domain-Specific Markup Languages: Since the user is at freedom to define his/her own tags, many domain-specific markup languages have been developed like, MathML and CML (Chemical Markup Language). The problem with various domain-specific markup languages is that of non-standardization while describing the resources on the Web.

But at the same time preventing this kind of flexibility and extensibility will again result in lack of inadequate resource description. Hence, there should be a common model/framework which can bridge the gap between these various schemas. It is at this stage that the RDF came into the picture which is the next layer in the Semantic Web pyramid.

RDF is a basic data model, not a language. The RDF model provides the description of Web documents (in other words rendering of metadata to the documents) in a natural manner so that the metadata can be shared across different applications. RDF expresses the meaning, which encodes in sets of triplets (resource/subject, predicate/property and object/value), each triplet being rather like the subject, verb and object of an elementary sentence. These triplets can be written using XML tags.

4.3.3 RDF Triplets

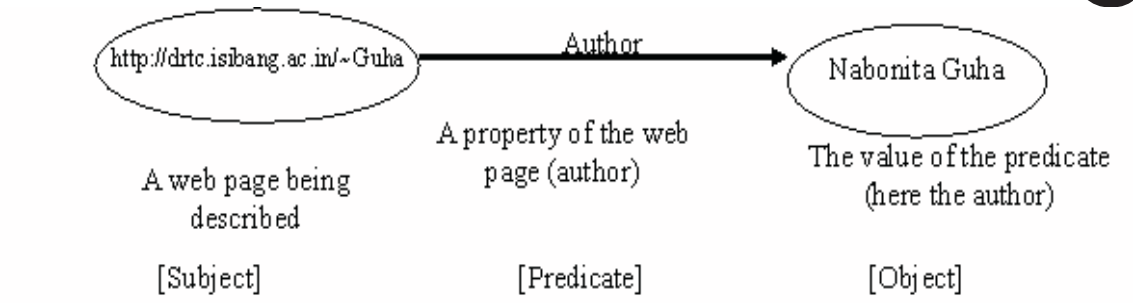
A simple RDF model has three parts,

- Subject/Resource: Any entity which has to be described is known as resource, also known as subject. It can be a 'web page' on Internet or a 'person' in a society.

- Predicate/Property: Any characteristic of resource or its attribute which is used for the description of the same is known as property or predicate. For example, a web page can be recognized by 'Title' or a man can be recognized by his 'Name'. So both are attributes for recognition of resource 'web page' and 'person' respectively.

- Object/Value: A property must have a value also known as object. For example, the title of DRTC webpage is "Documentation Research and Training Center", name of a person is "S. R. Ranganathan".

The combination of subject, predicate and object is said to be a 'Statement' or 'Rule'. For example, a statement, Nabonita Guha is the author of the web page http://drtc.isibang.ac.in/~Guha. This statement can be represented diagrammatically as follows:



The rdf:Description element makes a statement about the resource http:// drtc.isibang.ac.in/~Guha. Within the description the property is used as a tag, and the content is the value of the property.

The most important feature of RDF is that it is developed to be domain-independent, i.e. it is very general in nature and does not restrict/apply any constraint on any one particular domain. It can be used to describe information about any domain.

The RDF model imitates the class system of object-oriented programming. A collection of classes (as defined for a specific purpose or domain) is called a 'schema' in RDF. These classes are extensible through 'subclass refinement'. Thus, various related schemas can be made using the base schema. RDF also supports metadata reuse by allowing transmission or sharing between various schemas.

RDF versus RDFS Layers

An illustration of different layers involved in RDF and RDFS can be represented in the following way for a statement, Semantic web and its application is taught by Nabonita Guha.

The schema for this statement may contain classes such as lecturers, academic staff members, staff members, courses and properties such as is taught by, involves, etc. The above statement can be illustrated as follows. In the following figure, blocks are properties, ellipses above the dashed line are classes, and ellipses below the dashed line are instances.

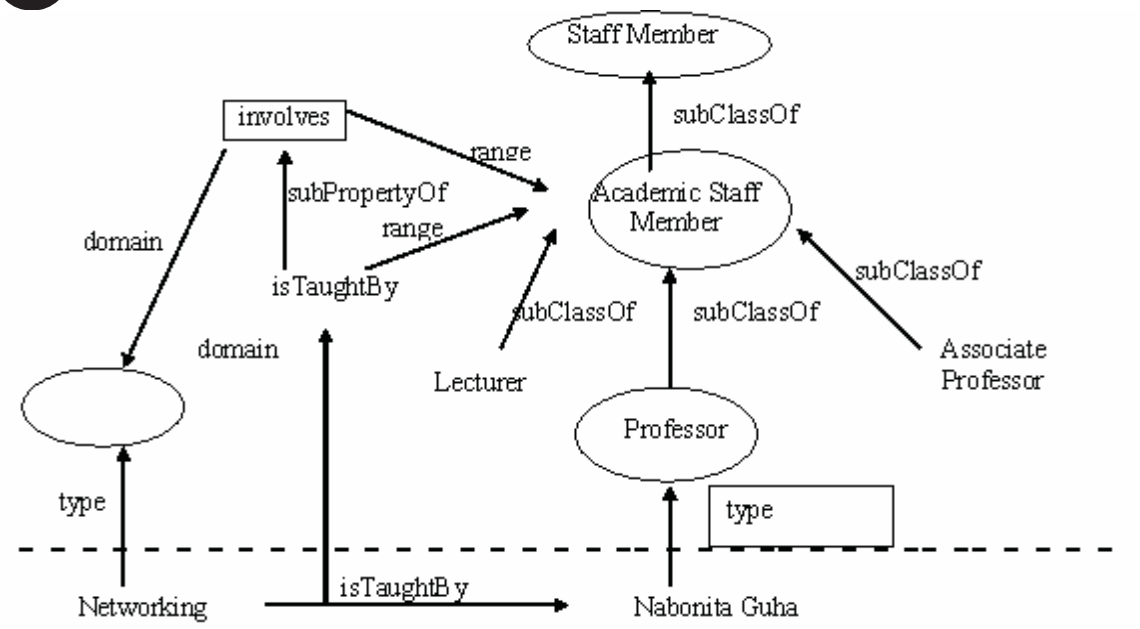


Fig: RDF and RDFS Layers

Limitations of RDF Schema

RDF and RDFS allow the representation of some ontological knowledge. The main modeling primitives of RDF/RDFS concern the organization of vocabularies in typed hierarchies: subclass and sub property relationships, domain and range restrictions, and instances of classes. However, a number of other features are missing, like,

- Local scope of properties: rdfs:range defines the range of a property, say eats, for all classes. Thus in RDF Schema we cannot declare range restrictions that apply to some classes only. For example, we cannot say that cows eat only plants, while other animals may eat meat, too.
- Disjointness of classes: Sometimes we wish to say that classes are disjoint. For example, male and female are disjoint. But in RDF Schema we can only state subclass relationships, e.g., female is a subclass of person.
- Boolean combinations of classes: Sometimes we wish to build new classes by combining other classes using union, intersection, and complement. For example, we may wish to define the class person to be the disjoint union of the classes male and female. RDF Schema does not allow such definitions.
- Cardinality restrictions: Sometimes we wish to place restrictions on how many distinct values a property may or must take. For example, we would like to say

that a person has exactly two parents, or that a course is taught by at least one lecturer. Again, such restrictions are impossible to express in RDF Schema.

- Special characteristics of properties: Sometimes it is useful to say that a property is transitive (like "greater than"), unique (like "is mother of"), or the inverse of another property (like "eats" and "is eaten by").

Thus we need an ontology language that is richer than RDF Schema, a language that offers the above features and more. In designing such a language one should be aware of the trade-off between expressive power and efficient reasoning support. Generally speaking, the richer the language is, the more inefficient the reasoning support becomes, often crossing the border of non-computability. Thus we need a compromise, a language that can be supported by reasonably efficient reasoners while being sufficiently expressive to express large classes of ontologies and knowledge.

4.3.4 What is ontology ? Ontology can be said to be the definition of entities and their relationship with each other. Ontologies define data models in terms of classes, subclasses, and properties. For example, a man is a subclass of human which in turn is a subclass of animals that is a biped i.e. walks on two legs.

Definitions

Neches (1991): Ontology defines basic terms and as well as the rules for combining terms and relations to define extensions to the vocabulary.

Advantages of Ontology

- Provide a shared understanding of domain.
- Useful for the organization and navigation of web sites.
- Useful for improving the accuracy of web searches.
- Web searches can exploit generalization/specialization information.

4.4 How Ontology Language Develops

The Web Ontology Working Group of W3C identified a number of characteristic use-cases for the semantic web that would require much more expressiveness that RDF and RDF Schema offer. The researchers in the United States (US) and in Europe identified the need for a more powerful language to build ontology. In Europe OIL (Ontology Interface Layer), an ontology language was developed. In US, DARPA (Defense Advanced Research Project Agency) had initiated a similar project called DAML (Distributed Agent Markup Language). Latter on these two have been merged and came up with a single ontology language, DAML+OIL.

DAML+OIL in turn was taken as the starting point for the W3C Web Ontology Working Group in defining OWL, the language that is aimed to be the standardized and broadly accepted ontology language of the semantic web.

4.5 Species of OWL

OWL comes in three sub language or flavors, called OWL Full, OWL DL (Description Logic) and OWL Lite.

4.5.1 OWL Full : It is the complete language and it uses all the OWL language primitives. It allows the combination of these primitives in arbitrary ways with RDF and RDF Schema.

Advantage

It is fully upward compatible with RDF, both syntactically and semantically: any legal RDF document is also a legal OWL Full document and any valid RDF/RDF Schema conclusion is also a valid OWL Full conclusion.

Disadvantage

It has become so powerful as to be undecidable, dashing any hope of complete (or efficient) reasoning support.

4.5.2 OWL DL: OWL DL supports a form of what is called description logic. Description logics apply certain carefully chosen restrictions to the kind of things that can be said in order to gain computing advantages. This allows to be sure that description logic processors can successfully compute results—in the jargon, it's "complete and decidable".

Advantage:

- It permits efficient reasoning support.

Disadvantage

- We lose full compatibility with RDF: an RDF document will in general have to be extended in some ways and restricted in others before it is a legal OWL DL document. Every legal OWL DL document is a legal RDF document.

4.5.3 OWL Lite: OWL Lite is OWL DL with more restrictions. For example, OWL Lite excludes enumerated classes; disjoint ness statements, and arbitrary cardinality. The idea is to make it easy to start with and easy to implement processors, so that people can begin using OWL Lite easily and later graduate to more complicated uses.

Advantage

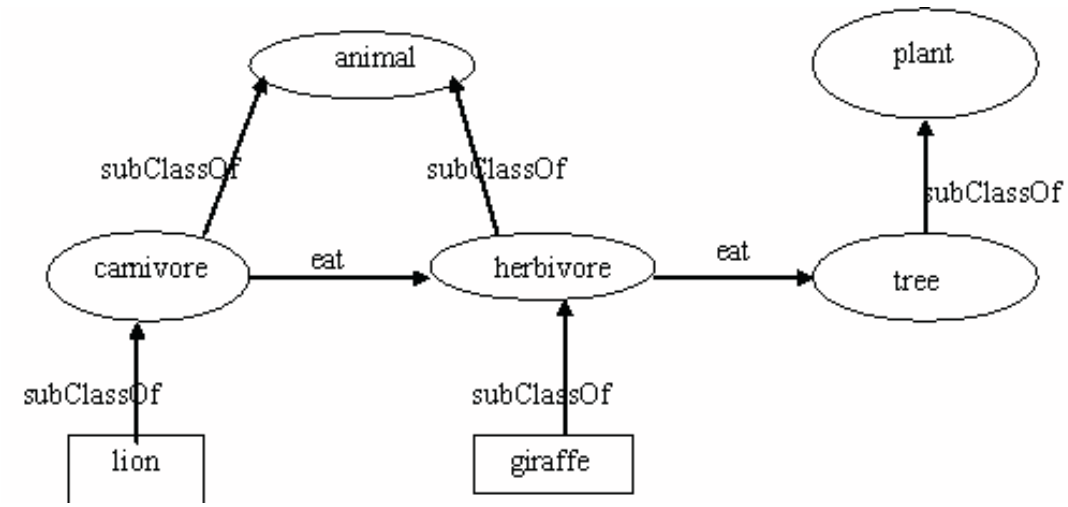
- It is easier to grasp (for users) and easier to implement (for tool builders).

Disadvantage

- The expressivity is more restricted.

Example of an Ontology

This example shows an ontology that describes African wildlife.



Classes and subclasses of the African wildlife ontology

The next layer in the Semantic. Web pyramid is logic and proof which enhance the ontology language further. Logic is the foundation of knowledge representation. It helps to establish the consistency and correctness of data sets and to infer conclusions that aren't explicitly stated but are required by or consistent with a known set of data.

Importance of Logic

- It provides a high-level language in which knowledge can be expressed in a transparent way. And it has a high expressive power.
- It has a well-understood formal semantics, which assigns an unambiguous meaning to logical statements.
- Automated reasoners can deduce (infer) conclusions from the given knowledge, thus making implicit knowledge explicit.
- There exist proof systems for which semantic logical consequence coincides with syntactic derivation within the proof system.
- Because of the existence of proof systems, it is possible to trace the proof that leads to a logical consequence. In this sense, the logic can provide explanations for answers.

Reasoning Example

1. X is a Cat
2. a Cat is a Mammal
3. a Mammal gives birth to live young

Therefore, X gives birth to live young

At the top of the pyramid is the trust layer, which is the high-level and crucial concept: the web will achieve its full potential only when users have trust in its operation (security) and in the quality of information provided. The trust layer will emerge through the use of digital signatures and other kinds of knowledge, based on recommendations by trusted agents or on rating and certification agencies and customer bodies.

Each layer is seen as building on the layer below. Each layer is progressively more specialized and also tends to be more complex than the layers below it. Thus the layers can be developed and made operational relatively independently.

4.6 Applications

The major industrial firms and academic and research institutions have started to think seriously about use and applications of semantic web technology. The semantic web technology has the potentiality to be applied in different areas. Some of the areas are horizontal information mapping for cross-domain resource discovery (implemented by Elsevier, a leading scientific publisher), data integration, e-learning, multimedia collection indexing and online procurement.

5. Conclusion

Most of the people think that semantic web is basically an alternative or parallel technology of Artificial Intelligence (AI). But it is not true. Actually AI aimed to develop intelligent systems emulating human intelligence whereas semantic web has come to just help the web users in their day-to-day work. But I also believe that we have to take help of AI to reach into the vision of Semantic Web. It is too early to predict exact products but semantic web has promising applications especially to achieve meaningful retrieval in a distributed information system like the World Wide Web.

But possibly the first success stories will not emerge in the open heterogeneous environment of the WWW, rather in intranets of large organizations. In such environments, central control may impose the use of standards and technologies, and possibly the first real success stories will emerge.

References

1. http://www.netlingo.com/lookup.cfm?term=semantic%20Web
2. http://www.gsn.org/web/research/internet/disadse.htm
3. http://www.definethat.com/define/3381.htm
4. http://www.webopedia.com/TERM/s/search_engine.html
5. Antoniou, Grigoris and Harmelen, Frank van. A semantic web primer. 2004: MIT Press, London.

6. Davis, John, Fensel, Dieter and Harmelen, Frank van. Towards the semantic web. 2003: John Wiley, West Sussex.

7. Resource Description Framework (RDF) Model and Syntax Specification: W3C Recommendation, 22 Feb. 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/#intro

8. Namespaces in XML. World Wide Web Consortium, 14th Jan. 1999. http://www.w3.org/TR/REC-xml-names/#sec-intro

9. Introduction To Mobile Education http://www.metc.pku.edu.cn/cgz/protect/mc2003/pdf/chapt05-2.pdf

10. Ross, Greg. Towards a semantic web, at http://insight.zdnet.co.uk/internet/0,39020451,39186052,00.htm

11. Fensel, Dieter and Musen, Mark A. The semantic web: a brain for humankind, at www.cs.umbc.edu/771/papers/ ieeeIntelligentSystems/introduction.pdf

12. The semantic web: an interview with tim berners-lee, at http://www.consortiuminfo.org/bulletins/semanticweb.php

13. Berners-Lee, Tim, Hendler, James and Lassila, Ora. The semantic web. Scientific American (May, 2001), at http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2

14. Lassila, Ora. Towards the semantic web, at www.w3c.rl.ac.uk/pastevents/TowardsTheSemanticWeb.pdf

15. Passin, Thomas B. Explorer's guide to the semantic web. 2004: Manning, Greenwich.

16. Sarkar, Ananya. Search Engine. On DRTC colloquim. 2006.

17. http://www.thejcdp.com/issue008/day/04day.htm

18. http://www.cornishwebservices.co.uk/search-engine-optimisation/disadvantages.shtml

19. http://www.gsn.org/web/research/internet/disadse.htm

BIOGRAPHY OF AUTHOR

Protiti Majumdar, ADIS Student, Documentation Research and Training Center, Indian Statistical Institute, 8th Mile Mysore Road, Bangalore- 560059.

Email : protiti@drtc.isibang.ac.in