Towards Intelligent Information Retrieval on Web

Ms. Vandana Dhingra
School of Engineering and technology
Apeejay Stya University
Gurgaon, India

Dr. Komal Kumar Bhatia
Associate Professor, CSE
YMCA UST
Faridabad, India

***Abstract*— The World Wide Web is an information resource with virtually unlimited potential. However, this potential is relatively untapped because it is difficult for machines to process and integrate this information meaningfully and today the WWW links more than 15 billion pages. The retrieval of relevant information on web is an issue that is of main concern. As the internet grew and became popular, more and more information is growing over the WWW has increased to such an extent that extracting relevant information became a major issue –anytime one searches on the web, it ends up in hundreds of links that provide information on a particular topic. Extracting useful information is left on user to decide. This paper emphasizes on where today's search is lacking and need of an intelligent information retrieval system on web and to explore its theoretical framework. This paper also presents the technologies/concepts required to make semantic web a reality like metadata, RDF, URI, XML, XMLS, Triples and Ontologies.**

Keywords*—** ***Semantic web, intelligent retrieval; Triples; Ontologies; RDF; URI; XML; SWD

I. INTRODUCTION

The World Wide Web is the greatest repository of information ever assembled by man. It contains documents and multimedia resources concerning almost every imaginable subject, and all of this data is instantaneously available to anyone with an Internet connection. The Web's success is largely due to its decentralized design: web pages are hosted by numerous computers, where each document can point to other documents, either on the same or different computers. As a result, individuals all over the world can provide content on the Web, allowing it to grow exponentially as more and more people learn how to use it.

However, the Web's size has also become its curse. Due to the sheer volume of available information, it is becoming increasingly difficult to locate useful information. Although directories (such as Yahoo!) and search engines (such as Google and Alta Vista) can provide some assistance, they are far from perfect. For many users, locating the "right" document is still like trying to find a needle in haystack.

Furthermore, users often want to use the Web to more than just locate a document, they want to perform some task. For example, a user might want to find the best price on a desktop computer, plan and book a vacation to a Goa, or make reservations at a moderately-priced Goa restaurant with the movie they plan to see that evening. Completing these tasks often involves visiting a series of pages, integrating their content and reasoning about them in some way. This is far beyond the capabilities of contemporary directories and search engines, but could they eventually perform these tasks? The main obstacle is the fact that the Web was not designed to be processed by machines. Although, web pages include special information that tells a computer how to display a particular piece of text or where to go when a link is clicked, they do not provide any information that helps the machine to determine what the text means.

Thus, to process a web page intelligently, a computer must understand the text, but natural language understanding is known to be an extremely difficult and unsolved problem.Futher in those direction researchers have begun to explore the potential of associating web content with explicit meaning, in order to create a Semantic Web. Rather than rely on natural language processing to extract this meaning from existing documents, this approach requires authors to describe documents using a knowledge representation language.

II. WHERE SEARCH IS LACKING –LEADING TO AN INTELLIGENT INFORMATION RETRIEVAL

Presently, search on web is keyword based i.e., information is retrieved on the basis of text search of all available matching url's/hyperlinks. This may result in the presentation of irrelevant information to the user. In the current web, resources are accessible through hyperlinks to web content spread throughout the world. These links make the physical connections and are not understood by the machines. So there is a lack of relationships which captures the meaning of the links for the machines to understand.

There is a need of having data on the web linked in a way that it can be interpreted by the machines. The idea is to create a more intelligent web by annotating pages on the web with their semantics, understandable by computer. Thus, if a machine could understand what a page was about, it would give the more relevant results.

An intelligent search engine needs to fetch the few pages from the billions available.

The problem issues three topics:

1 Accuracy of keywords entered.

2 The way the search engines map the Net.

3 Accurately tagging the web pages so that search engines know what a particular page is about. There is a need to replace the ineffective web searches by meaningful/intelligent and fast web searches [15]. It is based on the fundamental idea that web resources should be annotated with semantic mark-up that captures information about their meaning. The objective is to provide a common framework that allows data to be shared and reused across applications, enterprise, and community boundaries and which brings structure to the meaningful content of WebPages which results in creating an environment where software agents roaming from page to page readily carry out advanced searches.

III.EVOLUTION OF SEMANTIC WEB

In 2001, the original *Scientific American article* [1] on the semantic web described the evolution of a web that consisted largely of data for *human reading* to one for computer *manipulation*. The Semantic web is a web of actionable information –information derived from data through a semantic theory for the symbols interpretation. The Semantic theory focuses on a system of "meaning" in which the logical connection of terms establishes interoperability between systems.

Scientific American article assumed that it is simple to recruit the data to a particular use context but it's still difficult to achieve in today's web. Tim Berners-Lee proposed architecture of semantic web in 2001.

IV. SEMANTIC WEB VS CLASSICAL WEB

Semantic Web is the next-generation web of concepts linked to other concepts, rather than a collection of hypertext documents linked by keywords. In Classical web, An HTML anchor tag (link) is a keyword reference to another document, which supplies a word or phrase that links to another document, usually displayed as underlined text on a browser. But the link doesn't exactly say how the two documents are related to each other. HTML hyperlinks don't give any real indication about relationships between files, and the text in the link may be extremely vague. A new standard, the Resource Description Framework (RDF), makes it possible to be much more specific about how things are related to each other. In fact, RDF describes much more than documents—any entities or concepts can be linked together. This is the basic idea behind the Semantic Web—that concepts, rather than documents, can be linked together.

V. PROBLEMS WITH CURRENT WEB

Although *WWW* is truly incredible and it has provided features and benefits that have changed our world completely [3]. However, current web technologies are clearly insufficient to supports today's dynamic, distributed and robust computing needs. New web technologies are required primarily to structure the information, improve current search mechanism and expose the semantics of the information [4].
Following summarize some of the major limitations in our current web that encourages the need for a new vision and infrastructure for the web [3], [4].

A. Single Document Search [8]

One fundamental problem in information representation and retrieval from our current Web is that, usually information could be retrieved from a single web page and a single document, and it is extremely difficult to collectively retrieve information which is spread over more then one documents and several web pages [4].

B. Search Limited to Keywords (No semantics)

Many times our search returns no results because of the reason that keywords that we specify are not matched in the searched documents. But this happens even if the required information exists in searched documents but these documents use some different terminology and vocabulary.

C. Irrelevant and Excessive Information

Another problem caused by the keyword based search is that; usually an excessive amount of information is returned as a result of a keyword based query. Most of this information is irrelevant and it is very difficult and time consuming to separate relevant information from this excessive amount of retrieved data.

D. Semi structured Information Representation

Our current web is too document-centric to support sophisticated information representation [3]. The most amount of information available at our current web is either unstructured or semi structured. The information is based on HTML based free format web pages which are although very much suitable for direct human use and information exchange but not appropriate at all for the automated information exchange, retrieval and processing by software agents (machines).

VI. Concepts Used To Make Semantic Web A Reality

A. Metadata

It is information about information, which is widely used in real-world for searching. For example, you want to borrow some books on computer from a library. Usually a library will provide a lookup system which allows you to list books by author, title, subject and so on. This list contains lots of useful information: author, title, ISBN, date and most important, location of the book. You need some information (the book's location) you want to know and you use metadata (information about information, in this case: author, title and subject) to get it. However, metadata is not necessary: you can lookup the book you want to find one by one among all books in the library. Obviously this is not a wise way. In addition, the use of metadata is not just for searching although searching is the most common aim of metadata. There is some other useful information behind the scenes, which are important to business

B. RESOURCE DESCRIPTION FRAMEWORK (RDF)

Resource Description Framework is a framework for processing metadata and it describes relationships among resources with properties and values. It is built on the following rules:

a. **Resource**: Everything described by RDF expressions is called a resource. Every resource has a URI and it may be an entire web page or a part of a web page

b. **Property**: "A property is a specific aspect, characteristic, attribute, or relation used to describe a resource" – W3C, Resource Description Framework (RDF) Model and Syntax Specification. Note that a property is also a resource since it can have its own properties.

c. **Statements**: A statement combines a resource, a property and a value. These three individual parts are known as the "subject", "predicate" and "object".

Example:

Statements can be represented as a graph in RDF.Consider a simple example:

"Vandana is the author of the resource"	
http://vandana.dhingra/home/index.html	
This sentence has the following parts:	
Subject (Resource)	*http://vandana.dhingra/home/index.html*
Predicate (Property)	Author
Object (literal)	"Vandana "

Figure 1. Dividing the sentence into 3 parts

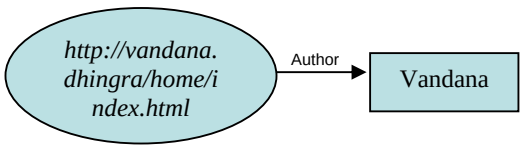

Figure 2. RDF Graph Representation

Vandana Dhingra is the creator of the resource
http://vandana.dhingra/home/index.html
can be represented in RDF by
<? xml version="1.0"?>
<rdf: RDF
xmlns:rdf=" *http://vandana.dhingra/2000/22* -rdf-syntax-ns#"
xmlns:s="http://description.org/schema/">
<rdf: Description about=" *http://vandana.dhingra/home/index.html* ">
<s: Creator>vandana</s: Creator>
</rdf: Description>
</rdf: RDF>

Figure 3. An Example -Coding in RDF Language

C. TRIPLES

RDF encodes in sets of Triples wherein each triple is the subject, predicate and object of an elementary sentence. Data is represented by subject-predicate-object triples [6] represented as <s.p.o> ie subject **s** has a predicate (or a relationship) , **p** with object **o**. Subject is the part that identifies the thing the statement is about. The part that identifies the property or the characteristics of the subject that the statement specifies is called the predicate. Object is the part that identifies the value of that property.

D. XML (Extended Markup Language), RDF/XML, XML Schema (XMLS)

RDF provides an XML based syntax called RDF/XML for recording and exchanging relationships. XML helps to create one's own tags, which has well defined meanings and it helps to add arbitrary structure to the documents but does not convey the meaning. The XML [7] standards give a syntactic structure for describing data.

E. ONTOLOGIES

Ontologies are the most important tool in knowledge representation, as they allow us to logically relate large amounts of data [5]. Ontologies are formal theories supporting knowledge sharing and reuse. They are used to explicitly represent semantics of semi-structured information. These enable sophisticated automatic support for acquiring, maintaining and accessing information. In the context of knowledge sharing, we use the term ontology to mean a specification of a conceptualization i.e. Ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents [3]. Ontology is a model of a particular domain, built for a particular purpose. Formally, ontology is the statement of a logical theory. Pragmatically, a common ontology defines the vocabulary with which queries and assertions are exchanged among agents.

1) Tools to Develop Ontologies

- Ontolingua,
- WebOnto,
- ProtegeWin,
- Adaptiva,
- Semantic Works 2008,
- COE, Conzilla2,
- HOZO,
- Onto Track,
- SWOOP,
- OntoSaurus,
- ODE and
- KADS22

F. WEB ONTOLOGY LANGUAGE (OWL)

It resulted in the development of OWL (Web Ontology Language). OWL is a set of XML elements and attributes, which have a standard meaning and are used to define terms and their relationships. It was designed to provide a common way to process the content of web information as well as to be read by the computers. OWL provides a schema over RDF and is a stronger language with greater machine interoperability than RDF.

G. DAML

The DARPA Agent Mark-up Language (DAML) Program started in 2000. DAML combines many language components of the Ontology Inference Layer (OIL) soon after it was started. The result of these efforts is DAML+OIL, a more robust language for general knowledge representation than RDF and RDFS. DAML is not a W3C standard, but many people in W3C participated in this program. DAML is kind of extension of RDF and RDFS, but it is not a data model. It not only provides stronger abilities to express constraints in schemas but also can build general knowledge representation, i.e. it is also an ontology language.

VII. STATE OF ART

Where is Semantic Web today? [10]
During the last few years, many improvements took place in semantic web related technologies.

- Improvement in the development status of ontology languages.
- Improved architecture of semantic web.
- Extensive work on semantic standards
- Approved RDF and OWL which provides a solid base to establish semantic applications.
- Development of domain –specific ontologies
- Evolution of tools for creating and publishing semantic information which makes it easier for non specialists to apply semantic web technology to their own fields /applications. Therefore, semantic web has opened a new window to various applications and systems that take benefit of machine understandable information and it has been widely accepted in IT branch with various research and industry projects originating from it.

VIII. RESEARCH ISSUES

Research issues being carried in various aspects like:

- How to effectively query
- How to develop Ontologies or reuse them and align and map between them? (major issue)
- How to construct a semantic web browser?
- How to develop crawlers for semantic web
- How to establish trust and provenance of the content?(Provenance-When, Where, and conditions)

REFERENCES

[1] Tim Berners-Lee , Nigel Shadbolt and Wendy Hall,"*The Semantic Web Revisited*," IEEE May/June 2006.

[2] Ram Mohan Rao, "*The future of search*," DIGIT Aug 2005.

[3] Michael Wilson and Brian, "*The Semantic Web: prospects and challenges,*"IEEE2006, http://ieeexplore.ieee.org/iel5/11087/35308/01678469.pdf?isnumber=35308&arnumber=1678469

[4] Frank van Harelen and Vrije, "*The Semantic Web: What, Why, How, and When,*"IEEE Distributed Systems Online March,2004,http://www.cs.vu.nl/~frankh/postscript/IEEE-DS04.pdf

[5] "*Using Ontologies in the Semantic Web:A Survey*" Li Ding, Pranam Kolari, Zhongli Ding, Sasikanth Avancha, Tim Finin, Anupam Joshi

[6] Sheila A. Mcllraith, Tran Cao Son and Honglei Zeng ," Semantic Web Services," IEEE Intelligent Systems 2001, http://ieeexplore.ieee.org/iel5/5254/19905/00920599.pdf

[7] C.Anantaram,"*Semantic-WebTechnology-the next generation internet,*"CSI Sept. 2006

[8] Yoo, Myaeng, Yun, Lee, "*Universal Information retrieval system in semantic web environment,*" IEEE 2005

[9] A Min , Andjomshoaa , Ferial and Roland ,"*Semantic Web: Challenges and new requirements*," IEEE 2005 CS(*Proceedings of 16th International Workshop on Database & Expert Systems Applications*)

[10] Li Ding and Tim Finin, "*Characterizing the Semantic Web on the Web*", in *Proceedings of the 5th International Semantic Web Conference*, (ISWC'06) November 2006

[11] Bhavani, Eric, Dock,"*Dependable Semantic Web*," IEEE CS 2002

[12] Li Ding, Tim Finin, Anupam Joshi, Yun Peng, Rong Pan, and Pavan Reddivari, *Search on the Semantic Web*, IEEE Computer, 10(38):62–69, 2005

[13] Martin Hepp "*Semantic Web and Semantic Web Services*," IEEE Internet Computer Society, March/April 2006.

[14] Michael C. Daconta, Leo J. Obrst, Kevin T. Smith, "*The Semantic Web-A guide to the future of XML, Web Services, and Knowledge Management*", *Wiley Publishing*", Inc.

[15] Li Ding and Tim Finin and Anupam Joshi and Rong Pan and R. Scott Cost and Yun Peng and Pavan Reddivari and Vishal C Doshi and Joel Sachs, "*Swoogle: A Search and Metadata Engine for the Semantic Web*" , in Proceedings of the 13th ACM Conference on Information and Knowledge Management , 2004

AUTHORS PROFILE

Ms.Vandana Dhingra, pursing Ph.d.from YMCAIE University have done M.tech(Information Technology),B.E. (Computer Science and Engineering) with First Class with distinction and is having experience of teaching in reputed Engineering colleges for 13 Years . She has administrative experience of being Head of Department of Computer Science and Engineering in Apeejay College of Engineering, Sohna for past five years. She has published research papers in IEEE Digital Library and various international and national conferences. Her subjects of Interest include Digital System Design, Operating Systems and Distributed Operating Systems. Her research interests are Semantic web and Ontologies.

Dr. Komal Kumar Bhatia received the B.E, M.Tech. and Ph.D. degrees in Computer Science Engineering with Honors from Maharishi Dayanand University in 2001, 2004 and 2009, respectively. Presently, he is working as Associate Professor in Computer Engineering Department in YMCA University of Science & Technology, Faridabad. He is guiding PhDs in Computer Engineering and his areas of interests are Search Engines, Crawlers and Hidden Web.