Search on Web- From Classical Web to Semantic Web

Jijy George, Sandhya .N., Suja George

Abstract*—The WWW is a vast information resource with enormous potential. The retrieval of relevant information from the web is a major issue because it is difficult for the machines to process and integrate the information. Internet is growing very fast as pages are added in a very fast pace. Searching on the web for a specific topic results in hundreds of pages and it is up to the user to extract the useful information from the result set. This paper presents insight on how current search engine works and also the potential gain of using current search engines. This paper further gives an overview of the challenges surrounding current search techniques and looks at the need of an intelligent information retrieval system on web. This paper also reviews the foundations required to make the search engine an intelligent one and also gives an insight on concepts like metadata, RDF, URI, XML, triples and ontologies

Index Terms ***:Semantic web, RDF, XML, Triples, Ontologies***

I. INTRODUCTION

The World Wide Web is the greatest depository of information ever assembled by man. It contains documents and multimedia resources related to most of the subject, and all of these data are readily available over internet to the users. The Web's successful usage is highly determent by its distributed design. The web pages are hosted by numerous computers, where each document may lead to other several documents, either on the same or different computers that have hosted/stored the required information. People all over the world can add content on the Web and make it to grow rapidly in its usage as more people start using it. Therefore it may warrant a situation where Web's size itself becomes a challenge. The large related data of information available may make it difficult situation to locate exact useful information out of it. If we do not know the exact URL, it is very difficult to locate information on the web. The web directories like Yahoo! And search engines like Google and Alta Vista can certainly assist to some extend but most times they are not definitive.

Search Engines are software programs that are used to search information or documents for specified keywords and then return a list of the documents where the keywords were found. If we do not know the exact URL, it is very difficult to locate information on the web. Therefore search engines are mandatory in our day to day life. Moreover it is an index of websites and it is created by special software called spiders or web crawlers [3].Search engines do not search the World Wide Web directly.

It searches the database of web pages selected from the millions/billions of pages residing on servers. Computer robot programs called spiders are used to select and create these databases .Once the spiders find pages; they pass it on to another computer program for Indexing. This program identifies the content, tests its link and stores it in the search engine's database. While you search the web using a search engine, you are always searching a stale copy of the real web page. When you click on links provided in a search engine's search results list, you retrieve from the server the current version of the page. There are three types of search engines. They are power-driven by robots [1,2] (crawlers, ants or spiders), power-driven by human submissions and the hybrid of these two. To find out the corresponding website and fetch the information crawler based search engines used crawlers [3] (automated software agents). It also find out site's meta tags (special HTML tag that provides information about a Web page) and follow the links. Finally Crawler will return all the information to a central depository where the data is indexed. Human powered search engines completely rely on human beings. The information which is submitted by human is put into index.

The users often use the Web not just to locate a document but may then also wish to have an analytical view on the information available. For example let's say if the user wants to get information not only on the price of the laptop available but also want to get the competitive price list of the laptop available. To complete these tasks, user has to visit series of pages, collate the information available to use it meaningfully. This is much beyond the capability of current available directories and search engines. So, could they eventually perform these desired tasks? The fact is that the Web was never designed to be manipulated 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. Further in those direction researchers have begun to explore the potential of associating web content with explicit meaning, in order to create a Semantic Web. Instead of relying on natural language processing to extract the meaning of the document, semantic web requires the authors to explain the document using a knowledge representation language.

II. HOW TO USE A SEARCH ENGINE

The working of each search engine is different. We can use the Help section of particular search engine to find out the best way of search.

Manuscript received March, 2013.

Prof. Jijy George, Department of Computer Science, St. Joseph's College (Autonomous), Shantinagar, Bangalore-560027, Karnataka.

Prof. Sandhya .N., Department of Computer Science, St. Joseph's College (Autonomous), Shantinagar, Bangalore-560027, Karnataka.

Prof. Suja George, Department of Computer Science, St. Joseph's College (Autonomous), Shantinagar, Bangalore-560027, Karnataka.

Search engines can be searched by entering keywords to the appropriate place. It index thousands or millions of websites. If you are entering more words, you will get the better result. While using search engines the following tips you can keep it in your mind

1. It will give you the result which is based on all the words you typed.
2. To get the exact result, better to use quotation marks around a phrase.
3. If you are using an advanced search, it will give more accurate/specific search options. For example, we can use Boolean expressions to find out a particular website or files.

2.1 How Search Engines Work

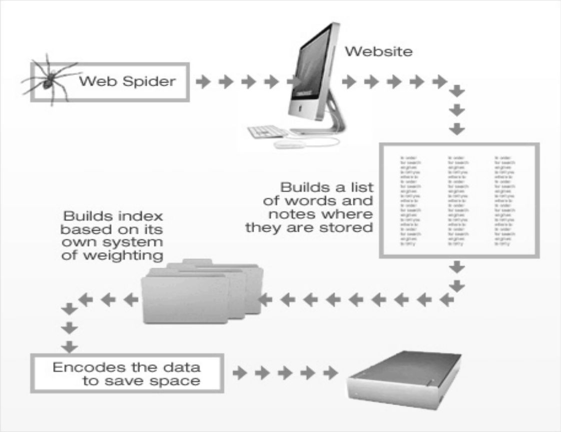
Web Spider
Website
Builds a list of words and notes where they are stored
Builds index based on its own system of weighting
Encodes the data to save space

Fig-1 How Search Engines Work

A search engine or Information retrieval system consists of the following four modules :

- A document processor [4]
- A query processor[4]
- A search and matching function[4]
- A ranking capability [4]

2.1.1 Document Processor

The document processor prepares, processes, and inputs the documents or sites against user's search. A well-formed, consistent format is required for further steps of processing. Therefore document processor merge all the data into a single consistent data structure so that it is easy to handle. It identifies potential index able elements in documents. Every search engine has its own rules. Document processor has to determine what action has to be taken by the tokenizer based on the rules which the corresponding search engine follows. In addition the document processor performs some or all of the following steps:

- ***Removing stop words***

A stop word list may consists of articles (*a, the*), conjunctions (*and, but*), interjections (*oh, but*), prepositions (*in, over*), pronouns (*he, it*), and forms of the "to be" verb (*is, are*). To remove these stop words we can use an algorithm that compares index terms in the documents against a stop word list and removes certain terms from index for searching.

- ***Stemming of terms.***

Stemming eliminates word suffixes and reduces the number of unique words in the index. It will also reduce the storage space for the index and increase the speed of the search process. For stemming, system may use either a strong stemming algorithm or a weak stemming algorithm. Through a strong stemming algorithm both inflectional suffixes and derivational suffixes are removed. Inflectional suffixes alone are removed through a weak stemming algorithm

- ***Index entries are extracted.***

From the original document the document processor extracts all the remaining entries through this step.

- ***Calculating and assigning weights.***

In this step each terms are assigned a weight depends on the search engine. Simple search engine use a binary value 1 for presence and 0 for absence. If the search engine is more sophisticated then the weighting scheme also more complex.

- ***Create Index or inverted file***

It is a data structure used to store the index information that will be searched for each query.

2.1.2 Query Processor

Query processors accept the query, select a plan for executing the query and then execute the chosen plan. Query processing may contain the following possible steps

- **Tokenizing.** A token is an alpha-numeric string which comes in between white space and/or punctuation. When a user inputs a query, the search engine must break it down into small understandable segments known as tokenizing.
- **Parsing the query**. Through this step the system needs to parse the query into query terms and operators. These operators may occur in the form of reserved punctuation or reserved terms in specialized format .Once this parsing is over, a search engine may take the list of query terms and search them against the inverted file.
- **Creating the query.** Depends on the search engine's query representation the system does its matching. If a statistically based matcher is used, then the query must match the statistical representations of the documents in the system. If a Boolean matcher is utilized, then the system must create logical sets of the terms connected by AND, OR, or NOT. After this point, a search engine may take the query representation and perform the search against the Indexed or inverted file.

2.1.3 Search and matching function

Depends on the model of information retrieval, the system will carry out their search and matching function. The commonly used technique is search the inverted file for documents which meets the query requirements.

2.1.4 A ranking capability

Here we are discussing about the features of a query which make for good matches

- **Frequency of term**: A particular query term appears how many times in a document are one of the most obvious ways of determining a document's relevance to a query.
- **Terms location:** If the terms are appearing on the title of a document or pages that match a query term are weighted more than terms appearing in the body of the document.

The query terms appearing in section headings or the first paragraph of a document may also be considered.

- **Link analysis:** Link analysis is purely based on how each page is well-connected according to the definition by Hubs and Authorities. Hub documents contains a large no of 'out links' and Authority documents have a high number of 'in-links'
- **Popularity**: Popularity helps to utilizes data on the frequency with which a page is chosen by all users.
- **Date of Publication:**. The search engines present results beginning with the most recent

2.2 Advantages of search engines

1. Since it is created by software its indexes may be very large.
2. Search is done by entering keywords in the appropriate place.
3. Comparing to subject directories, it is much larger.
4. For each search you will get a large no of results.

2.3 Disadvantages of search engines

1. Sites are not evaluated for quality or relevance. Based on the keywords it is randomly selected and indexed.
2. While selecting keywords we need to be more specific.
3. It may result many pages to check.
4. Sites/pages may not be arranged irrelevant order. Sometimes the most useful sites may come at the end of the list. There are chances of getting irrelevant sites also.

III. NEED OF INTELLIGENT SEARCH ENGINE

So far we discussed about traditional search engine. Traditional search engine works purely on keyword based. This may result in the presentation of irrelevant information to the user. Therefore a new technology called "Semantic Search" has developed .In this search the conceptual meaning of the query is considered rather than the literal meaning of the keyword. Moreover it has the ability to operate without any artificial means. It is not depending the ranking of the page, keywords or tags and concentrate more on the contextual meaning of the query. Certain extend its ability to understand the query may come close to the human brain of thinking .Its search results are more of real time results. Semantic search engine understands the context of the query in a better way and give relevant result. The traditional search engines like Google, Yahoo, Bing etc. matches the queries based on keywords, ranking of the page and inbound link measurement algorithms. But a semantic search engine is done its search based on the meaning and semantics of the query, so that you may get a better relevant result.

IV. HISTORY OF SEMANTIC WEB

The concept of the *Semantic Network Model* was formed in the early sixties by the cognitive scientist Allan M. Collins, linguist M. Ross Quillian and psychologist Elizabeth F. Loftus in various publications as a form to represent semantically structured knowledge[14]. It extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users. The term was coined by Tim Berners-Lee the inventor of the World Wide Web and director of the World Wide Web Consortium ("W3C"), which oversees the development of proposed Semantic Web standards. He defines the Semantic Web as "a web of data that can be processed directly and indirectly by machines."

Many of the technologies proposed by the W3C already existed before they were positioned under the W3C umbrella. These are used in various contexts, particularly those dealing with information that encompasses a limited and defined domain, and where sharing data is a common necessity, such as scientific research or data exchange among businesses. In addition, other technologies with similar goals have emerged, such as microformats.

V. ADVANTAGES OF SEMANTIC SEARCH OVER TRADITIONAL SEARCH

Semantic Web is an information repository where concepts in documents are linked to similar concepts in other documents. In contrast in Classical web the hypertext documents are linked by a set of keywords. Here a web page is linked to another page with the help of HTML anchor tag, usually represented as underlined text, which is a keyword reference that maps a word or phrase in one document to another .This linking doesn't give any indication about the real relationships between the documents.

In semantic web a document is linked to another document by relating the concept in one document with a concept in another document. This can be achieved in semantic web by using a new standard, the Resource Description Framework (RDF). In fact, RDF can be used to link any entities or concepts together.

The fundamental idea behind the Semantic Web is that here the web is a collection of concepts that are linked together and is not a collection of documents that are linked together.

5.1 Morphological variations of tenses and plural forms can be handled very easily

Semantic search can handle morphological variations easily. For example when we type words like read, reading or read will all lead to exactly same page. There will be no difference between "how to read an article?" or "reading an article."

5.2 Differentiating words and put their right meaning according to the context

Semantic search has special algorithm to differentiate between words and put their correct meaning for the context. At the same time it will give almost similar results for synonyms which match with the current context. For example words like cure, treat or heal as synonyms so results yielding to "cure for Diabetics" will be same as results to "treat/treatment for Diabetics" and "heal for diabetics."

5.3 Appropriate question answering system: Traditional systems are mostly based on the ranking of the page and it will display the related web pages or documents. But semantic search is having a question answering system and it will give a single result which describes the query.

5.4 Concept and knowledge Matching: When the user gives a query in semantic search engine, it will display all the possible solution to the concept of that query .For example, if you have a query "Installation of Windows 95?" the results shows the installation of windows 95 and the consumer preview of windows 95 .

5.5 Capable for bring down the exact sentence for the given query: Semantic search has the ability to bring down the exact paragraph/sentence from a huge content. For example, content describes the features of IBM ThinkPad X32 notebook [5] which contains the description about its processor, I/O, memory, applications and other related features. Imagine that you have a query "what is the processor in IBM ThinkPad X32 notebook?" then the semantic search will directly point you to paragraph which describes the processor of IBM ThinkPad X32 notebook.

VI. COMPONENTS OF SEMANTIC WEB

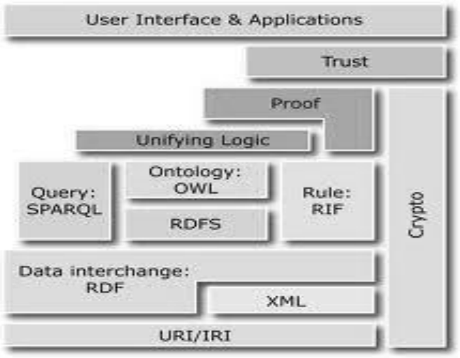

Fig-2 Components of semantic web

6.1 Metadata

Metadata is information about information which can widely use in real-world for searching. For example, you want to buy something from a shop; the shop owner will give you a catalog to provide a lookup system which allows you to find items along with their price, type, etc. Metadata will help us to make searching easier and faster. 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.

6.2 The Resource Description Framework

The RDF (Resource Description Framework)[8][11] is a language for describing information and resources on the web. The information in the web can be put into RDF files. The intelligent agents (web spiders) that travel through the web search, discover, pick , collect, analyze and process information from the web. The Semantic Web uses RDF to describe web resources.RDF are designed to be read and understood by machines.RDF is not designed for being displayed to people.RDF is written in XML.

6.3 Resource Description Framework is a framework for manipulating metadata and it explains relationships among resources with properties and values. The rules used to build are:

Resource: Whatever an RDF expression is called a resource. All resources have a URI and it may be an entire web page or a part of a web page

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.

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

6.4 Ontologies

On the Semantic Web, ontologies [8] are used to define the concepts and the relationship between these concepts for a specific problem. Ontologies are used to characterize the concepts that can be used in a specific area of concern, classify possible relationships, and define possible checks on using those concepts. In reality ontologies can be very complex (with several thousands of terms) or very simple (describing one or two concepts only).

Ontologies are the basic building blocks for inference techniques on the Semantic Web [8].

6.4.1 Role of Ontologies in Semantic Web

The ontologies on the Semantic Web help data assimilation. When ambiguities exist on the terms used in the different data sets, or when a bit of extra knowledge may lead to the discovery of new relationships, ontologies are helpful. Consider, for example, the application of ontologies in the field of medicine. Doctors use them to represent knowledge about symptoms, diseases, and treatments. Pharmaceutical companies use them to embody information about drugs, dosages, and allergies. Combining facts from the medical and pharmaceutical professionals with patient data results in a wide range of intelligent applications such as decision support tools that search for possible treatments, systems that monitor drug efficiency and possible side effects, and tools that support epidemiological research.

The application decides whether to use complex or simple ontologies., Some applications need an contract on general terminologies, without any firmness imposed by a logic system. Finally, some applications may need more complex ontologies with complex reasoning procedures. It all depends on the need and the goals of the applications.

There are various techniques to describe and define different forms of ontologies in a standard format. These include RDF and RDF Schemas [8], Simple Knowledge Organization System (SKOS)[8],Web Ontology Language (OWL)[8], and the Rule Interchange Format (RIF)[8]. The choice among these different technologies depends on the complexity and strictness required by a specific application.

6.4.2 Ontology Example

A general example may help. A bookseller may want to combine data coming from different publishers. The data can be imported into a common RDF model, e.g., by using converters to the publishers’ databases. However, one database may use the term “author”, whereas the other may use the term “creator”. To make the integration complete and extra definition should be added to the RDF data, describing the fact that the relationship described as “author” is the same as “creator”. This extra piece of information is ontology, though this is a simple one.

In a more complex case the application may need a more detailed ontology as part of the extra information. This may include formal description on how authors are to be uniquely identified (eg, in a US setting, by referring to a unique social security number), how the terms used in this particular application relate to other datasets on the Web (eg, Wikipedia or geographic information), how the term “author” (or “creator”) can be related to terms like “editors”, etc.

6.5 Triples

RDF codes in groups of Triples where each triple is the subject, predicate and object of a basic sentence. Data is represented by subject-predicate-object triples [11] represented as <s.p.o> ie subject **s** has a predicate (or a relationship) , **p** with object **o**. Subject refers to the statement meant. Predicate identifies the property or the characteristics of the subject that the statement specifies.

Object is the part that identifies the value of that property.

6.6 XML (Extended Markup Language), RDF/XML, XML Schema (XMLS)

RDF supports 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 [13] standards give a syntactic structure for describing data.

6.7 OWL

The **Web Ontology Language** (**OWL**)[8] is a family of knowledge representation languages for writing ontologies. The languages are characterized by formal semantics and RDF/XML-based serializations for the Semantic Web. OWL is endorsed by the World Wide Web Consortium and has attracted academic, medical and commercial interest.

6.8 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. WHERE IS SEMANTIC WEB TODAY?

During the last few years, many developments are taking place in semantic web related technologies. There are developments happening in the status of ontology languages. The architecture of semantic web is being improved. Extensive work is done on semantic standards. The RDF and OWL standards are approved and therefore it provides solid base to establish semantic applications. Domain specific ontologies are being developed. Tools for creating and publishing semantic information are developed .This makes it easier for non specialists to apply this technology in their own fields.

The semantic web has opened a new window to various applications and systems that take advantage of machine understandable information and it has been widely accepted in various areas of research and projects.

7.1 Top 10 best semantic search engines[6][7]

The best 10 semantic search engines are Hakia, Kosmix, SenseBot, Kngine, Cognition, Swoogle, DuckDuckGo , Cluuz,Factbites, Evri

VIII. FUTURE OF SEMANTIC WEB

This technology has commercial applications. The companies like Intelliseek are trying to give users with business intelligence over the Web, including mining corporate Web sites for competitive intelligence information.

Use of search agents online -- infobots, scouring the Web and sifting relevant sites for users -- would alone provide a "major boost in productivity at work and at home,"[15].

Some serious computer scientists, although cautious about the promise of the Semantic Web, are ultimately optimistic that it will be everything developers are hoping for -- an online source for all of the knowledge that humanity has created in science, business and the arts.

"Of course, hype always outruns reality in these things," [15]. "But in this case, I think reality will plug along and catch up."[15]

IX. CONCLUSION

The retrieval of information from current search engines has certain issues that have to be taken care. The successful usage of classical or traditional web is highly determent by its distributed design. The web pages are hosted by numerous computers, where each document may lead to other several documents, either on the same or different computers that have hosted/stored the required information. People all over the world can add content on the Web and make it to grow rapidly in its usage as more people start using it. Therefore it may warrant a situation where Web's size itself becomes a challenge. The large related data of information available may make it difficult situation to locate exact useful information out of it. So there is a need of an intelligent search engine which satisfies the high expectations of users exists. The development of semantic web goes in this direction. The semantic web has opened a new windowpane to various applications and systems that take advantage of machine understandable information and it has been widely accepted in various areas of research and projects. Semantic web provide a "major boost in productivity at work and at home".

ACKNOWLEDGMENT

A work of this nature requires the blessings and patience of a number of people. We would like to acknowledge the guidance of all our colleagues in this work. We would also like to thank god for his blessings and acknowledge his presence in all our efforts.

REFERENCES

1. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5655402&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D5655402
2. Search Engine Tutorial. Berkley (CA) University Library http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SearchEngines.html
3. http://www.webopedia.com/DidYouKnow/Internet/2003/HowWebSearchEnginesWork.asp
4. http://www.infotoday.com/searcher/may01/ liddy.htm
5. http://www.notebookreview.com/ default.asp? newsID=2285
6. http://www.webgranth.com/top-10-semantic-search-engines-best-alternative-to-google-search-engine-to-get-more-accurate-results
7. Research Buzz Web Site. http://www.researchbuzz.com/
8. http://www.w3.org/standards/ semanticweb/ontology
9. "Using Ontologies in the Semantic Web:A Survey" Li Ding, Pranam Kolari, Zhongli Ding, Sasikanth Avancha, Tim Finin, AnupamJoshi
10. C.Anantaram,"Semantic-WebTechnology-the next generation internet,"CSI Sept. 2006
11. 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
12. C.Anantaram,"Semantic-WebTechnology- the next generation internet,"CSI Sept. 2006
13. 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
14. http://www.wikipedia.org/wiki/ *Semantic_Web*
15. http://www.technewsworld.com/ *story/31199.html*