Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Analysis of Exploring Information from Search Engines in Semantic Manner

Prathyusha Kanakam[1], S. Mahaboob Hussain[2], Dr. Sumit Gupt[3], Dr. D. Surya Narayana[4]
[1,2]*Assistant Professor of CSE, Vishnu Institute of Technology, Bhimavaram, India,*
[3]*HOD & Professor of CSE, Vishnu Institute of Technology, Bhimavaram, India,*
[4]*Principal & Professor of CSE, Vishnu Institute of Technology, Bhimavaram, India,*

Abstract— In this Internet Era, every one having web on their particular gadgets like mobiles, laptops, computers, tabs and so on in order to explore the information what they looking for. But, the present search engines are not providing the relevant information exactly. To overcome this problem, this paper proposed a concept of search engines which evaluates the search queries given by user and react accordingly to furnish the information in a semantic manner i.e., based on the meaning of the search query. This Semantic searching of Web will provide intelligent access to heterogeneous, distributed information, enabling software products (agents) to mediate between user needs and the information sources that are available for enabling users to find, share, and combine information more easily. This new technology may apply in the fields of knowledge management, content mining of web and electronic commerce. This paper proposed the concept of semantic web-enabled web services which will help to bring the semantic web to its full potential.

Keywords— content mining; search engines; semantic search; semantic web;

I. INTRODUCTION

There is a need of search engines in this e-world. For example there is lots of great and useful information in a book library, but it's impossible to examine all the books personally. Not even the most indefatigable web-surfer could hyperlink to all the documents in the aptly named World Wide Web. There are billions of pages on the Web.

There are some software programs like robots, spiders or crawlers which are used to examine the documents in the web, this job will be done by search engines. A robot is a piece of software that automatically follows hyperlinks from one document to the next around the Web. A robot sends information back to its main site when it discovers a new site and it will be indexed. These robots are also used to update previously catalogued sites. Spiders and bots are software programs used by search engines which are used to survey the Webs and help to build their databases. These programs retrieve web documents and it was analysed. Search engine index is built by the data collected from the web pages. The query is searched from a search engine site, by searching the index of search engine of all analysed web pages. The best URLs are then returned as hits and ranked in order with the best results at the top [8].

There are different strategies and different type of searching methodologies and some of them are discussed below.

A. Key word Searching

Most of the search engines do their text query and retrieval using keywords. The most common searches on the web is in the form of text searches. Full-text indexing systems generally pick up every word in the text except commonly occurring stop words such as "a," "an," "the," "is," "and," "or," and "www." Some of the search engines discriminate upper case from lower case; others store all words without reference to capitalization.

Some search engines handle words and simple phrases. In its simplest form, text search looks for pages with lots of occurrences of each of the words in a query, stop words aside. The more common a word is on a page, compared with its frequency in the overall language, the more likely that page will appear among the search results. Hitting all the words in a query is a lot better than missing some.

Search engines also make some efforts to "understand" what is meant by the query words. For example, most search engines now offer optional spelling correction. And increasingly they search not just on the words and phrases actually entered, but they also use stemming to search for alternate forms of the words (e.g., speak, speaker, speaking, spoke).

When ranking results, search engines give special weight to keywords that appear:

- High up on the page
- In headings
- In BOLDFACE (at least in Inktomi)
- In the URL
- In the title (important)
- In the description
- In the ALT tags for graphics.
- In the generic keywords meta tags (only for Inktomi, and only a little bit even for them)
- In the link text for inbound links.

More weight is put on the factors that the site owner would find it awkward to fake, such as inbound link text, page title (which shows up on the SERP -- Search Engine Results Page), and description.

The Problems with Keyword Searching is:

- Keyword searches have a problem to distinguish between words that are spelled the same way, but mean something different (i.e. pen drive, pool drive, car drive and etc,). This often results in hits that are completely irrelevant to our query.
- Search engines also cannot return hits on keywords that mean the same, but are not actually entered in your query. A query on heart disease would not return a document that used the word "cardiac" instead of "heart."
- The main disadvantages of Keyword searching are:
- Effective keyword searching requires some training and practice in using the search protocols.
- Most words have many synonyms and related concepts, thus keyword searches will probably retrieve unrelated records that are not related to the topic.

B. Basics and Advance Searching

There are two different types of searches in the sites like basic and refined or advanced. In the basic search we simply enter keywords without using dropdown menus with additional options. Basic search may be some times quite complex depending upon search engines. But refining options of an advanced search is differ from search engine to another, it includes to give more weight to one search term than you give to another, the ability to search on more than one word and this advance search also to exclude words that might be likely to muddy the results.

To refine searches many of the search engines allow using Boolean operators. These are the logical terms AND, OR, NOT, and the proximal locators, NEAR and FOLLOWED BY.

- Boolean AND makes all the terms specified must appear in the documents, i.e., "heart" AND "attack."
- Boolean OR makes at least one of the terms specified must appear in the documents, i.e., bronchitis, acute OR chronic.
- Boolean NOT makes at least one of the terms specified must not appear in the documents.
- Some Search engines use the characters + and - instead of Boolean operators to include and exclude terms.
- NEAR means that the terms entered should be within a certain number of words of each other.
- FOLLOWED BY means that one term must directly follow the other. ADJ, for adjacent, serves the same function.
- Phrases: is a very important task in a search engine having ability to query on phrases. Those that allow it usually require that you enclose the phrase in quotation marks, i.e., "space the final frontiers."

C. Relevancy Rankings

Basing upon confidence or relevancy rankings most of the search engines will return results. Search engines list the hits according to how closely they think the results match the query. Frequency and the positioning of keywords are considered in some search engines to determine relevancy, reasoning that if the keywords appear early in the document, or in the headers, this increases the likelihood that the document is on target. Another method is to determine which documents are most frequently linked to other documents on the Web.

D. Information on Meta Tags

Some search engines are now indexing Web documents by the Meta tags in the documents' HTML. There is a lot of conflicting information out there on meta-tagging. The different search engines look at Meta tags in different ways. Some rely heavily on Meta tags; others don't use them at all. The general opinion seems to be that Meta tags are less useful than they were a few years ago, largely because of the high rate of spamdexing (web authors using false and misleading keywords in the Meta tags).

The words that appear at the top of the document more highly scored than the words that appear at the bottom by most of search engine algorithms. Some search engines give preference to the words that appear in HTML header tags (H1, H2, H3, etc). Which helps to provide a file name to your page that constitute one of your prime keywords, and to include keywords in the "alt" image tags.

All the major search engines have slightly different policies. While designing a website and meta-tagging the documents, one should check out what the major search engines say in their help files about how they each use Meta tags for better accessing.

E. Concept Based Searching

Concept-based search systems try to determine what it mean, not just what it say, unlike keyword search systems. Concept-based search returns hits on documents that are "about" the subject/theme that explores, even if the words in the document don't precisely match the words you enter into the query. Clustering systems are builded by using different methods which are very complex, depends on sophisticated linguistic and artificial intelligence theory and so on. and the software associated with it determines meaning by calculating the frequency with which certain important words appear or When several words or phrases that are related to a particular concept also appear close to each other in a text. The search engine concludes that the piece is "about" a certain subject by statistical analysis.

F. Page Ranking

Search engine ranking algorithms are maintained secretly by the companies, for at least two reasons: to protect their methods from their competitors and they also want to make it difficult for web site owners to manipulate their rankings. That said, a specific page's relevance ranking for a specific query currently depends on three factors:

- Its relevance to the words and concepts in the query.
- Its overall link popularity.
- Whether or not it is being penalized for excessive Search Engine Optimization (SEO).

II. DIFFERENT TYPES OF SEARCHES

In these e-world days, Search Engine plays a crucial role in retrieving and systematically arranging relevant data for various purposes. But search results are giving irrelevant and redundant information to the search queries given by the user. Here, Content Mining of Web [1] takes its part in producing the search results in semantically manner. Basically, the way we search a query is of two types: Semantic Search and syntactic search as shown in below Fig. 1.

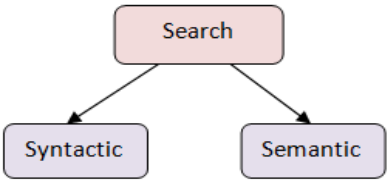

Fig. 1. Classification of Searching

A syntactic web search engine [2] is designed to search for information may consist of web pages, images, information and other types of files on the World Wide Web and FTP servers and the search results presented as a list are called hits. Some search engines also mine data available in databases or open directories. These search engines operate algorithmically and web directories are maintained by human editors so they are the mixture of algorithmic and human input.

Semantic search is technique for searching of data in which a search query aims not only to find keywords, but also to determine the intent and contextual meaning of the words that a person used for searching. It evaluates and understands the search phrase to provide meaningful search results and to find the most relevant results in a website, database or any other data repository.

In 2001 the inventor of the World Wide Web, Tim Berners-Lee, described his vision of the Semantic Web in an article co-authored with James Hendler and Ora Lassila [3] (Berners-Lee et al., 2001). The authors imagined that the Semantic Web would bring structure to the content of Web pages and enable computers to perform sophisticated tasks for people. Inspired by this vision, researchers throughout the world have been engaged in researching about the Semantic Web over the past few years.

III. SEARCH ENGINES AND THEIR WORKING

Web search engine is a software system designed for searching information on the World Wide Web ((WWW) which is a system that interlinks hypertext documents accessed via the Internet) and the search results are often referred as search engine results pages (SERPs) that contains information about web pages, images, information and other types of files and some search engines also mine data available in databases or open directories.

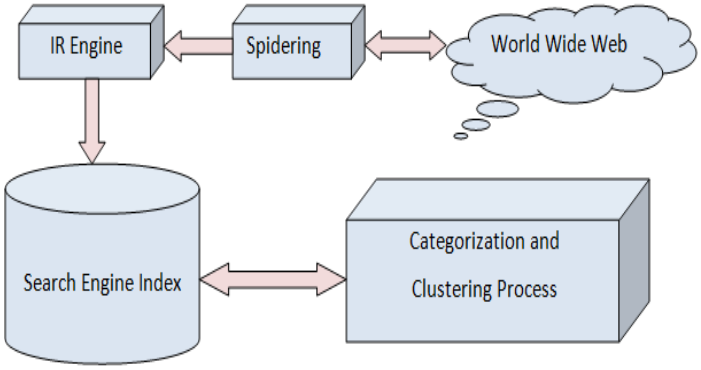

Fig .2. Architecture of Search Engine

Fig. 2 shows the architecture of search engine and it briefly explains how a search engine works. A search engine operates in the following order:

A. Web Crawling

The World Wide Web is automatically browsed by a computer program in a methodical way is called a Web Crawler. The antonyms of Web Crawlers are ant, bot, worm or Web spider. The process of scanning the WWW is called Web crawling or spiderling. To provide up-to-date data to the users Web Crawling is used by Search engines. Bots creates a copy of all the visited pages for later processing by a Search Engine and then it will then index the downloaded pages in order to provide fast searches.

IR (Information Retrieval) Engine is responsible for retrieving and indexing of queries from the user and based on the input given by the user. Firstly, the search engine stores information about many web pages, which are retrieved from the HTML mark up of the pages [4] by a Web crawler (sometimes also known as a spider) which is a special part of search engine that crawls every link on the site by browsing the internet (the crawling is the process of traversing the web by repeatedly following hyperlinks and storing downloaded pages for subsequent processing).

Data about these web pages are stored in an index database for use in later queries and then analyzes the contents of each page to determine how it should be indexed (for example, words can be extracted from the titles, page content, headings, or special fields called meta tags). Search engines, like Google, store all or some part of the source page (referred to as a cache) as well as information about the web pages, whereas others, like AltaVista, stores every word of every page they find.

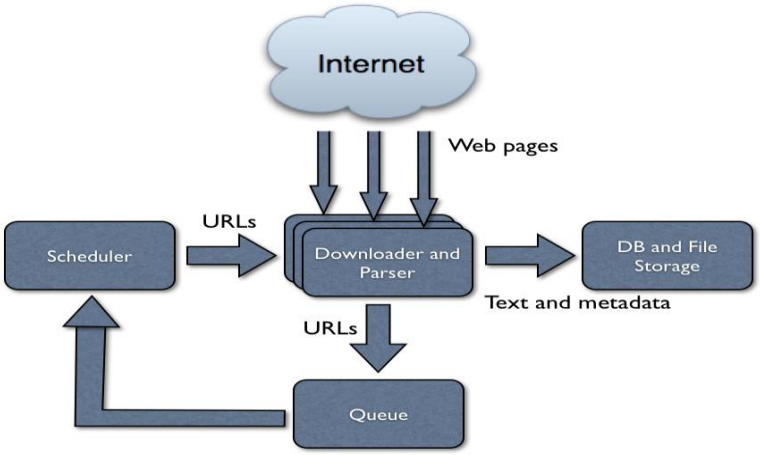

Fig. 3. Web Crawler Artichtecture

Web Crawlers are also used for automating tasks on websites such as checking links or validating HTML code as in Fig. 3. It usually starts with a list of URLs to visit which are called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit [11]. Before a search engine can tell you where a file or document is, it must be found. When a spider is building its lists, the process is called Web crawling [12].

B. Indexing

Indexing is the second major step that a search engine takes to deliver information and maintains a copy of all the content during the crawl process, and stores it in an index for easy retrieval [13, 14] as shown in Fig. 4.

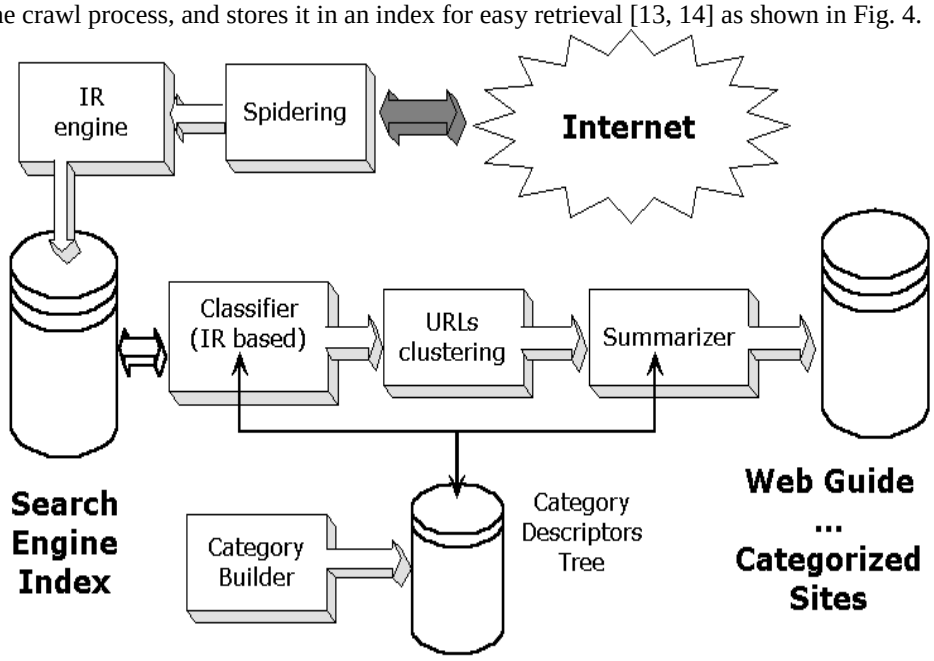

Fig. 4. Automated Categorization and Abstracting of Web sites
(Source from [13])

Search engine indexing is the process of collecting, parses and stores of data for later use by itself [5]. The actual Search Engine Index is the place where all the data is stored that is collected by the search engine. It provides the results for search queries, and results pages within it. Without a search engine index, the search engine would take considerable amounts of time and effort each time a search query was initiated, as the search engine would have to search not only every web page or piece of data that has to do with the particular keyword used in the search query, but every other piece of information it has access to, to ensure that it is not missing something that has something to do with the particular keyword.

IV. SEARCH ENGINES AND THEIR USAGE STATISTICS

There are different Search Engines available in the market and some of them are presented in this paper and they are tabulated in Table I along with their description.

TABLE I. LIST OF SEARCH ENGINES

Search Engine	Description

Google	Google is the world's most popular search engine and it was launched in 1997.
bing	Bing Search: Its an Microsoft's product and it entered into the burgeoning search engine market.
YAHOO! SEARCH	Yahoo! Search: The 2nd largest search engine on the web
altavista	AltaVista: It as built by researchers at Digital Equipment Corporation's Western Research Laboratory in 1995,. From 1996 powered Yahoo! Search, since 2003 - Yahoo technology powers AltaVista.
cuil	Cuil: Cuil was a search engine website (pronounced as Cool) developed by a team of ex-Googlers and others from Alta vista and IBM. Cuil, termed as the 'Google Killer' was launched in July, 2008 and claimed to be world's largest search engine, indexing three times as many pages as Google and ten times that of MS. Now defunct.
excite	Excite: Now an Internet portal, was once one of the most recognized brands on the Internet. One of the famous 90's dotcoms.
Go	Go.com: The Walt Disney Group's search engine is now also an entire portal. Family-friendly!
HOTBOT	HotBot was one of the early Internet search engines (since 1996) launched by Wired Magazine. Now, just a front end for Ask.com and MSN.
alltheweb	AllTheWeb: Search tool owned by Yahoo and using its database, but presenting results differently.
galaxy	Galaxy: More of a directory than a search engine. Launched in 1994, Galaxy was the first searchable Internet directory. Part of the Einet division at the MCC Research Consortium at the University of Texas, Austin
AOL Search	search.aol: Now powered by Google. It is now official.
Live Search	Live Search (formerly Windows Live Search and MSN Search) Microsoft's web search engine, designed to compete with Google and Yahoo!. Included as part of the Internet Explorer web browser.
LYCOS	Lycos: Initial focus was broadband entertainment content, still a top 5 Internet portal and the 13th largest online property according to Media Metrix.
GIGABLAST	GigaBlast was developed by an ex-programmer from Info seek. Giga blast supports nested boolean search logic using parenthesis and infix notation. A unique search engine, it indexes over 10 billion web pages.
Alexa	Alexa Internet: A subsidiary of Amazon known more for providing website traffic information. Search was provided by Google, then Live Search, now in-house applications run their own search.

The below Fig.5 shows the current statistics (January 2014 to April 2014) of the desktop search engines usage by the users. Search engines are used for the various purposes in the present era. Many of the search engines are coming into the picture by providing the best results for the quires. So, there is lot of changes occurring in the usage statistics. Google search engine is the most frequent search engine used by the users. But, slowly the usage is coming down year by year due to the other search engine showing their efficiencies in generating the relevant results for the queries.

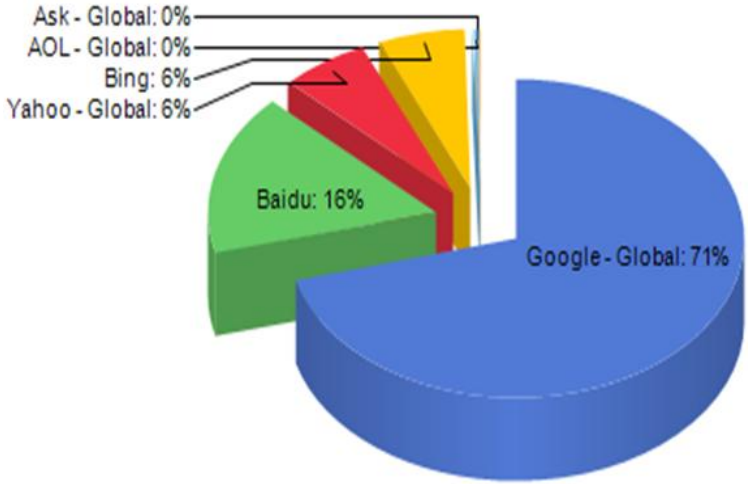

Fig. 5. Search Engine usage statistics for Jan 2014- April 2014.

The below Table II shows the correspondence changes in the search engines from January 2014 to April 2014.

TABLE II. COMPARISION OF SEARCH ENGINES USAGE

Search Engine	Year 2013	Year 2014
Google - Global	79.82%	70.30%
Baidu	6.40%	16.37%
Yahoo - Global	7.03%	6.16%

Bing	4.89%	6.09%
AOL - Global	0.35%	0.26%
Ask – Global	0.47%	0.15%
Excite - Global	0.02%	0.04%

The below Fig.6 gives a brief idea about the usage statistics of each search engines in the present market and it is calculated the average users of each search engine. For example Google have 79.82% in 2013 and 70.30% in 2014. It clearly shows the decrement of the users for Google search engines by 9.52 % for a year. Likewise statistics has being changed for all the search engines users due to the competition between the search engines by providing relevant information to the given queries like the semantic search engines.

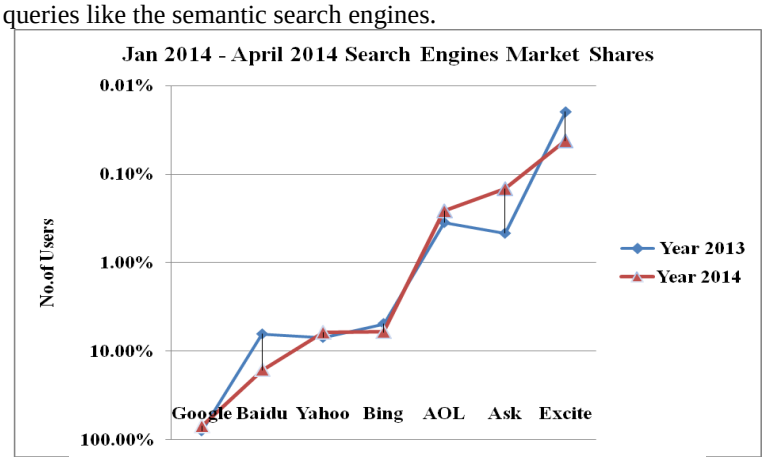

Fig. 6.Jan 2014- April 2014 Search Engines Market Shares.

The effectiveness of each search engine as illustrated in the diagram is calculated using two metrics. They are:

Time-On-Site: it depends upon how long the average visitors remain on the site after a search engine referral. By this measure one can say, how relevant these search engine results are.

Page-Depth: It depends upon average number of pages registered by an user.

V. SEMANTIC SEARCH ENGINES

Usually search engine finds the keywords in the web pages. If they are relevant to the search query that webpage will be shown in search results. Whereas semantic search not only look for keywords but it also do contextual analysis, (Fig. 5). It checks whether the query meaning is actually matching with content of the post. For this contextual analysis Google uses a knowledge database which contains all the queries searched by people. Normally people use different truncated forms of the same query. The aim of the search engine is to answer the user's query and with semantic search you are able to get exactly what are you are looking for.

Fig 5: Semantic Search

Semantics is the process of communicating enough meaning to result in an action. Semantic search [6, 7] is the ability of a search engine to determine what you mean when you search for something and provide you with results that don't necessarily match the words you used in your search query.

There are three features of semantic search:

- Turning documents into meaningful interchangeable data,
- Reflects a rising use expectation nurtured by modern technology.
- Presents a unique challenge for its enabling technologies.

Semantic search is an application of the Semantic Web to search. Search is both one of the most popular applications on the Web and an application with significant room for improvement. We believe that the addition of explicit semantics

can improve search. Semantic Search attempts to augment and improve traditional search results (based on Information Retrieval technology) by using data from the Semantic Web.

The below Fig. 6 illustrates a semantic Search Engine with an example. When keyword is given as input to the semantic search engine it produces the results according to the meaning of that particular keyword or a query posed by the user. Here, user searches for the keyword: apple and the Search Engine observes the meaning of the keyword and searches for that keyword in various databases like apple as software company, apple as fruit, apple as Clothing Store and apple as import/export business.

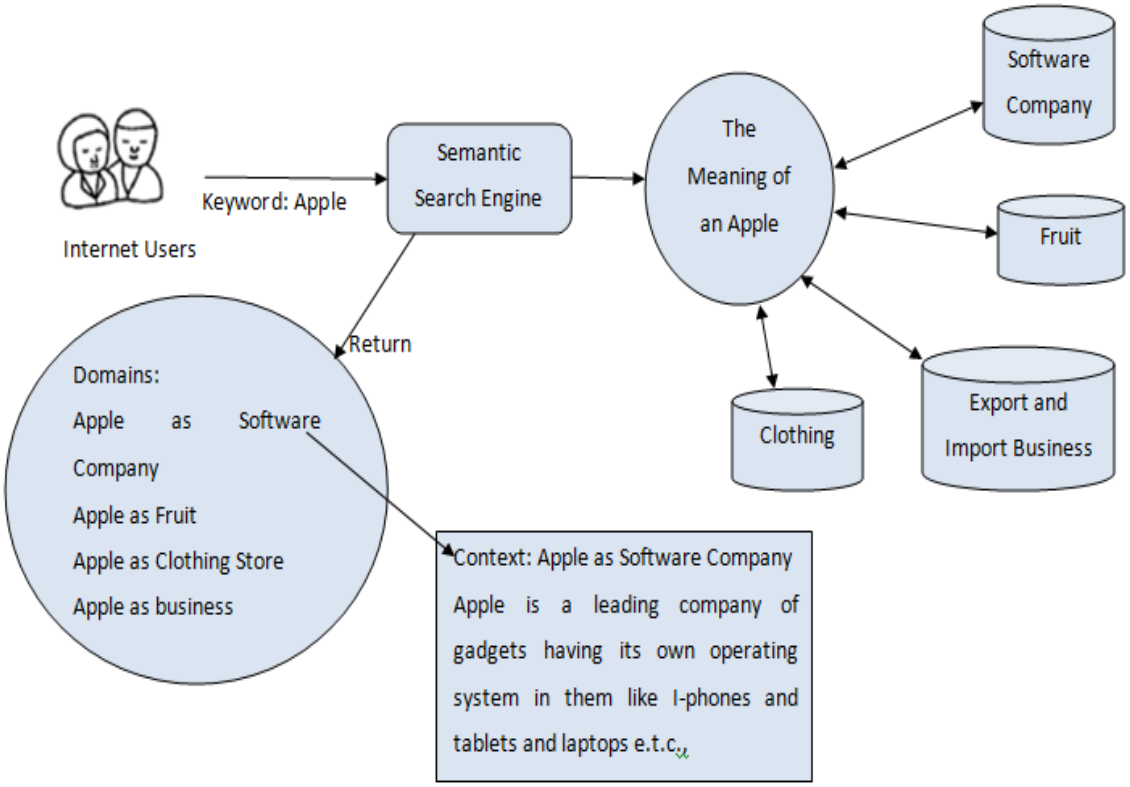

Fig 6: Working of Semantic Search Engine

Then content of particular web page that the user posed was returned to the user in an easier manner.

A. Semantic Search in Various applications:

As discussed earlier Semantic Search may used in the fields of Knowledge Management and Electronic Commerce.

1) Knowledge Management: Efficient management of knowledge plays key role in maintaining the competitiveness of organizations. Traditional knowledge management is now facing new problems triggered by the web; information overload, inefficient keyword searching, heterogeneous information integration and geographically-distributed intranet problems. These problems will be tackled by the modern technology known as Semantic Web Technology.

2) E-Commerce: The three main challenges in applying semantic web technology to Electronic commerce: Efficient Alignment of Ontologies, Versioning of Ontologies, and population Ontologies.

The below Table II represents various types of search engines and their applications in different fields with distinctive names.

B. Different Semantic Search Engines available in market:

The ideal search engine [10] should match the search queries to the exact context and return results within that context.

The next section describes briefly about the Semantic Search Engines which are the main prominent for the users for the relevant results and the reason for changes in statistics of users.

This statistical data report is collected from the browsers of site visitors by the exclusive on demand networks of share PostClients and the HitsLink analytics of market share statistics for internet technologies [16].

TABLE III. VARIOUS TYPES OF SEARCH ENGINES

Search Engine Name	Description of the Search Engine	Specialty
VikiTron	Semantic mathematics, chemistry and knowledge engine.	Math, numbers, chemistry, geography
Firmly	Specializes in auto-tagging companies websites in economic sectors	Business search engine
Invention Machine's Goldfire	Specializes in surfacing concepts from various document types, enterprise applications, technical and deep web sites, worldwide patent literature, consumer	Decision engine for researchers, engineers, and scientists.

	sentiment and more.	
Sophia Search Limited	Specialises in auto-tagging of content for semantic search and discovery	Search engine
Symbolab	Specializes in scientific search	Scientific search engine
True Knowledge (now Evi)[9]	Specialises in knowledge base and semantic search	Answer engine
Yummly	Semantic web search engine for food, cooking and recipes	Food related
Browse.lt	Fast and comfortable web search engine	Web, video, images, news
Swoogle	Searching over 10,000 ontologies	Semantic web ontologies. Indexes over 4 million semantic web documents.
Falcons	Full semantic search engine	Provides keyword-based search for objects, concepts (classes and properties), ontologies, and RDF documents on the semantic web.
International Digital Media Archive	Semantic document search engine	Archive of semantic metadata extracted from thousands of documents.

While Google, Yahoo and Live continue to hold sway in search, here are the engines that take a semantics (meaning) based approach, the end result being more relevant search results which are based on the semantics and meaning of the query, and not dependent upon preset keyword groupings or inbound link measurement algorithms, which make the more traditional search engines easier to game, thus including more spam oriented results.

Here is a wrap up of some of the top semantic search engines which we've covered previously, and some updates on their research.

1) HAKIA: It is built around 3 evolving technologies. They are OntoSem (HAKIA's repository of concept relations),QDEX (HAKIA's replacement for inverted index), Semantic Rank Algorithm.

2) KOSMIX: The Search company has takes its categorization concept further by providing users with a dashboard of content, aptly called – " Your guide to the Web".

3) SENSEBOT: The technology powering this engine creates a summary of the top results that are returned for a user query, often negating the need to drill down into the URLs to get the information that one is seeking. Semantic Engines LLC, the company behind the engine provides a variety of products around this technology.

4) COGNITION SEARCH: The Cognition Search NLP Product is a solution companies can use to extract relevant results from their content. The application of this technology could range from better search across the enterprise to fetching more relevant ads

5) SWOOGLE: It is especially for the semantic web. It indexes documents developed on the concepts and standards for semantics (such as the RDF Format).

VI. CONCLUSION AND FUTURE WORK

This paper illustrates the concept of Web Search Engine, various types of Searches and it mainly focuses on Semantic Search and the working of Semantic Search engines. Semantics has a vast usage in these internet days and user need relevant information for their posed Search Queries and Semantics can be applied to various fields such as Knowledge Management and E-Commerce.

In Future, detailed design of Semantic Search Engine will be conveyed and how query clustering will help in grabbing the useful information according to the search queries of the users.

REFERENCES

[1] Ms. Shital C. Patil, Prof. R. R. Keole," The Role of Web Content Mining and Web Usage Mining in Improving Search Result Delivery", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.3, March-2014, pg. 7-14.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] Sanchika Gupta & Dr. Deepak Garg, "Comparison of Semantic and Syntactic Information Retrieval System on the Basis of Precision and Recall" , International Journal of Data Engineering (IJDE), Volume (2) : Issue (3), 2011.

[3] Zhang, J. (2007). "Ontology and the Semantic Web", Proceedings of the North American Symposium on Knowledge Organization. Vol. 1. Available: http://dlist.sir.arizona.edu/1897/.

[4] Web Search Engines: http://en.wikipedia.org/wiki/Web_search_engine.

[5] What is Search Engine Index? : http://www.brickmarketing.com/define-search-engine-index.htm.

[6] Semantic Search : http://google.about.com/od/s/g/semantic_search.htm.

[7] Anusree.ramachandran, R.Sujatha, "Semantic search engine: A survey" , R Sujatha et al, Int. J. Comp. Tech. Appl., Vol 2 (6), 1806-1811.

[8] How search engine works- http://www.monash.com/spidap4.html.

[9] http://evi.com . Accessed 2014 April 16.

[10] Semantic search engines that will change the world of search http://www.searchenginejournal.com/semantic-search-engines/9832/

[11] Web crawlers. Googlebot.On 11.14.09, In Reviews, Tutorials, By Adriano, http://www.milkaddict.com/web-crawlers-googlebot/
[12] How Internet Search Engines Work by Curt Franklin, http://computer.howstuffworks.com/internet/basics/search-engine1.htm
[13] Towards Automated Categorization and Abstracting of Web Sites ,Giuseppe Attardi, Automated Categorization and Abstracting of Web sites, http://www.di.unipi.it/~gulli/categorization/automatedCategorization.html
[14] Search Engine Chronicle:Chris Barton, http://www.searchenginechronicle.com/crawling-vs-indexing-vs-ranking/#sthash.9IxN36ue.dpuf
[15] Market Share Statistics for Internet Technologies: http://www.netmarketshare.com/

o0o

