Trends in Web-Based Search Engine

Kuyoro Shade O, Okolie Samuel O, Kanu Richmond U and Awodele Oludele
Babcock University, Nigeria

ABSTRACT

The use of web search engines is an essential part of the ordinary life. It is difficult to underestimate the tremendous role they have for internet users. Over the years several useful web-based search engines such as Lycos, Excite, AltaVista, Google, Yahoo, Bing and the likes emerge. This work gives an insight into the trend of web-based search engine, diverse ways in which it works, and its future.

Keywords: *Search engines, information retrieval, World Wide Web, Lycos and AltaVista*

1. INTRODUCTION

The volume of information available online today is mind-boggling. Internet users have become so dependent on search engines such as Google, Yahoo and the likes in surfing the net that it is difficult to believe that few years back these do not even exist. Search engine is designed to search for information on the World Wide Web and FTP servers. The search results are generally presented in a list of results called hits. The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Search engines operate algorithmically, i.e. a mixture of algorithmic and human input.

Today, the World Wide Web data is enormous and is rapidly changing; it cannot be confined in the rigid structure of the library. The format of the information is extremely varied, and the individual bits of data -coming from blogs, articles, web services of all kinds, picture galleries, and so on form an infinitely complex virtual organism. In this environment, finding information easily necessitates something more than the traditional structures of data organization or classification.

Web-based search engines allow millions of people from all over the world to search for important information about anything ranging from articles, news, sport, history, products, to multimedia and many more. Search engines are keys to any successful internet promoting strategy. Web-based search engine has been of great importance to everyone that has access to the internet service. It has many impacts on people in their different levels of lifestyle and social responsibility. Possible benefactors and beneficiaries include companies that sell products and services, users of the web. Most Web search engines are commercial ventures supported by advertising revenue and, as a result, some employ the practice of allowing advertisers to pay money to have their listings ranked higher in search results. Those search engines which do not accept money for their search engine results make money by running search related ads alongside the regular search engine results.

The explosive growth of the World Wide Web has proven to be a double-edged sword. While an immense amount of material is now easily accessible on the Web, locating specific information remains a difficult task. This work gives insight into the trends of web-based search engine, how it works and its future. The remaining section are arranged as follows: Section 2.0 gives a description of the historical background of web-based search engines, Section 3.0 presents how search engines works, Section 4.0 discusses the present state of web-based search engines, Section 5.0 highlights the features of web-based search engines and gives list of typical web-based search engines in use today, Section 6.0 describes the future expectations of web-based search engines and Section 7.0 gives the conclusion.

2. HISTORICAL BACKGROUND OF WEB-BASED SEARCH ENGINE

During the early development of the web, there was a list of web servers edited by Tim Berners-Lee and hosted on the CERN web server. The first tool used for searching on the Internet was called Archie. The name stands for "archive" omitting the "v". [2] It was created in 1990 by Alan Emtage, Bill Heelan and J. Peter Deutsch, Computer Science students at McGill University in Montreal. The program downloaded the directory listings of all the files located on public anonymous FTP (File Transfer Protocol) sites, creating a searchable database of file names; however, Archie did not index the contents of these sites since the amount of data was so limited it could be readily searched manually [1].

The rise of Gopher (created in 1991 by Mark McCahill at the University of Minnesota) led to two new search programs, Veronica and Jug head. Like Archie, they searched the file names and titles stored in Gopher index systems. Veronica (Very Easy Rodent-Oriented Net-wide Index to Computerized Archives) provided a keyword search of most Gopher menu titles in the entire Gopher listings. Jug head (Jonzy's Universal Gopher Hierarchy Excavation and

Display) was a tool for obtaining menu information from specific Gopher servers. While the name of the search engine "Archie" was not a reference to the Archie comic book series, "Veronica" and "Jug head" are characters in the series, thus referencing their predecessor.

In the summer of 1993, no search engine existed yet for the web, though numerous specialized catalogues were maintained by hand. Oscar Nierstrasz at the University of Geneva wrote a series of Perl scripts that would periodically mirror these pages and rewrite them into a standard format which formed the basis for W3Catalog, the web's first primitive search engine, released on September 2, 1993.[3]

In June 1993, Matthew Gray, then at MIT, produced what was probably the first web robot, the Perl-based World Wide Web Wanderer, and used it to generate an index called 'Wandex'. The purpose of the Wanderer was to measure the size of the World Wide Web, which it did until late 1995. The web's second search engine Aliweb appeared in November 1993. Aliweb did not use a web robot, but instead depended on being notified by website administrators of the existence at each site of an index file in a particular format.

Jump-Station (released in December 1993) used a web robot to find web pages and to build its index, and used a web form as the interface to its query program. It was thus the first WWW resource-discovery tool to combine the three essential features of a web search engine (crawling, indexing, and searching). Because of the limited resources available on the platform on which it ran, its indexing and hence searching were limited to the titles and headings found in the web pages the crawler encountered.[4]

One of the first "full text" crawler-based search engines was WebCrawler, which came out in 1994. Unlike its predecessors, it let users search for any word in any webpage, which has become the standard for all major search engines since. It was also the first one to be widely known by the public. Also in 1994, Lycos (which started at Carnegie Mellon University) was launched and became a major commercial endeavor. Soon after, many search engines appeared and vied for popularity. These included Magellan (search engine), Excite, Infoseek, Inktomi, Northern Light, and AltaVista. Yahoo was among the most popular ways for people to find web pages of interest, but its search function operated on its web directory, rather than full-text copies of web pages. Information seekers could also browse the directory instead of doing a keyword-based search.

In 1996, Netscape focused on giving a single search engine an exclusive deal to be the featured search engine on Netscape's web browser. There was so much interest that instead a deal was struck with Netscape by five of the major search engines, where for $5 million per year each search engine would be in rotation on the Netscape search engine page. The five engines were Yahoo!, Magellan, Lycos, Info seek, and Excite.[5]

Search engines were also known as some of the brightest stars in the Internet investing frenzy that occurred in the late 1990s. Several companies entered the market spectacularly, receiving record gains during their initial public offerings. Some have taken down their public search engine, and are marketing enterprise-only editions, such as Northern Light. Many search engine companies were caught up in the dot-com bubble, a speculation-driven market boom that peaked in 1999 and ended in 2001.

Around 2000, Google's search engine rose to prominence. The company achieved better results for many searches with an innovation called Page Rank. This iterative algorithm ranks web pages based on the number and Page Rank of other web sites and pages that link there, on the premise that good or desirable pages are linked to more than others. Google also maintained a minimalist interface to its search engine. In contrast, many of its competitors embedded a search engine in a web portal.

By 2000, Yahoo was providing search services based on Inktomi's search engine. Yahoo acquired Inktomi in 2002 and Overture (which owned Allthe Web and AltaVista) in 2003. Yahoo switched to Google's search engine until 2004, when it launched its own search engine based on the combined technologies of its acquisitions.

Microsoft launched MSN Search in the fall of 1998 using search results from Inktomi. In early 1999 the site began to display listings from Look smart blended with results from Inktomi except for a short time in 1999 when results from AltaVista were used instead. In 2004, Microsoft began a transition to its own search technology, powered by its own web crawler (called MSNBOT). Microsoft's rebranded search engine, Bing, was launched on June 1, 2009. On July 29, 2009, Yahoo and Microsoft finalized a deal in which Yahoo Search would be powered by Microsoft Bing technology.[1]

Web search engines faced a number of difficult problems in maintaining or enhancing the quality of their performance. These problems are either unique to this domain, or novel variants of problems that have been studied in the literature. The search engine literature is founded in the area of information retrieval. The vast amount of data available online has initiated extensive research on algorithms and the architecture of search engines.

Generally, search engines' common goal is to make people of different background have easy access to information, which is far from their normal reach. Through the years from the days of Archie till this present time of

MSN, Google, Yahoo and others, search engine have improve to a better standard as the technology improves globally.

3. HOW WEB-BASED SEARCH ENGINES WORK

Search engines work by storing information about many web pages, which they retrieve from the html itself. These pages are retrieved by a Web crawlers also known spider -an automated Web browser which follows every link on the site (Figure 1). Data about web pages are stored in an index database. The purpose of an index is to locate information as quickly as possible. Some search engines, such as Google, store all or part of the source page- cache as well as information about the web pages, whereas others, such as AltaVista, store every word of every page they find. The cached page always holds the actual search text since it is indexed, so it can be very useful when the content of the current page has been updated and the search terms are no longer in it. Increased search relevance makes the cached pages very useful, even beyond the fact that they may contain data that may no longer be available elsewhere [1].

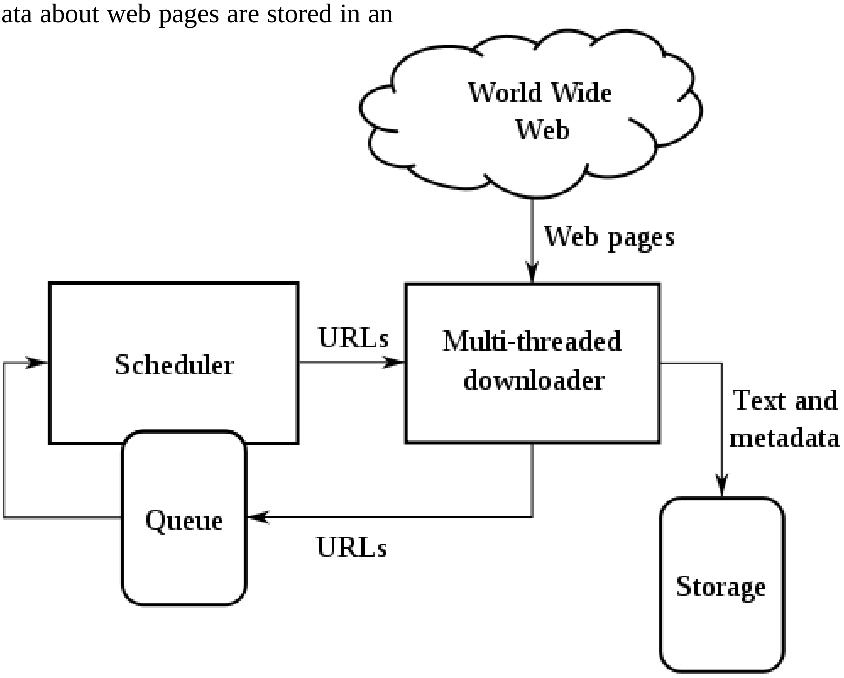

Fig 1: High-level Architecture of a standard Web Crawler[1] Source: dnet - PhD. Thesis of Carlos Castillo

When user enters query into a search engine using keywords, the engine examines its index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. The index is built from the information stored with the data and the method by which the information is indexed. There are currently no known public search engines that allow documents to be searched by date. Most search engines support the use of the Boolean operators AND, OR and NOT to further specify the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The search engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search which allows users to define the distance between keywords. There is also concept-based searching where the research involves using statistical analysis on pages containing the words or phrases searched for. Also, natural language queries allow the user to type a question in human-readable form.

The usefulness of a search engine depends on the relevance of the result sets it returned. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another [6]. The methods also change over time as Internet usage changes and new techniques evolve. There are two main types of search engine that have evolved: A system of predefined and hierarchically ordered keywords that humans have programmed extensively [8], and a system that generates an "inverted index" by analyzing texts it locates. This second form relies much more heavily on the computer itself to do the bulk of the work.

Goggling information on the World Wide Web is such a common activity today that it is hard to believe that few years back this word does not exist. Search engines are now an integral part of human lifestyle, but this has not always the case. Few years back, systems for finding information were driven by data organization and classification performed by humans. Such systems are not entirely obsolete --libraries still keep their books ordered by categories, author names, etc. Yahoo itself started as a manually maintained directory of web sites, organized into categories. Some of the problems identified with every search engine are Content quality and Duplicate host:

3.1 Content Quality

There are many troubling issues concerned with the quality of the content on web. The web is full of noisy, low-quality, unreliable, and indeed contradictory content. A reasonable approach to information for relatively high-quality content would be to assume that every document in a collection is authoritative and accurate, design technique for this context, and then tweak the techniques incorporate the possibility of low-quality content. The democratic nature of content creation on the web leads to a corpus that is fundamentally noisy and of poor quality, and useful information emerges only in a statistical sense. In designing a high quality search engine, one has to start with the assumption that a typical document cannot be “trusted” in isolation; rather, it is the synthesis of a large number of low-quality documents that provides the best set of results. It would be very helpful for web search engines to be able to identify the quality of web pages independent of given user request. There have been link-based approaches, for instance page ranking that helps to estimate the quality of web pages. [6][9]

3.2 Duplicate Hosts

Web search engines try to avoid crawling and indexing duplicate and near duplicate pages, as they do not add new information to the search results and clutter up the results. The problem of identifying duplicates within set of crawled pages is well studied. However, if a search engine can avoid crawling, the duplicate content in the first place, the gain is even larger. Predicting whether a page will end up being a duplicate of an already-crawled page is chancy work, but the problem becomes more tractable if we limit it to finding duplicate hosts, that is, two hostnames that serve the same content. [6][9] One of the ways that duplicate hosts can arise is via artifact of the Domain Name System (DNS) where two host names can resolve to the same physical machine.

4. WEB-BASED SEARCH ENGINES: THE PRESENT

Over the years we have seen several web-based search engines, diverse ways in which they work, and how they have are created. Google, Yahoo, and Bing and the likes are in existence now; which handles the queries after processing the keywords, thus called keyword-based search engines. They search information given on the web page. Recently, some research groups start delivering results from their semantics-based search engines; however most of them are still in their initial stages.

The World Wide Web Worm (WWWW) [13] was one of the first web-based search engines. It was subsequently followed by several other academic search engines, many of which are now public companies. Compared to the growth of the Web and the importance of search engines, there are few documents about recent search engines. However, there has been a fair amount of work on specific features of search engines, which can get results by post-processing the results of existing commercial search engines, or produce small scale "individualized" search engines. There has also been lots of research on information retrieval systems, especially on well controlled collections.

Creating a search engine which scales even to today’s web presents many challenges. Fast crawling technology is needed to gather the web documents and keep them up to date. Storage space must be used efficiently to store indices and, optionally, the documents themselves. The indexing system must process hundreds of gigabytes of data efficiently. Queries must be handled quickly, at a rate of hundreds to thousands per second [18].

These tasks are becoming increasingly difficult as the Web grows. However, hardware performance and cost have improved dramatically to partially offset the difficulty. There are, however, several notable exceptions to this progress such as disk seek time and operating system robustness.

5. FEATURES OF WEB-BASED SEARCH ENGINES

The following features are the basic ways of evaluating Web-based search engines:

5.1 Web Indexes

Whenever a Web search request is issued, it is the web index generated by Web robots or spiders, not the web pages themselves, that is used in retrieving information. Therefore, the composition of Web indexes affects the performance of a Web search engine. There are three components regarding the makeup of a Web index, namely, coverage, update frequency and the portions of Web pages indexed (e.g., titles plus the first several lines, or the entire Web page)[14]. The magnitude of these three components depends largely on the power and sophistication of the hardware and software that make the Web index or database. On the other hand, larger coverage, frequent updates and fulltext indexing do not necessarily mean better Web search engines in other measurements.

5.2 Search Capability

A competent Web search engine must include the fundamental search facilities that Internet users are familiar with, which include Phrase searching, truncation, and limiting facilities (e.g., limit by field). Because the searching capabilities of a Web search engine ultimately determine its performance, absence of these basic functions will severely handicap the search tool.

5.3 Retrieval Performance

Retrieval performance is traditionally evaluated on three parameters: precision, recall and response time. While these three variables can all be quantitatively measured, extra caution should be exercised when one judges the relevance of retrieved items and estimates the total number of documents relevant to a specific topic in the Web system.

5.4 Output Option

This evaluation component is examined from two perspectives. One is the number of output options a Web search engine offers; the other deals with the actual content of the output. Sometimes, one search engine may appear quite impressive in one aspect, but in reality it cannot satisfy its users because of its weakness in the other facet of output evaluation criterion. The output content, to a certain degree, is decided by the way a search engine is constructed.[14]

5.5 User Effort

User effort refers to the documentation and interface.Well-prepared documentation and a user-friendly interface play a notable role in users' selection of Web search engines. The attractiveness of each Web search engine is expressed, to its users, mainly in its documentation and interface. In other words, users will not use a search engine unless they are comfortable with its interface, and able to read and comprehend its documentation when consulted.

6. TYPICAL WEB-BASED SEARCH ENGINES

6.1 Google

Google is designed to be a scalable search engine especially with extremely large data sets. It makes efficient use of storage space to store the index. Its data structures are optimized for fast and efficient access. The primary goal is to provide high quality search results over a rapidly growing World Wide Web. Google employs a number of techniques to improve search quality including page rank, anchor text, and proximity information. Furthermore, Google is a complete architecture for gathering web pages, indexing them, and performing search queries over them. In designing Google, the rate of growth of the Web and technological changes were put into consideration. [12].

6.2 Hakia

Hakia is a general purpose search engine that search structured text like Wikipedia. Hakia calls itself a “meaning-based search engine”. They try to provide search results based on meaning match, rather than by the popularity of search terms. The presented news, Blogs, Credible, and galleries are processed by hakia's proprietary core semantic technology called QDEXing. It can process any kind of digital artifact by its Semantic Rank technology using third party API feeds. A single query by the user brings results from any repository including Web, News, Blogs, Video, Images, Hakia Galleries and also from Credible Sources. For short queries the site displays results in categories, instead of a standard list as shown in current search engines. For longer queries, Hakia highlights relevant phrases or sentences. The results are somehow relevant and reliable but Hakia does not reveal it’s inside technology. Hakia take the searched query and find the results in many categories for example from galleries, videos etc. so it took more time than the usual search engines in the retrieval of results.[21][25]

6.3 Sense Bot

Sense Bot represents a new type of search engine that prepares a text summary in response to the user's search query. Sense Bot extracts the most relevant results using Semantic Web technologies from the Web. It then summarizes the results together for the user as per topic. It uses text mining algorithms to parse (human readable) Web pages which lead to identification of key semantic concepts. The coherent summary is then performed from multi documents that are retrieved. This summary itself becomes the main result of the search. Although the search results are still not relevant, because the summarized result may divert the results from actual demands of the user; the sources from which the results are coming are usually the news agencies so reliability is also somehow missing.[25]

6.4 Power Set

Power set does not search simply on keywords alone, but also try to understand the semantic meaning behind the search phrase as a whole. Power set's first product is a search and discovery experience for Wikipedia. It attempts to use natural language processing to understand the nature of the question and return pages containing the answer. It gives more accurate results, and aggregates information from across multiple articles. The results returned by Power set are most reliable and relevant than all the other semantic search engines however its scope is only limited to articles of Wikipedia.[23][25]

6.5 Deep Dyve

Deep Dyve is a powerful, professional research engine that grants users access to the expert content from the Deep Web - the part of the internet that is not indexed by traditional search engines. It indexes every word in a document, but also computes the factorial combination of words and phrases in the document and uses some industrial strength statistical techniques to assess the “informational impact” of these combinations. The presentation of search results is very complex. It presents the users with many advanced options for refining, sorting or saving the search. The search results are however relatively easy to navigate. The results presented are only for the paid customers, they are not available for the general public.[24][25]

7. WEB-BASED SEARCH ENGINE: THE FUTURE

A large-scale web search engine is a complex system, upon which further improvements must be made. Some simple improvements include query caching, smart disk allocation, and subindices.

Another area which requires much research is updates. There is need for smart algorithms to decide what old web pages should be re-crawled and what new ones should be crawled. One promising area of research is in using proxy caches to build search databases, since they are demand driven. Simple features supported by commercial search engines like boolean operators, negation, and stemming should be added. Other features such as relevance feedback and clustering etc should be explored. Web search engine is a very rich environment for research ideas. There are far too many improvements that can enhance it in the future. The following are few expectations of better improved web-based search engines: high quality search, scalable architecture and a first-class research tool.

8. CONCLUSION

The search engine is a very important tool for people to obtain information on Internet While the information retrieval technology has been studied for decades, the searching results still cannot satisfy people’s needs, both in quantity and quality. Most Web search engines only make use of the text on a Web page, ignoring other rich source of information, such as the hyperlinks among pages or the Web usage information. With the rapid development of Internet and the continuous increase of the information, introducing the Web mining technology into the traditional search engines is one of the most important challenges in the fields of information retrieval and artificial intelligence.

REFERENCES

[1] Wikipedia, 2011, http://en.wikipedia.org/wiki/Web_search_engine Retrieved May 1st 2011.

[2] "Internet History - Search Engines" (from Search Engine Watch), Universities Leiden, Netherlands, September 2001, web: Leiden U-Archie.

[3] Oscar Nierstrasz (2 September 1993). "Searchable Catalog of WWW Resources (experimental)"

[4] "Archive of NCSA what's new in December 1993 page". Web.archive.org. 2001-06-20. http://web.archive.org/web/20010620073530/http://archive.ncsa.uiuc.edu/SDG/Software/Mosaic/Docs/old-whats-new/whats-new-1293.html. Retrieved 2012-05-14.

[5] Browser Deals Push Netscape Stock Up 7.8%. Los Angeles Times. 1 April 1996

[6] Henzinger, M. R., Motwani, R., and Silverstein, C. Challenges in web search engines. SIGIR Forum 36, 2 (2002), 11–22

[7] Aaron M W (2005), Search Engine Optimization, SEO Books. (pp. 1-3).

[8] B. McBride, Jena: A semantic web toolkit. IEEE Internet Computing, Vol. 6, No. 6, pp. 55-59, 2002.

[9] Eric Prudhommeaux, Presentation of W3C and Semantic Web, 2001, http://www.w3.org/2001/Talks/0710-ep-grid

[10] "Google Technology". Google.com. http://www.google.com/technology/. Retrieved 2011-05-27.

[11] Jaimie S, Cristian D (2007), Professional Search Engine Optimization with PHP - A Developer’s Guide to SEO, Wiley Publishing, Inc. (pp. 1-4).

[12] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The Page Rank Citation Ranking: Bringing Order to the Web. Manuscript in progress. http://google.stanford.edu/~backrub/pageranksub.ps [Page 98]

[13] Oliver A. McBryan. GENVL and WWWW: Tools for Taming the Web. In Proceedings of the First International Conference on the World Wide Web. CERN, Geneva (Switzerland), May 25-26-27 1994.

http://www.cs.colorado.edu/home/mcbryan/mypapers/www94.ps

[14] Notes, Greg R. (July/August 1995a). Searching the World-Wide Web: Lycos, WebCrawler and More. Online, 19(4), 48-53.

[15] Monika R. Henziger (2002), Challenges in Web Search Engines page 4&5, September 2002

[16] Oliver A. Mc Bryan. GENVL and WWWW: Tools for Taming the Web. First International Conference on the World Wide Web. CERN, Geneva (Switzerland), May 25-26-27 1994. http://www.cs.colorado.edu/home/mcbryan/mypapers/www94.ps

[17] Seo Wizard (2011)http://www.seowizard.com/The-Importance-of-Search-Engine-Optimization-SEO.htmlaccessed May 1st 2011.

[18] Sergey Brin and Lawrence Page, “The Anatomy of a Large-Scale Hyper textual Web Search Engine”, Computer Science Department, Stanford University, Stanford, CA 94305, USA sergey@cs.stanford.edu and page@cs.stanford.edu

[19] Timothy B, Martin E B (2008), PHP and My SQL Create-Modify-Reuse, Wiley Publishing, Inc. (pp. 87-91).

[20] [Weiss 96] Ron Weiss, Bienvenido Velez, Mark A. Sheldon, Chanathip Manprempre, Peter Szilagyi, Andrzej Duda, and David K. Gifford. HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering. Proceedings of the 7th ACM Conference on Hypertext. New York, 1996.

[21] http://company.hakia.com/termsofservicemeet.html retrieved May 21,2012

[22] http://www.sensebot.net/ Retrieved May21,2012

[23] "Powerset Founders". Archived from the original on 2007-10-27. http://web.archive.org/web/20071027074129/http://www.powerset.com/team. Retrieved 2012-05-21.

[24] http://www.deepdyve.com/retrieved May 21, 2012

[25] Ritu Khatri, Kanwalvir Singh Dhindsa, Vishal Khatri Investigation and Analysis of New Approach of Intelligent Semantic Web Search Engines International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-1, Issue-1, April 2012