International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166

Volume 3 Issue 2 May 2012

A COMPARATIVE ANALYSIS OF VARIOUS WEB SEARCH ENGINES

Pooja Choudhary

Dept.of Computer Science & Engg., UIET, Kurukshetra University Kurukshetra,India

Abstract

The main source of an information system is search engine .The layer occupied by relational database management system, in traditional business, environment is supplemented or replaced with a search engine or the technology used to build search engine. The SQL (Structured Query Language) which was used for query for information are now replaced by keyword or searches for structured, semi-structured, or unstructured data. Information is maintained in data tier in a typical multi-tier or entire where it can be stored and retrieved from a database or file system. The data tier is queried by the logic or business tier when information is needed using a data retrieval language like SQL. In a search –oriented architecture the data tier may be replaced or placed behind another tier which contains a search engine and search engine index which is queried instead of the database management system.

Keywords: Search engine, World Wide Web, SQL.

1. Introduction: Search engines are programs that search web pages or documents for a particular keywords or where the keywords were found. A Search engine is really a collection of programs, however, the term is often used to specifically describe systems like Google, Bing and Yahoo! Search that enable users to search for documents on the World Wide Web. The components and tasks of web search engines, Crawling or spidering is an automated process to gather the data with web spiders. They can be pictured as little spiders and are also known as crawlers, robots, software agents, web agents, wanderers, walkers, or knowbots [Clay & Esparza, 2009]. They are named after special software robots, this type of search service is

called “spider-based” or “crawler-based” search engine. Spiders process the web page and give us information. The web pages are found by them by URL which is given by a web page holder to notify their web page, or through hypertext links embedded in most web pages [Sherman & Price, 2001]. In the latter case, spiders start by crawling a few web pages and follow the links on those pages. After fetching the pages they point to, they follow the links that are on the last pages. The same process will be continued until they have indexed a certain part of the web that includes pages they store across many machines, what leads to the next task. Indexing is the second part of search engines. It is the process of “taking the raw data and categorizing it, removing duplicate information, and generally organizing it all into an accessible structure” [Clay & Esparza, 2009].The stored full-text indexes of the crawled web pages are organized in a database, typically in an inverted index data structure [Sherman & Price, 2001]. It is ultimate for keyword based queries, so that documents that include the typed keywords can be quickly retrieved [17].

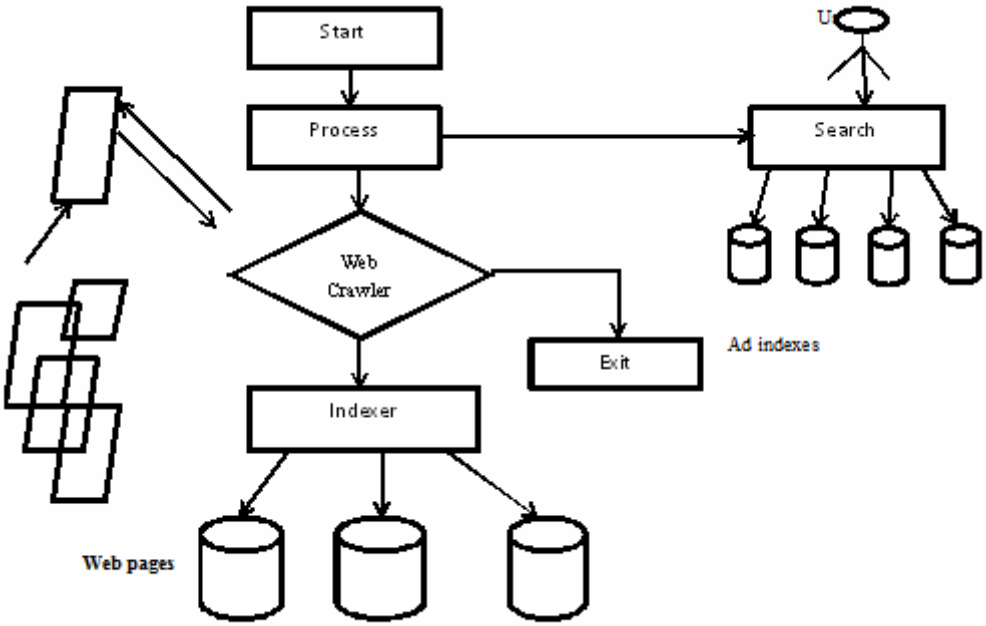

Figure1: Components of Search engine

Webmasters have taken many advantages of the web, especially for business commitments. A lot of power will be put into search engine optimization (SEO) or maximizing search engine visibility, online marketing strategies [Clay & Esparza, 2009][17].

2. Google Architecture

GOOGLE was the first web search engine known to apply link analysis on a large scale, although all web search engines currently make use of it. Page rank is a score assigned to each page by them, Which can be interpreted as the fraction of time that a random web surfer will spend on that webpage when following the outlinks from each page on the web. Another interpretation is that when a page links of another page, it is effectively casting a vote of confidence. Page rank calculates a page's importance from the votes cast for it. HITS is another technique employing link analysis which scores pages as both hubs and authorities, where a good hub is one that ,links to many good authorities and good authority is one that is linked from many good hubs.

In Google, downloading of web pages is done by several distributed crawlers. There is a URL server that sends lists of URLs to the crawlers. The web pages that are crawler sent to the crawler are the transfer to the store server. The store server then compresses and stores the web pages into a repository. Every web page has an associated ID number called a docID which is assigned whenever a new URL is parsed out of a web page.

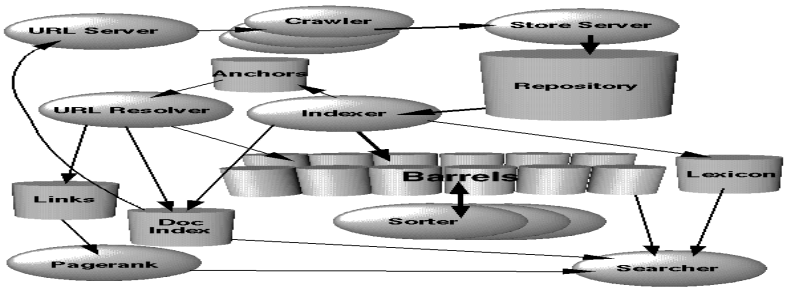

Figure2: Google Architecture [8]

The indexing function is performed by the indexer and the sorter. The indexer performs a number of functions. It reads the repository, uncompresses the documents, and parses them. Each document is converted into a set of word occurrences called hits. The hits record the word, position in document, an approximation of font size, and capitalization. The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index. The indexer performs another important function. It parses out all the links in every web page and stores important information about them in an anchors file. This file contains enough information to determine where each link points from and to, and the text of the link. The URL resolver reads the anchors file and converts relative URLs into absolute URLs and in turn into docIDs. It puts the anchor text into the forward index, associated with the docID that the anchor points to. It also generates a database of links which are pairs of docIDs. The links database is used to compute Page Ranks for all the documents. The sorter takes the barrels, which are sorted by docIDand routes them by wordID to generate the inverted index. This is done in place so that little temporary space is needed for this operation. The sorter also produces a list of wordIDs and offsets into the inverted index. A program called DumpLexicon takes this list together with the lexicon produced by the indexer and generates a new lexicon to be used by the searcher. The searcher is run by a web server and uses the lexicon built by DumpLexicon together with the inverted index and the Page Ranks to answer queries[8].

3. Characteristics of Search Engines

Boolean Operations: AND means the search results must have both terms – often it is typed in UPPER CASE, but not always. AND provides some particular information about a topic and thereby decreasing the number of sites regarding that topic. Example: communication AND transport will look for web sites about both communication and transport. OR means the search results can have just one of the terms. OR provide wide results, increases number of websites and hence broadens your search Example: communication OR transport will look for web sites that mention either communication or transport. NOT reduces number of results and the probability of showing results regarding the second term are highly reduced. Example:

communication NOT transport will open web sites about communication excludes results including transport.

Non Boolean Searching: If the search terms are directly written without any space and works like OR, then it will show results for all topics specified. Example: communication transport will look for web sites about communication that also mention transport [15].

4. Comparative Analysis of various Search Engines:

Since web designers are interested, major search engines as extremely relevant, because they want their site to be visited most and therefore to be placed in a place where they can get a lot of traffic. Thus, those engines are most appropriate for SEO strategies. Google is recognized as the largest search engine worldwide. A search survey by ComScore, a leader in measuring digital world, proves that this statement is acknowledged: In 2009, Google dominated 66.8% of worldwide search with 87,809 searches, followed by Yahoo! with 9,444 searches, the Chinese search engine Baidu with 8,534 searches, and Bing that ranked fourth with 4,094 searches [comScore, 2010].

Studies of Hitslink by Net Application show in its market share rankings of search engines that the positions of the last two years still remain in August 2010. Google ranks first again with 84.73% market share, and outperforms Yahoo! (6.35%), Baidu (3.31%), and Bing (3.30%), while the other engines only capture a total of 1.32%. Baidu recently outpaced Bing, while the other engines only capture a total of 1.32%. Baidu recently outpaced Bing, namely from July to August 2010 with 1% [Net Applications, 2010] [17].

Top 10 Search Properties by Searches Conducted December 2009 vs. December 2008 Total Worldwide, Age 15+ - Home & Work Locations Source: comScore qSearch			
	Searches (MM)		
	Dec-2008	**Dec-2009**	**Percent Change**
Worldwide	*89,708*	*131,354*	*46%*
Google Sites	55,638	87,809	58%
Yahoo! Sites	8,389	9,444	13%
Baidu.com Inc.	7,963	8,534	7%
Microsoft Sites	2,403	4,094	70%
eBay	1,327	2,102	58%
NHN Corporation	1,892	2,069	9%
Yandex	992	1,892	91%
Facebook.com	1,023	1,572	54%
Ask Network	1,053	1,507	43%
Alibaba.com Corporation	1,118	1,102	-1%

Figure4.1: Top properties of Searches Conducted [17]

Webdevelopersnotes.com with a total of 13,304 participants is currently making an online survey in September 2010, voting on what search engine they think is the best in the world. The majority of the voters estimated Google as the best search engine, while only one fourth of the total voters chose either Yahoo, Bing, AOL, or Ask. The search engine Baidu, which is the most used search engine in China, has taken the third rank, but it is only available in the Chinese version and not widespread globally at the moment.

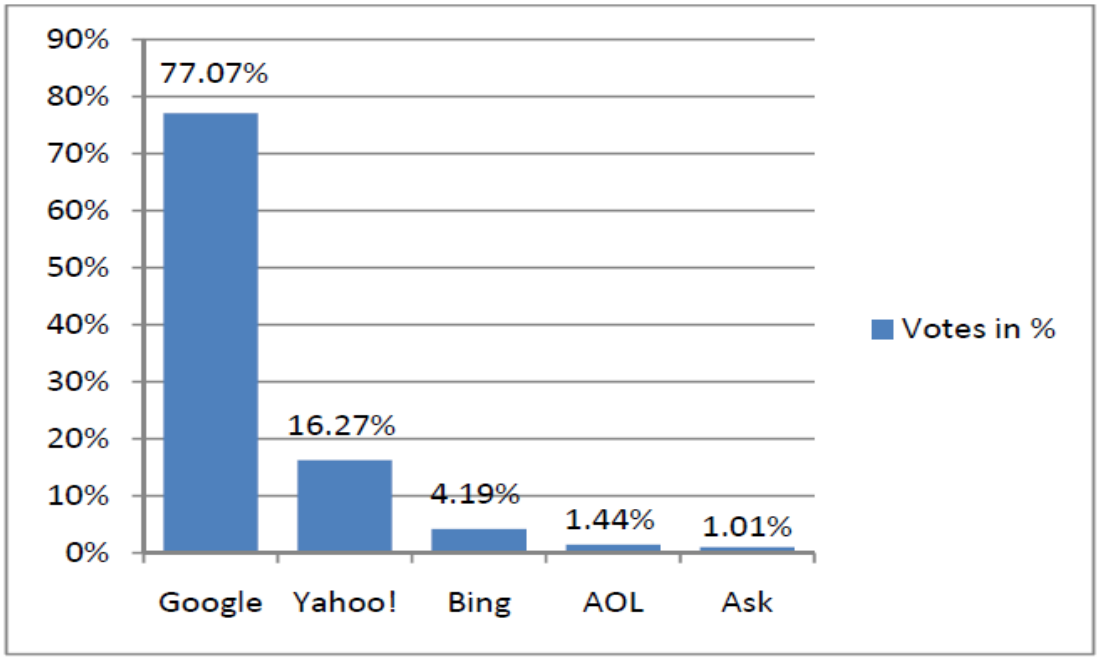

Figure 4.2 Survey for the best search engine [17]

The search engine Baidu, which is the most used search engine in China, has taken the third rank, but it is only available in the Chinese version and not well-known generally at the moment. Consequently, in this paper, it should not be examined more in detail [17].

In this section following aspects of search engines are analyzed: The database size, the actuality, the capabilities, and the technology of the results of search engines.

(a)The database size: The web has such an enormous growth that it cannot be counted. The exact number of web pages that is in a search engines is not known either. However, there are some estimations of the size of data in each search engine**.** Lewandowski [2005a] says the only way to find out the size of the web is to evaluate the size that is based on a representative sample [17]. He indicates that Google has indexed around 8 billion documents, whereas estimation of Yahoo's index shows 5-7 billion, and formerly MSN search 4-5 billion documents. Those numbers, however, were from 2005**.**Google, Yahoo, and Bing work on reaching the Deep Web. Google involves the most important documents of the Hidden Web for patent data manually, but it has also developed a technology that allows an automatic approach of resources [Lewandowski, 2005a; Madhavan, Ko, Kot, Ganapathy, Rasmussen, & Halevy, 2008].Yahoo in contrast, has its Content Acquisition Program in which it includes documents of the Deep Web through partnerships as content providers [Olsen, 2004].There is no clear information about Bing, but it has also been exploring the Invisible Web, for example in Microsoft Research, a testbed for information extraction from Deep Web was proposed [Yamada, Craswell, Nakatoh, &Hirokawa, 2004].

(b)Actuality: In a research, Lewandowsky, Whalig and Meyer-Bautor [2005] tested the frequency in which the indices are updated by Google, Yahoo and MSN. During forty-two days, they observed four different groups with nine or ten websites, updated every day, to find out whether these search engines are able to index current contents on a daily basis. Googlebot, Google's web crawling robot updates many sites daily and is the fastest concerning index quality. The amount of time for re-crawling usually depends on the link popularity and on the frequency how often the web page changes [Google Guide, 2007]. MSN on the other hand updates the index with MSNbot frequently, while Yahoo seems to update with its crawler Yahoo Slurp in a chaotic way (Lewandowski et al., 2005).

(c) Capabilities: **(I)** Google supports the Boolean operators AND and OR, as well as the removing and including functions "-" and "+". Furthermore, it also features phrase search with quotation marks, wildcards, reducing and website specification, as well as Meta words search [Google, 2010c].

(II) Yahoo also gives the possibility to use the Boolean operators and the function to require orexclude words. It gives the permission of the use of quotation marks, wildcards, stemming to expand search results, and Meta words. In addition, time can be saved by Yahoo! Shortcuts, symbols and keywords for specialized answers that appear directly on the results page, such as calculator or gas prices [Yahoo, 2010a].

(III) As for Bing, like the other two search engines, Boolean operators can be applied, as well as Meta words, and placing words into quotation marks. It does not mention wildcards, but rather stemming. Similar to Yahoo Shortcuts, it has instant answers to get the information quickly [Microsoft, 2007].

(d) Technology: (I) Google uses the famous and patented query-independent factor called Page Rank that determines the link popularity. This algorithm, named after Larry Page, analyzes the whole link structure of the web and assesses which pages are most important. Google assumes that votes for a page's importance can be assigned to links. The more important Google believes a page is, that means the more votes it has, the higher is its Page Rank, and thus, its listing on the SERP. The link should ideally be from pages that are as relevant as possible. Not only the quantity of hyperlinks that point into a page, called inlinks, is a ranking factor, but also relevant content and the quality of the pages [Langville & Meyer, 2006]. With a query-dependent that is called Hypertext-Matching Analysis, Google evaluates the full content of web pages as well as the content of neighboring web pages to determine if they are relevant to the conducted query [Google Inc., 2010].

(II) As for Yahoo! Search, the algorithm has some similarity to the Google algorithm. It's been claimed that Page rank affects the Yahoo ranking at some point. The ranking consists of the analysis of web page text, keywords in the title, description, source, and associated links. Most importantly, the title must contain major keywords. It can be advantageous to also include them in the description and category. Another part of its algorithm is click-popularity, what means that the number of document clicks from a results page will be counted [Yahoo! Help, 2010].

(III) The Bing ranking is completely automated, its algorithm is complex and never human mediated. Every time it updates the index, it changes the relevance rankings. Bing emphasizes

on keywords, especially in the domain name and URL, and great, original “analyzes the quality and quantity of index able webpage content, the number, relevance, and authoritative quality of websites that link to web pages, and the relevance of the website’s content to keywords” [Bing Webmaster Center, 2010]. In Bing community [DeJarnette, 2009], guidelines for SEO are given for improving site ranking. Bing mainly content that are directed to the desired audience. Additionally, the architecture of the content should be well organized with regard to images to help MSNbot read and crawl the site [17].

International Journal of Computing and Business Research (IJCBR)
ISSN (Online) : 2229-6166

Volume 3 Issue 2 May 2012

Search Engine Comparison Table			
	Google	**Yahoo!**	**Bing**
Database			
Index size	Rank 1 (> 12 bn pages) Includes Deep Web	Rank 2 (> 9 bn pages) Deep Web	Rank 3 (> 0.9 bn pages) Deep Web
Crawler name/ Actuality	Googlebot Fastest, many daily updates	Yahoo Slurp No clear frequency	MSNbot Frequent updates
Capability			
Search Operators Advanced Search	• Boolean operators • - to remove • + to include • “” quotation marks for exact words • Wildcards • Stemming • Meta word search • Advanced Search Form • 46 languages	• Boolean operators • - to remove • + to include • “” quotation marks for exact words • Yahoo shortcuts • Wildcards • Stemming • Advanced Search Form • 32 languages	• Boolean operators • - to remove • “” quotation marks for exact words • Instant answer • Stemming • Advanced Search Form • 42 languages
Technology			
Speed	Shown for every query Google Instant “search-before-you-type”	Not exactly known	Not exactly known
Ranking	PageRank, hypertext matching analysis	Keywords, click-popularity	Automated, emphasis mainly on keywords, inlinks

Figure 4.3: Search Engine Comparison Table [17]

5. Challenges in Search Engine

PRIMO DESIGN CONCEPTS Decoupled achitecture: Following to the principle of decoupled architecture, Primo includes a publishing platform that enables it to harvest data from resources controlled by the library, such as the library catalog, digital repositories, knowledge bases that

describe the library's electronic holdings, and course management systems. This data is typically bibliographic data representing items of any type that the library offers - for example, books, journals, articles, images, music scores, videos, and audio - and may also include textual objects, such as PDF files stored in a local repository. The publishing platform normalizes the harvested data (that is, makes the data conform to one set of rules), converts all the data to one format, and enriches it with information such as book-cover images, abstracts, and tables of contents from third-party sources such as Amazon.com and Syn Solutions : In addition, the publishing platform detects duplicate records and groups all the similar records (for example, multiple editions of the same book) to display them to the user as one entity in the initial search results.3 The outcome of this process is the Primo index, which is optimized for Primo searches. After the initial building of the index, the publishing platform also maintains it; whenever the contents of the source repository change, the publishing platform harvests the data that was added or changed and takes it through the publishing process. Publishing platform notice duplicate records and collects all similar records to display them to user at one place only [1].

(a) Click fraud: It is considered to represent any kind of fraud that exploits pay-per-click markets. Any intentional click on a pay-per-click advertisement is conceived as fraudulent if no intention of a conversion exists (JANSEN, 2006; KITTS et al., 2006)[3]. In other words, the perpetrator is not interested in the products, services, or the content of the advertised Website. A conversion is generally referred to a click on an advertisement that leads to a predefined action. In the view of an advertiser, this positive result can be the visit of a Website, the request of information material, the registration of a new customer, or the conclusion of an e-commerce transaction. Based on this definition of click fraud, a classification of click fraud types is presented according to the motivation and the form of the click fraud conducted. Click fraud motivation can be differentiated into damnification and enrichment. Damnification refers to a perpetrator aiming to harm the company by assaulting the advertising campaign. In contrast, enrichment is click fraud directed towards a personal gain. An example of this case is a partner of the search engine provider causing click fraud in order to increase advertising compensation. Click fraud can be conducted manually by individuals clicking on an ad or automatically by computer programs [10].

Solutions: To prevent ppc click fraud we can use different prevention tools like click defense is acitesaggressive numbers and is primarily aimed at prevention through accurate tracking. Click forensics is a firm that will audit your clicks.

(b) Content quality: The web is full of noisy, low quality, unreliable & unneeded content and quality evaluation is evaluating the quality of different ranking algorithm is a notoriously different problem.

Solution: Commercial search engine have the benefit of large amount of user behavior data they can use to help evaluate ranking. Typical document cannot be trusted in isolation rather than it is the synthesis of a large no. of low quality documents that provides the best set of results.

(c) Web Conventions: Most creators of web pages seem to follow simple rules without anybody imposing these rules on them. For e.g they use the anchor text of a link to provide a description of target page. The main issue here is to identify the various conventions that have evolved organically $ to develop technique for accurately determining when the conventions are being violated.

Solution: The various conventions that have evolved organically should be identified and techniques should be developed for accurately determining when the conventions are being violated.

(d)Duplicate hosts: Web search engine try to avoid crawling & indexing duplicate & near duplicate pages, as they do not add new information to the search results & clutter up the results. That duplicate hosts can arise is via an artifact of the domain name system (DNS) where two hostnames can resolve to the same physical machine.

Solutions: The best way is to send ARP packets to that IP address. All the machines with same IP will revert with the MAC address. If you are using any security solution block those MAC address or if you are using some tools to manage your asset you can clearly find out which users have the duplicate IPs.

(e) Synonym problem: Search with synonyms is a challenging problem for Web search, as it can easily cause intent drifting. Synonym discovery is context sensitive. Although there are quite a few manually built thesauri available to provide high quality synonyms (Fellbaum, 1998), most of these synonyms have the same or nearly the same meaning only in some senses. If we simply replace them in search queries in all occurrences, it is very easy to trigger search intent drifting. Thus, Web search needs to understand different senses encountered in different contexts [14].

Solution: Cross references in deterministic search systems solve the synonym problem by using a controlled vocabulary (such as the Library of Congress Subject Headings) that select one term for each concept and provide cross references from variant terms. For example, a search on "false teeth" leads to a cross reference that says "see Dentures."

(f) Homonym problem: Search systems are created by trained humans (like library catalogers) who apply rules and reasoning to create search platforms that are precise and provide high recall. These systems are an excellent (and yes, more expensive) alternative to stochastic search systems, such as web search engines [16].

Solution: Deterministic search also solves the homonym problem by creating unique labels for concepts with the same name, as in "Jaguar (Animal)" and "Jaguar (Automobile)."This technology is called deterministic search, which is a type of search that uses one-to-one matching between a search query and metadata (including cross references) that function as substitutes for information resources. This is the type of search that is done in an online library catalog.

Deterministic search systems are created by trained humans (like library catalogers) who apply rules and reasoning to create search platforms that are precise and provide high recall. These systems are an excellent (and yes, more expensive) alternative to stochastic search systems, such as web search engines.

6. Advantages of Search Engines:

Simple searching: Web search engines and internet services do not require a significant degree of searching literacy on the part of the user and provide results that are usually good enough without precision searching. Features such as spelling corrections and suggested alternatives to search queries that yield few results relieve the user from the burden of entering highly accurate search terms. Web search interfaces are typically very simple and require no expertise for conducting a keyword search - the kind of search that is most popular on the web. Furthermore, recent improvements in the way in which results are organized and displayed on the screen have introduced 'faceted browsing', a method by which users can easily narrow down their searches to obtain more precise results.

Vast, heterogeneous content: Web search engine and internet explorer like provide more ease to its users in vast coverage and heterogeneity of resources in comparison to the libraries. It may be lacking behind in efficiency guaranteed by the material of library. Today people prefer online shopping.

Online availability of material: People are attracted to the online information and find the needful items within seconds. Internet provides its users facilities like listening audio, zooming in and out of images, focusing on particular paragraph. Although libraries also provides such things to some extent but are not much preferred for finding electronic materials.

Interactive experience: After the comparison of an users experience in both library system and internet services such as the characteristics of Web 2.0 as a platform for cAmazon.com, del.icio.us, Facebook, Flickr, MySpace, and YouTube, it is found that the later proves much better as it serves as a platform which connects people and with information it also provides right to share reviews, have discussions; the characterstics Web 2.0 as a platform for collaboration are only now emerging in such systems.

Conclusion:

These papers briefly describe the theory of search engine. The ALANTIC online search engine laid the foundation stone of web search engines. But in the present day time Google is most advance search engine. In this we have come across the various fields of search engine like its characterstics, advantages, challenges and their solutions etc. Various communities like IR are

working of the solutions of present day challenges faced of web search engine. We analysedgoogle, yahoo, bing and described their various components that how they work. We have made comparison between search engine on the basis of database and actuality.

REFERENCES

[1] TamarSadeh " *User centric solutions for scholarly research in the library*." LIBER QUARTERLY, ISSN 1435-5205 © LIBER 2007. All rights reservedIgitur. Utrecht Publishing & Archiving Services.

[2] Abhishek Das, Ankit Jain "*Indexing the World Wide Web* : *The Journey So Far*."Publisher: IGI Global, Pages: 1-24.

[3] Henzinger et. al. " *Challenges in web search engine*."ACM SIGIR Forum, 2002 -dl.acm.org.

[4] Jansex et. al. " *Determining the user intent of web search engine queries*."WWW '07 Proceedings of the 16th international conference on World Wide Web ACM New York, NY, USA ©2007 table of contents ISBN: 978-1-59593-654-7.

[5] http://www.webopedia.com.

[6] http://www. google.com.

[7] http://docs. google. Com.

[8] SergeryBrin and Lawrence page "*The Anatomy of a Large Scale Hyper textual WebSearch Engine*".WWW7 Proceedings of the seventh international conference on World Wide Web 7Journal Computer Networks and ISDN Systems archive Volume 30 Issue 1-7, April 1, 1998.

[9] Monika R.Henzinger"*Algorithmic in Web Search Engines*" Internet Mathematics Vol.1, No.1:115-126.

[10]Rainer Olbrich ,Carsten D. Schultz "*Search Engine Marketing and Click Fraud*"Business Information Review (2003) Volume: 20, Issue: 4, Pages: 195-202 ISSN: 02663821 DOI: 10.1177/0266382103204005.

[11] Bernard J.Jansen, Amanda Spink "*Searching multimedia federated contentwebcollections*". Emerald Group Publishing Limited Online Information Review, Vol. 30 Issue: 5, pp.485 – 495.

[12] Spink et. al."*Overlap among major web Search engines*".ITNG '06 Proceedings of the Third International Conference on Information Technology: New Generations IEEE Computer Society Washington, DC, USA ©2006ISBN:0-7695-2497-4.

[13] www.cloaking.com.

[14] Xing Wei et al. "*Search with synonyms: Problems & solutions*" Yahoo! Labs 701 First Avenue, Sunny vale, California, USA, 94089.

[15]Jason Tramer http:// itknowledgeexchange.techtarget.com/ita.

[16] Monika Hezinger, Jeffrey Beall "*The solution to weaknesses of web search engine"* Science Publishing Online august 2007.

[17] Prof. Dr. Andreas Meier "*Meta Search Engine Analysis"*University of FribourgFaculty of Economics Social Sciences Information Systems Research Group Meta-Search Engine Analysis Seminar ThesisSeptember 2010.