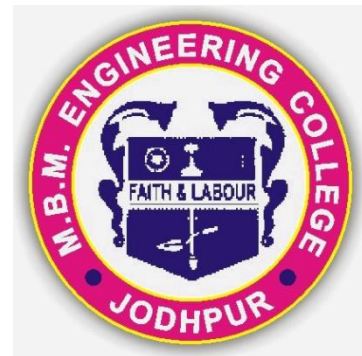# A
Seminar
On

"Search Engine"

Submitted for the partial fulfillment of
B.E. Final Year Information Technology

Submitted By:-
Govind Salvi
Roll No.-2471
E. No.–07/05766

Seminar Guide:-
Asst.Prof. N.L. Sharma
(M.B.M Engg.college Jodhpur)

Department of Computer Science and Engineering
MBM Engineering College, Faculty of Engineering,
Jai Narain Vyas University
Jodhpur (Rajasthan)
Session 2010-11

CERTIFICATE

This is to certify that report entitled "**Search Engine**" submitted to Department of Computer Science and Engineering, MBM Engineering College, JNV University in partial fulfillment of the requirements for the award of the degree of B.E. Final Year IT is the bonafied record of the work done by Govind Salvi under my supervision and guidance.

Signature of the guide

Asst. Prof. N. L. Sharma

Department of Computer Science and Engineering
MBM Engineering College, Faculty of Engineering,
Jai Narain Vyas University

DECLARATION

I, **Govind salvi** hereby declare that this Seminar titled "**Search Engine**" is a record of original work done by me under the supervision and guidance of **Mr. N. L. Sharma,** I further certify that this Seminar work has not formed the basis for the award of the Degree/Diploma/Associate ship/Fellowship or similar work to any candidate of any university and no part of this report is reproduced as it is from any other source without seeking permission.

Signature of the student

Govind Salvi

Roll No.-2471

E. No.–07/05766

Information Technology

Abstract

The **Web search engine** is the searching tool that is basically designed to search the information on the **world wide web** that will allows to access any kind of information you want to access and basically it uses the FTP(file transfer servers) which transfers the particular file when the request made by you and shows the stock of result that matches with your request and list of stock is presented in a linear manner such that high quality resources are on the top position of search engine while others are lower than it.

In order to improve their web engine, many companies occurs in the market as the web portal for their competitors and gain strength as the public search engine but Bing, Google and Yahoo has gain strength among them because they uses the innovation technique of page ranking which uses the algorithm which searches for the page rank of the particular web page by taking the consideration of that particular web page along with the website ranking to which that web page belongs.

CONTENTS

Seminar: Search Engine

Chapter 1:- Introduction

Search Engine, software program that helps users find information stored on a personal computer, or a network of computers, such as the Internet. A user enters search terms, typically by typing a keyword or phrase, and the search engine retrieves a list of World Wide Web (WWW) sites, personal computer files, or documents, either by scanning the content stored on the computers or computer networks being searched or by parsing (analyzing) an index of their stored data.

Search engines are most often used to find pages, files, news, images, and other data on the Web. Some of the most popular Web search engines include Google Inc., Microsoft Network (MSN) Search, and Yahoo! Inc. Each can be accessed from any Web browser, and each can be used for free. These engines operate by building—and regularly updating—an enormous index of Web pages and files. This is done with the help of a Web crawler, or spider, a kind of automated browser that perpetually trolls the Web, retrieving each page it finds. Pages are then indexed according to the words they contain, with special treatment given to words in titles and other headers. When a user inputs a query, the search engine then scans the index and retrieves a list of pages that seem to best fit what the user is looking for. Search engines often return results in fractions of a second.

Generally, when an engine displays a list of results, pages are ranked according to how many other sites link to those pages. The assumption is that the more useful a site is, the more often other sites will send users to it. Google pioneered this technique in the late 1990s with a technology called Page Rank. But this is not the only way of ranking results. Dozens of other criteria are used, and these will vary from engine to engine.

Many times, search results will also include what are called sponsored links, links that are ranked high in the search results or are prominently displayed because third-party companies pay a fee to the search engine. More often than not, sponsored links are labeled as such, but inexperienced Internet users often have trouble distinguishing between sponsored pages and unsponsored results. Sponsored links provide search engines with their primary source of revenue.

Figure 1.1 Yahoo Search Results (www.yahoo.com)

Chapter 2:-How Do Web Search Engines Work

Search engines are the key to finding specific information on the vast expanse of the World Wide Web. Without sophisticated search engines, it would be virtually impossible to locate anything on the Web without knowing a specific URL. But do you know how search engines work? And do you know what makes some search engines more effective than others?

When people use the term search engine in relation to the Web, they are usually referring to the actual search forms that search through databases of HTML documents, initially gathered by a robot.
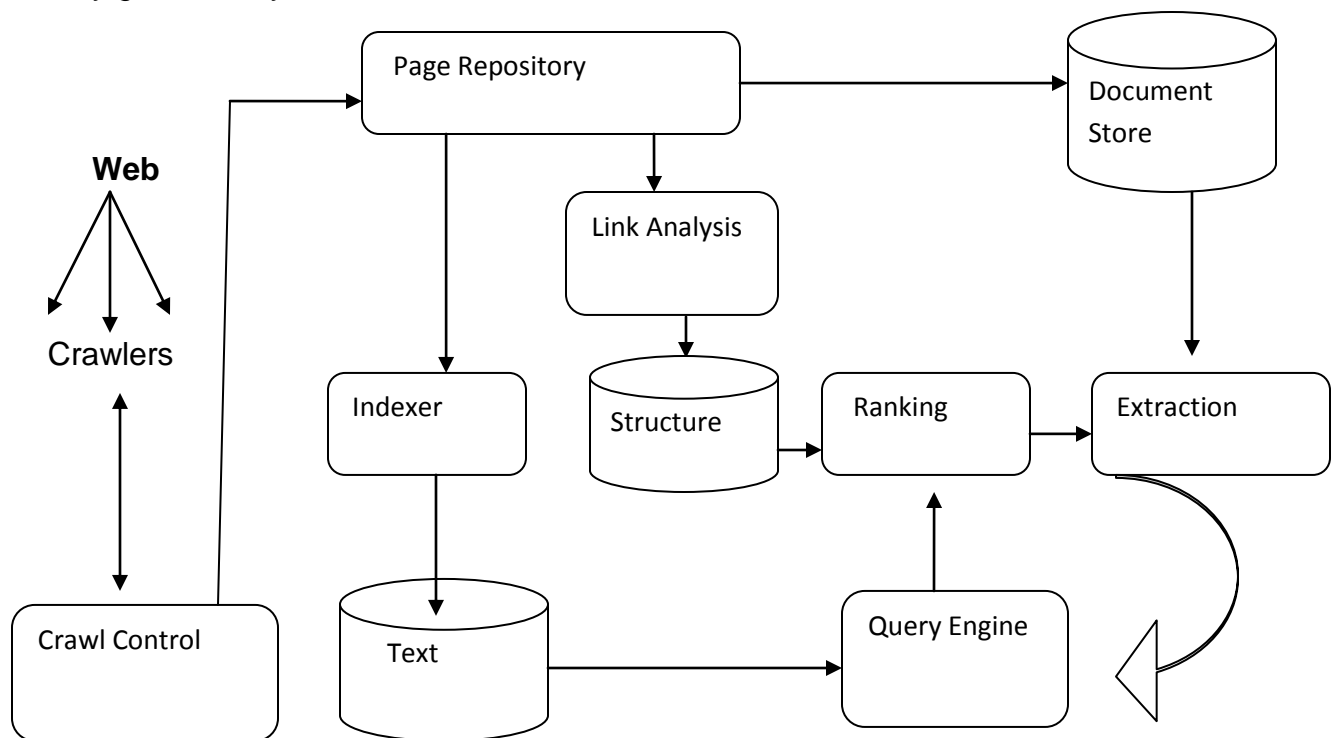
Figure 2.1 Search Engine Architecture

There are basically three types of search engines: Those that are powered by robots (called crawlers; ants or spiders) and those that are powered by human submissions; and those that are a hybrid of the two.

Crawler-based search engines are those that use automated software agents (called

crawlers) that visit a Web site, read the information on the actual site, read the site's meta tags and also follow the links that the site connects to performing indexing on all linked Web sites as well. The crawler returns all that information back to a central depository, where the data is indexed. The crawler will periodically return to the sites to check for any information that has changed. The frequency with which this happens is determined by the administrators of the search engine.

Human-powered search engines rely on humans to submit information that is subsequently indexed and catalogued. Only information that is submitted is put into the index.

In both cases, when you query a search engine to locate information, you're actually searching through the index that the search engine has created —you are not actually searching the Web. These indices are giant databases of information that is collected and stored and subsequently searched. This explains why sometimes a search on a commercial search engine, such as Yahoo! or Google, will return results that are, in fact, dead links. Since the search results are based on the index, if the index hasn't been updated since a Web page became invalid the search engine treats the page as still an active link even though it no longer is. It will remain that way until the index is updated.

So why will the same search on different search engines produce different results? Part of the answer to that question is because not all indices are going to be exactly the same. It depends on what the spiders find or what the humans submitted. But more important, not every search engine uses the same algorithm to search through the indices. The algorithm is what the search engines use to determine the relevance of the information in the index to what the user is searching for.

Figure 2.2 Context sensitive spell check (google search)

One of the elements that a search engine algorithm scans for is the frequency and location of keywords on a Web page. Those with higher frequency are typically considered more relevant. But search engine technology is becoming sophisticated in its attempt to discourage what is known as keyword stuffing, or spamdexing.

Another common element that algorithms analyze is the way that pages link to other pages in the Web. By analyzing how pages link to each other, an engine can both determine what a page is about (if the keywords of the linked pages are similar to the keywords on the original page) and whether that page is considered "important" and deserving of a boost in ranking. Just as the technology is becoming increasingly sophisticated to ignore keyword stuffing, it is also becoming savvier to Web masters who build artificial links into their sites in order to build an artificial ranking.

Chapter 3:- Search Engine Basics

i.) Meta-tag:-

A special HTML tag that provides information about a Web page. Unlike normal HTML tags, meta tags do not affect how the page is displayed. Instead, they provide information such as who created the page, how often it is updated, what the page is about, and which keywords represent the page's content. Many search engines use this information when building their indices

ii.) Keyword stuffing

A SEO technique used by Web designers to overload keywords onto a Web page so that search engines will read the page as being relevant in a Web search. Because search engines scan Web pages for the words that are entered into the search criteria by the user, the more times a keyword appears on the Web page the more relevancy the search engine will assign to the page in the search results (this is only one way that search engines determine relevancy, however.) Search engines often penalize a site if the engine discovers keyword stuffing, as this practice is considered poor netiquette, and some search engines will even ban the offending Web pages from their search results.

There are several methods of keyword stuffing. One way is to insert repeating keywords within the *input type="hidden"* field Meta tag or the *keyword* tag so that the keywords are not seen by the user but are scanned by the search engine. Another way is to make text in the body of the Web page *invisible text,* or *hidden text,* by making the text the same color as the page??s background, rendering the text invisible to the user unless the user highlights the text. This method is called *invisible keyword stuffing* .

Keyword stuffing also is referred to as keyword loading and *spamdexing.*

iii.)   Algorithms

A formula or set of steps for solving a particular problem. To be an algorithm, a set of rules must be unambiguous and have a clear stopping point. Algorithms can be expressed in any language, from natural languages like English or French to programming languages like FORTRAN.

We use algorithms every day. For example, a recipe for baking a cake is an algorithm. Most programs, with the exception of some artificial intelligence applications, consist of algorithms. Inventing elegant algorithms -- algorithms that are simple and require the fewest steps possible -- is one of the principal challenges in programming

The most widely used Algorithm for Search Engines is

TF-IDF Ranking Algorithm

TF- Stands for term Frequency: - How many times a word appears in a document

IDF-Stands for Inverse document frequency:-

Number of documents/number of documents containing the term (Actually the logarithm of this.)

TF-IDF of a keyword in a page = **TF * IDF**

Example:-

100 web pages. Keyword: *mbm engg. college*

#1 has 8 mentions. TF = 8.

#2, 17, 19, 76 have 4 mentions. TF = 4.

20 pages have 1 mention. TF = 1.

IDF = $\log_2 (100 / 25) = 2$

TF-IDF of *mbm engg. College* in #1 = 8 * 2 = 16 High!

TF-IDF of *mbm engg. College* in #2, 17, 19, 76 = 4 * 2 = 8 not so high

TF-IDF of *mbm engg. College* in 20 others = 1 * 2 = 2Small

TF-IDF of *mbm engg. College* in all the rest = 0 * 2 = 0 Irrelevant

Some Other Popular algorithms are

- Vector space
- Probabilistic (binary interdependence)
- Fuzzy set theory
- Bayesian statistical analysis
- Latent semantic indexing
- Neural networks
- Machine learning

iv.) Spider (crawlers)

A program that automatically fetches Web pages. Spiders are used to feed pages to search engines. It's called a spider because it *crawls* over the Web. Another term for these programs is WebCrawler.

Because most Web pages contain links to other pages, a spider can start almost anywhere. As soon as it sees a link to another page, it goes off and fetches it. Large search engines, like Alta Vista, have many spiders working in parallel.

Chapter 4:- Google's Working Principle

The following steps are a part of the Google Search Framework:

a. Crawling

b. Indexing

c. Query Processing

d. Ranking

Crawling: This step initiates the round-the-clock activity of search engines the accumulate. Data of the new & updated content from all the web pages on the web. Crawlers / Spiders / Robots /Bots are Search Engine specific software that collects data from the Web.

Indexing: This step is required for converting the unstructured page data (as the Crawler sees it.

In the web/Google terminology, it is also called the Lynx view) TO a structured format so that it can be used efficiently for Query processing.
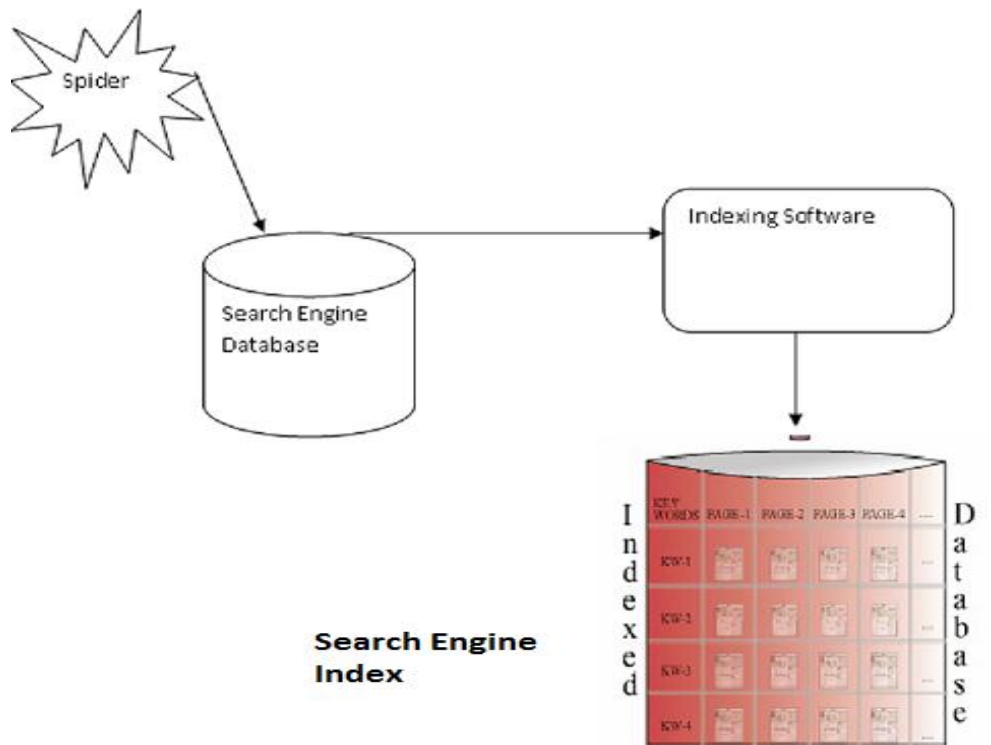
Figure 4.1 Indexing (www.telezent.com)

Query Processing: This step takes the User query (Keyword) as its input, which the user enters into the SEARCH BOX. The outcome of this step is the short listed pages from trillions of pages

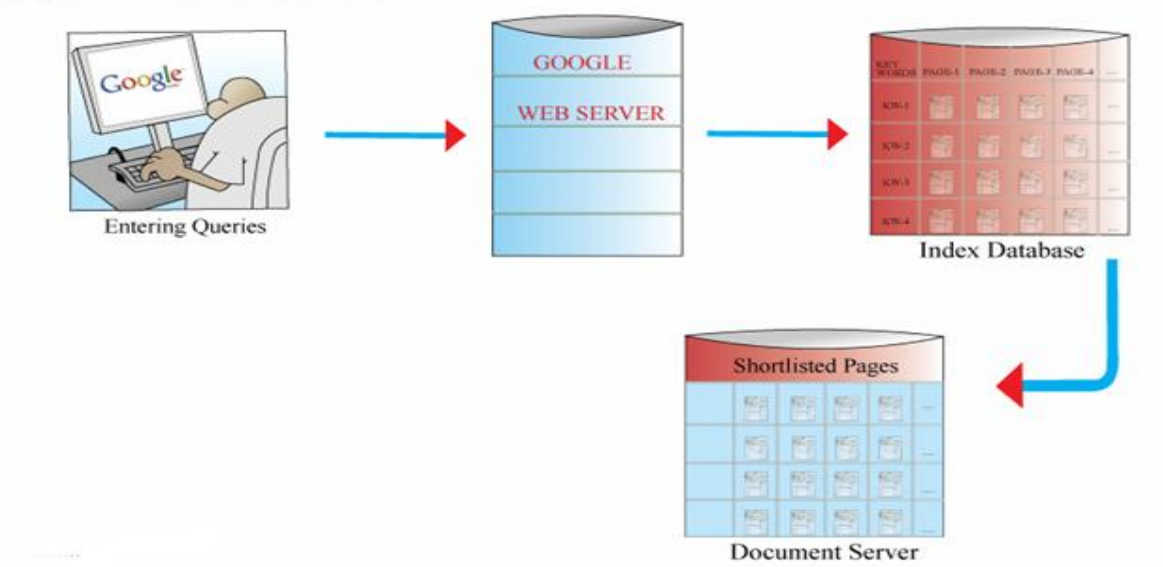and which are relevant for this query.

Figure 4.2 query processing (www.telezent.com**)**

Ranking: This is the final step of the search process. Google algorithm comes into picture here
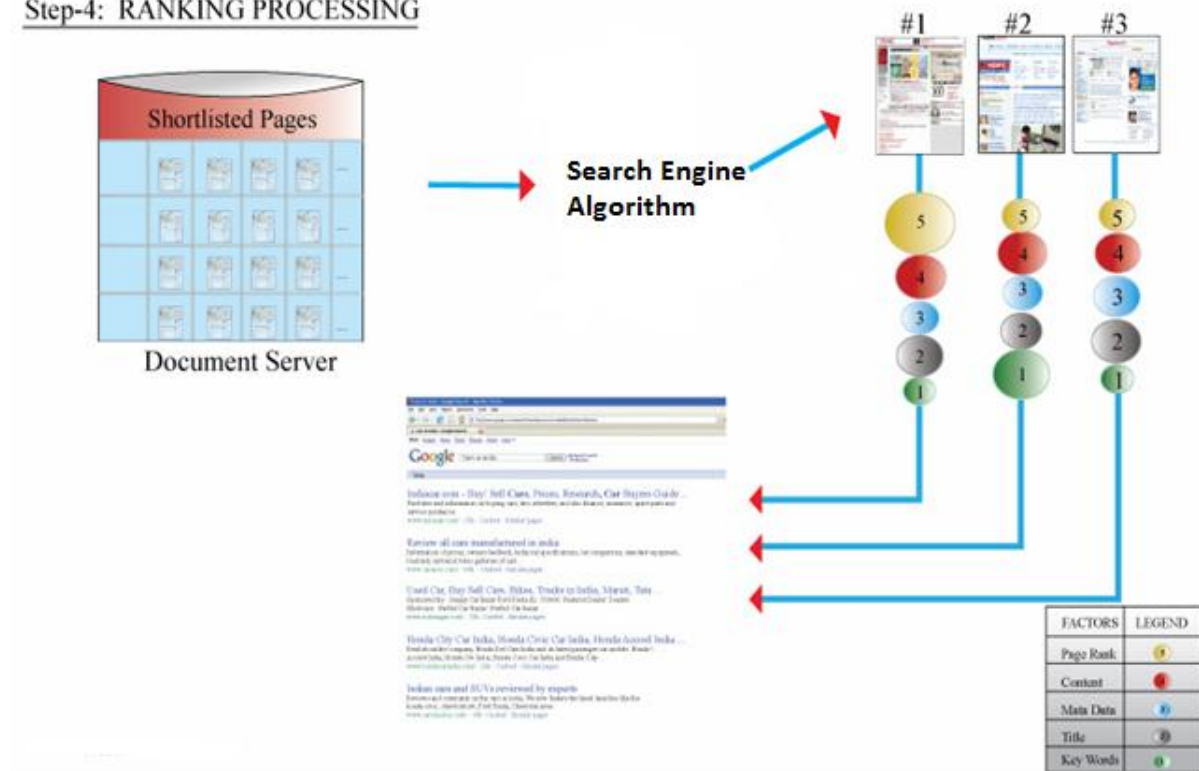
Figure 4.3 Ranking (www.telezent.com**)**

And applies around 200 factors to compare the short listed pages from the previous step. The ranked pages appear in order of relevance on the SERP.

The technology behind Google's great results

As a Google user, you're familiar with the speed and accuracy of a Google search. How exactly does Google manage to find the right results for every query as quickly as it does? The heart of Google's search technology is Pigeon Rank, a system for ranking web pages developed by Google founders Larry Page and Sergey Brin at Stanford University.

Figure 4.4(Google image)

Building upon the breakthrough work of B. F. Skinner, Page and Brin reasoned that low cost pigeon clusters (PCs) could be used to compute the relative value of web pages faster than human editors or machine-based algorithms. And while Google has dozens of engineers working to improve every aspect of our service on a daily basis, Pigeon Rank continues to provide the basis for all of our web search tools.

Why Google's patented Pigeon Rank works so well

Pigeon Rank's success relies primarily on the superior trainability of the domestic pigeon (Columba livia) and its unique capacity to recognize objects regardless of spatial orientation. The common gray pigeon can easily distinguish among items displaying only the minutest differences, an ability that enables it to select relevant web sites from among thousands of similar pages.

By collecting flocks of pigeons in dense clusters, Google is able to process search queries at speeds superior to traditional search engines, which typically rely on birds of prey, brooding hens or slow-moving waterfowl to do their relevance rankings.

When a search query is submitted to Google, it is routed to a data coop where monitors flash result pages at blazing speeds. When a relevant result is observed by one of the

19
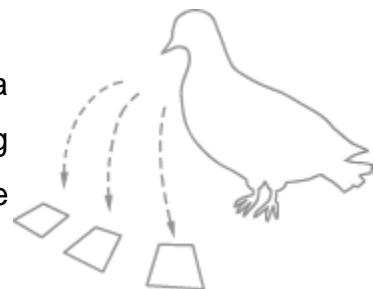
pigeons in the cluster, it strikes a rubber-coated steel bar with its beak, which assigns the page a Pigeon Rank value of one. For each peck, the Pigeon Rank increases. Those pages receiving the most pecks, are returned at the top of the user's results page with the other results displayed in pecking order.

Integrity

Google's pigeon-driven methods make tampering with our results extremely difficult. While some unscrupulous websites have tried to boost their ranking by including images on their pages of bread crumbs, bird seed and parrots posing seductively in resplendent plumage, Google's Pigeon Rank technology cannot be deceived by these techniques. A Google search is an easy, honest and objective way to find high-quality websites with information relevant to your search.

Data
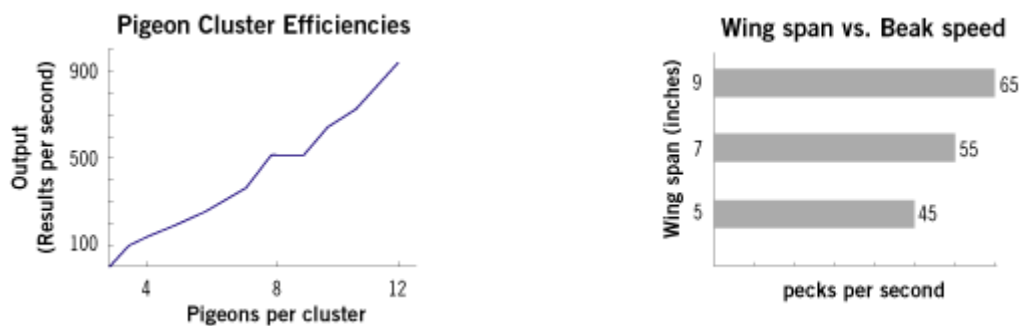
Figure 4.5(statistics by Google group)

Chapter 5:- How Bing web search engine works?

Figure 5.1 (www.bing.com)

As we are know that behind this search engine, a big name is there that is Microsoft which is responsible for this search engine working and due to this fact only that It is among the top 3 web search engines.

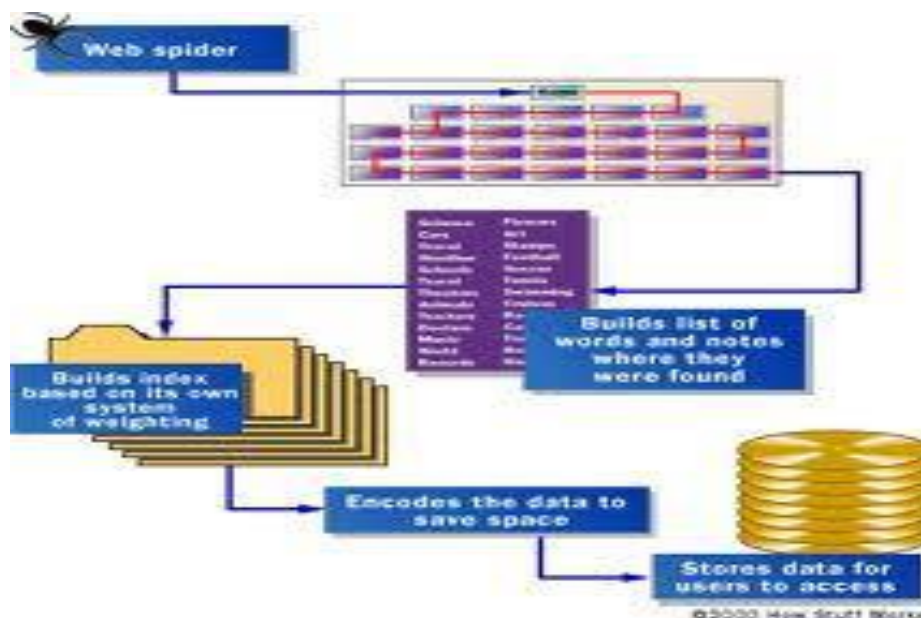The operation carried out by the web searching Bing engine is explained as follows-

Figure 5.2   Working Principle of Bing

a.)   The Bing search engine makes a list of the web pages which consist of the information about any topic and all this information can be retrieve from the HTML language itself which is consider to be the client side language for the data accessing of the website through the web-pages of that website

b.)   The pages in Bing search engine can be retrieved by the web crawler which is known with the spider name of the automatic web browser that links the every link of the website with the browser and exclusions of the robots.txt can be taken with the indexing which will arrange the contents of each page according to the user. For example- Special words can be selected from the title, headings or special fields which is known as meta data tags and these special words are then indexed with the index database for future queries that can be either a single word or a sentence and the purpose of it is to allow the information to access easily.

c.) It stores the part of the source page and information about the web page in its cache that holds the actual search text during the indexing which is helpful to access the contents of the current page after updation and the worst thing is that problem of link rot occurs which increases the usability of the user by returning it on the same web page.

d.) The least astonishment principle is implemented in it according to which the user returns on the same web page in which he/she was earlier surfing and this increases the importance of that web page**.**

e.) In case when any user enters a query in the search engine then the keywords of the user are examined and engine starts indexing and provides a list of the best matching results by providing the data of the whole file which is indexed by the engine according to the short summary that contains the title and different tags that is source of the attraction for the user.

f.) The indexing is done from the information that is stored in the database by using the gate function that uses the gates operations of the Boolean such as AND, OR and NOT and specify the search engine and looks for the same spell of the word by providing the advanced features such as proximity search by using the keywords for advancement.

g.) Sometimes, the concept based searching is also carried out which includes the statistical analysis that includes the words or phrases for the language query which would be understandable and the search engine importance is depend upon the relevance of the result set and increases the internet usage by means of the predefined and hierarchical ordered keywords that are programmed by the user.

h.) The system also generates the inverted index by analyzing the text that locates with in the file and for that indexing of the files many search engines offers revenue for the advertiser to pay money according to the listed ranking.

Chapter 6:- Comparison between Google and Bing

Ultimately, both search engines are going to provide you with the information that you desire in an easy and efficient manner. At the moment, it appears that Google still has most of the advantages over Bing. If you test Bing out you aren't likely to be disappointed, but you aren't necessarily going to be converted either. And when it comes to deciding to **make a website** and figuring out how to promote it, either will work.

Here's a quick breakdown of how the two engines compare and contrast.

Visuals

Google features an infamously minimalist web page design, and they don't exactly add a lot of aesthetic flair once you've inputted your search keywords. Bing, on the other hand, clearly aims to provide a more visually appealing search experience- something they announce from the beginning with the beautiful high quality photographs that greet you on their landing page.

While Bing looks nicer, most people aren't necessarily concerned about looks when it comes to their online searches. Google's overall philosophy is notoriously concerned with performance over aesthetics, and their streamlined web page template contrasts with Bing's the same way their sleek browser Google Chrome contrasts with Microsoft's bloated Internet Explorer.

Search Results

The main way that Google and Bing differentiate from each other is the way they organize their search results on the web page background. Both turn up very similar results for the same search term, and both provide essentially the same search options and will be of help when you make a website and SEO optimize it, but they simply organize their results slightly differently.

Google provides a more extensive breakdown of sources for your search results than Bing does. For example, once you input a Google search you can then look to the left of the results and select whether you only want to search for your keyword in Books, in Blogs, in News, or a number of other content sources. You get these same options regardless of the keyword you search for.

Bing also allows you to narrow the source of your search content, but they individualize these sources depending on what you're searching for. If you are searching for the term "President" you are going to see different search sources to choose from than if you search for the term "San Francisco."

Both search engines provide standards like image searches and video searches.

Research and the Sidebar

The extra search functions are the main difference between Google and Bing at this point in their evolution.

With Bing, you get a quick and easy listing of "related searches" to your main search query. Both search engines feature an auto-complete function on their searches. Bing also contains your most recent search history, which you can clear at will. Overall, their sidebar is designed around minimizing the number of individual searches you need to employ to get the information you want, and to making it as fast as possible to jump around your different recent searches. As it's fairly common to repeatedly search for the same terms when you are doing research, this is a welcome feature.

On the other hand, Google provide more in depth search functions on their sidebar, giving them the edge when you are doing research on recent news stories or discussions about a topic. The real key here is the fact that you are able to winnow down your searches to a specific **domain name** or time frames. You can search for the latest news, or for news that has popped up in the last 24 hours, 3 days, week, month, year... or even a custom range of dates.

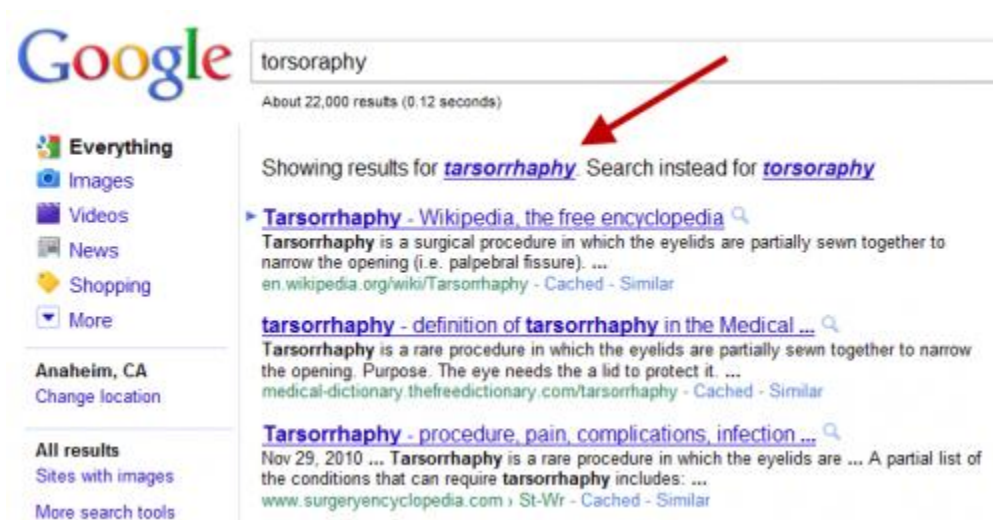For example, consider a search for torsoraphy, which causes Google to return this:

Figure 6.1 http://searchenginewatch.com/

In the example above, Google's searched for the correct spelling — tarsorrhaphy — even though torsoraphy was entered. Notice the top listing for the corrected spelling is a page about the medical procedure at Wikipedia.

Over at Bing, the misspelling is NOT corrected — but somehow, Bing manages to list the same Wikipedia page at the top of its results as Google does for its corrected spelling results:
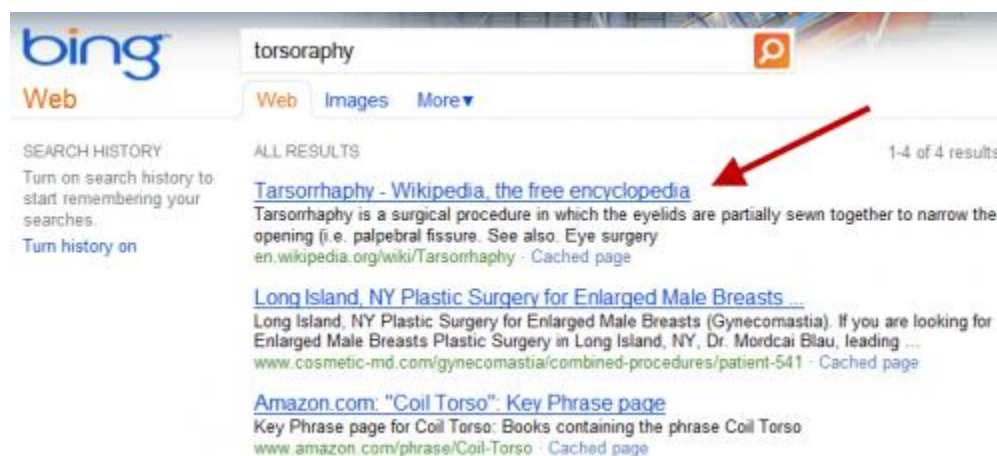

Figure 6.2 http://searchenginewatch.com/

Got it? Despite the word being misspelled — and the misspelling not being corrected — Bing still manages to get the right page from Wikipedia at the top of its results, one of four total pages it finds from across the web. How did it do that?

It's a point of pride to Google that it believes it has the best spelling correction system of any search engine. Google even claims that it can even correct misspellings that have never been searched on before. Engineers on the spelling correction team closely watch to see if they're besting competitors on unusual terms.

So when misspellings on Bing for unusual words — such as above — started generating the same results as with Google, red flags went up among the engineers.

To Sting A Bing

To verify its suspicions, Google set up a sting operation. For the first time in its history, Google crafted one-time code that would allow it to manually rank a page for a certain term (code that will soon be removed, as described further below). It then created about 100 of what it calls "synthetic" searches, queries that few people, if anyone, would ever enter into Google.

These searches returned no matches on Google or Bing — or a tiny number of poor quality matches, in a few cases — before the experiment went live. With the code enabled, Google placed a honey pot page to show up at the top of each synthetic search.

The only reason these pages appeared on Google was because Google forced them to be there. There was nothing that made them naturally relevant for these searches. If they started to appeared at Bing after Google, that would mean that Bing took Google's bait and copied its results.

This all happened in December. When the experiment was ready, about 20 Google engineers were told to run the test queries from laptops at home, using Internet Explorer, with Suggested Sites and the Bing Toolbar both enabled. They were also told

to click on the top results. They started on December 17. By December 31, some of the results started appearing on Bing.
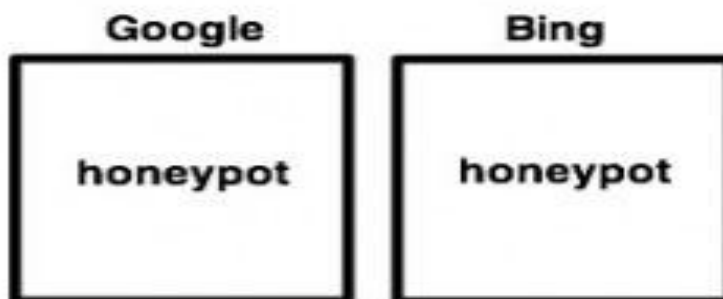
How The Bing Sting Worked

Google   Bing

0 matches   0 matches
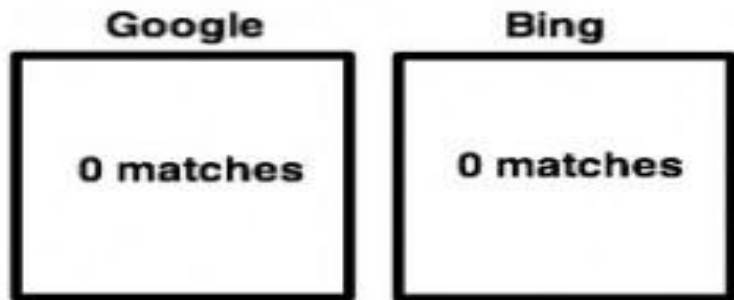
(1) Neither Google or Bing have a match for an unusual word

Google   Bing

honeypot   0 matches
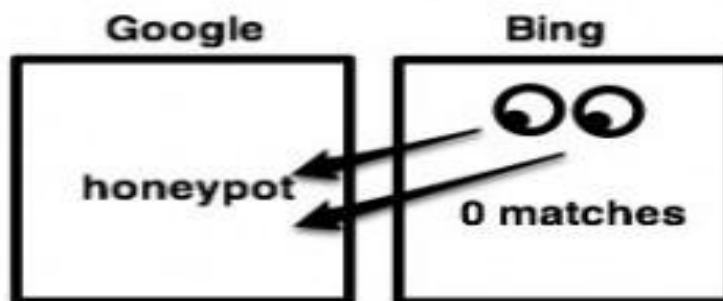
(2) Google makes "honeypot" page appear for word, which Bing sees via Internet Explorer user at Google

Google   Bing

honeypot   honeypot

(3) Bing begins ranking honeypot page for the unusual word

Figure 6.3 http://searchenginewatch.com/

Here's an example, which is still working as I write this, hiybbprqag at Google:
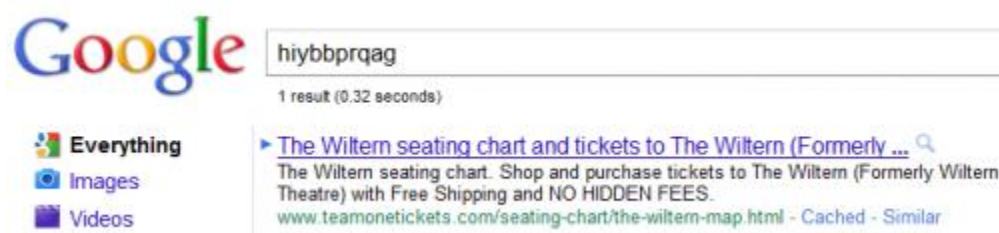
Figure 6.4

and the same exact match at Bing:

Figure 6.5 http://searchenginewatch.com/
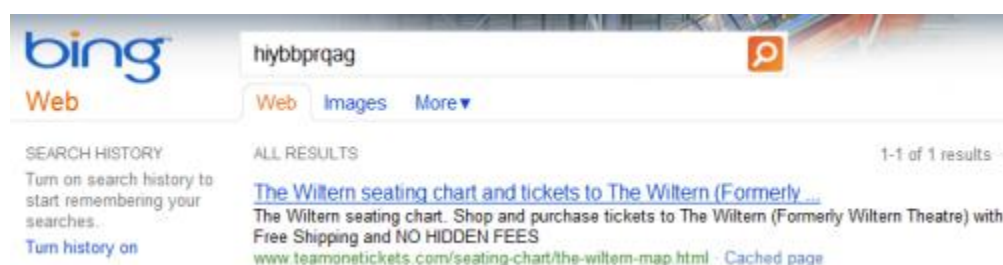
Here's another, for mbzrxpgjys at Google:

and the same match at Bing:

Here's one more, this time for indoswiftjobinproduction, at Google:

And at Bing:

To be clear, before the test began, these queries found either nothing or a few poor quality results on Google or Bing. Then Google made a manual change, so that a specific page would appear at the top of these searches, even though the site had nothing to do with the search. Two weeks after that, some of these pages began to appear on Bing for these searches.

It strongly suggests that Bing was copying Google's results, by watching what some people do at Google via Internet Explorer.

Chapter 7:- Search Engine Controversies

As the 21st century began, controversy erupted over the issue of search engine censorship. In late 2005 and early 2006, Cisco Systems, Inc., Google, MSN, and Yahoo! came under fire for cooperating with the Chinese government in censoring Internet content or providing assistance to trace political dissidents who used the Internet in China. Reporters without Borders, a France-based organization that promotes freedom of the press, accused Yahoo! of helping the Chinese government identify two dissidents through their Yahoo! e-mail accounts. One dissident, who posted essays on the Internet discussing political corruption in China, was sentenced to eight years in prison for "inciting subversion." Another, who had used a Yahoo! e-mail to send information about the Tiananmen Square protest, was sentenced to ten years in prison.

In February 2006 a congressional committee, the House Subcommittee on Africa, Global Human Rights and International Operations, heard testimony on the issue. The

Subcommittee heard reports that Google's search engine in China filtered out certain search terms, such as human rights and Tiananmen, and that MSN had shut down a Chinese citizen's blog at the request of the government and disallowed blogs on its MSN Spaces servicing China that contained terms forbidden by the Chinese government. Cisco Systems was criticized for providing hardware used to filter, or censor, Internet information. All four companies defended their practices by saying they were forced to abide with Chinese laws. Company representatives also stated their belief that it was the role of government, not business, to promote democracy and human rights in China. Google's representative said the company found it an unsatisfactory compromise to create Google.cn but decided it could make a "meaningful, though imperfect, contribution to the overall expansion of access to information in China." The chair of the House subcommittee, Representative Christopher H. Smith, a Republican of New Jersey, said he would introduce legislation called the Global Online Freedom Act to restrict an Internet company's ability to censor information regardless of another country's laws.

Chapter 8:-Conclusions

i.) Search engine plays important role in accessing the content over the internet, it fetches the pages requested by the user.

ii.) It made the internet and accessing the information just a click away.

iii.) A quality search process begins with quality metadata. It's that old principle: Garbage in, garbage out. Metadata is about giving a structure the content. For example, if every document is assigned keywords or or classified by Geography, the reader will get a much more accurate return from his or her search.

iv.) Search engines are the mortar of the Intranet. As important as they are, their implementation must be given high priority with the necessary time allotted for research and development.

v.) The need for better search engines only Increases.

vi.) Since search is such a common activity, the search box should appear on every page of your web site.

vii.) The initial target of the basic search should be the contents of the entire web site.

viii.) The basic search should allow for Boolean commands ("and," "or"), although this does not need to be explained.

Part II Bibliography

The data are collected from the following sources…

Web links

i.)   Wikipedia.   Till seminar

ii.)   www.telezent.com   1-may-2011

iii.)   http://searchenginewatch.com/   24-april-2011

iv.)   http://homeforprofits.com/search-engines/working-of-a-search-engine/
        15-april-2011

v.)   http://computer.howstuffworks.com/internet/basics/   15-april-2011

Books

Computer Networking Top down Approach by Kurose and Ross.

 "Search Engine." Microsoft® Encarta® 2009 [DVD]. Redmond, Microsoft Corporation, 2008.

Microsoft ® Encarta ® 2009. © 1993-2008 Microsoft Corporation.