Chapter-2

Search Engines

CHAPTER TWO
SEARCH ENGINES

2.1 INTRODUCTION

As we know that search engines play a vital role in electronic commerce environment, so it is necessary to understand the functioning of search engines. When a user fires a query to the search engine, he/she is actually not searching the whole web but instead searches the index of the search engine over the web [36]. There are so many limiting factors which restrict the user searching the whole web every time such as cost, speed and capabilities.

The steps to search a specific piece of information by a general query are as follows: An Internet user submits a query by typing a phrase or word in the search box.

1. Regarding the query, search engine looks through all the pages that it keeps in its database.
2. Search engine identifies the relevant web pages
3. Results are displayed on the Search Engine Results Page **(SERP)** in an order, starting with the most relevant results to least relevant pages.

Within a fraction of second, the whole search process complets but more complex algorithms are running behind the search process.

There are three major parts of te Web search engines.

1. Web Crawler
2. Indexer
3. Query Processor

Crawling is a computerized programmed process to collect the data with web spiders. Crawlers are also known as wanderers, software agents, robots, walkers, or knowbots [27].

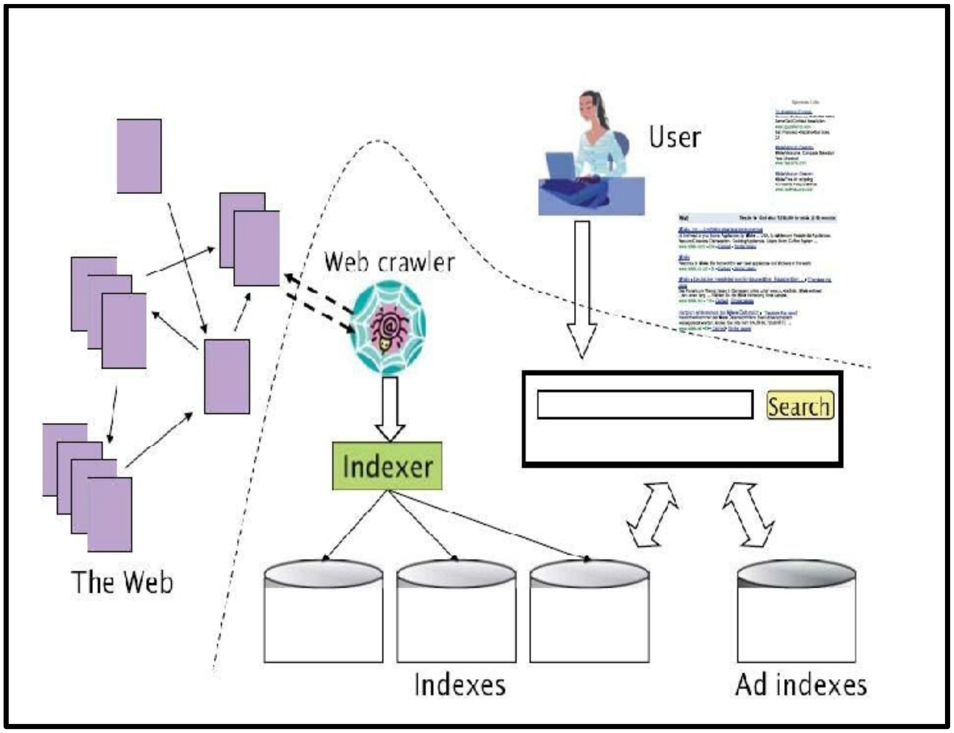

Figure 2.1: Components of Web Search Engine

Source [Manning, Raghavan and Schutze, 2009]

A full text index is built by the web crawler by fetching the web pages and making list of words as phrases in a database of the search engine. The URLs are added to the list to report the existence of a web page [36] or follow the link structure of the web. Spiders generally begin by crawling a few web pages and track the links on those pages. Spider follows the links till it reaches to the last web page. The process of indexing the URLs is a continuous process and takes place until a substantial part of the web pages are indexed. The pages indexed so far are stored across many computer systems throughout the world.

Indexing is the process of “taking the raw data and categorizing it, removing duplicate information and generally organizing it all into an accessible structure” [27].

The indexes identified so far are stored and organized in the database and generally an inverted index data structure is used [36]. Inverted index is helpful in keyword based queries and it helps in retrieving the results quickly.

Given below is an example of inverted index data structure. The keywords are stored in an alphabetic order. In this example, there are four phrases. Each word in the phrase is assigned a numbers. The first number is the identifier for every phrase. Numbers from one through four are assigned to each phrase. The location of the word within the phrase is indicated by the second number. Words like “and”, “is”, “the” or “you” make no sense and are known as stop words. These words are redundant by some search engines as they reduce the performance of the search engines.

Keyword/word	**Doc #, Position**		
and	(4,2)		
beautiful	(3,3)		
is	(3,2)		
life	(1,3)	(2,3)	(3,1)
live	(1,5)	(2,1)	
love	(1,1)	(2,5)	
the	(1,2)	(2,2)	
you	(1,4)	(2,4)	

Figure 2.2 Inverted Index Data Structure

Phrases:

1. “Love the life you live”
2. “Live the life you love”
3. “Life is beautiful”
4. “Live and love”

It is quite impossible to index the whole web not only due to technical limitations but cost incurred in indexing the whole web will also be very large so it is not feasible to crawl the whole web [76].

The query processor which is most important component of the web search engine consists of the form, relevant documents in the database with corresponding search query and the results-output formatter which is the search engine results page [36].
The special techniques are used by the search engines to locate the relevant documents indexed for a specific query. "The major differentiator of one search engine from another lies in the way relevance is calculated" [36]. The weight is assigned to different criteria by every search engine and complex algorithms are used to choose which results are to be displayed to the user in what order [27].

Search engines employ complex mathematical equations for ranking the results retrieved by search engines to attain optimization. Results are ranked by search engines using query dependent factors (on the page criteria) and query independent factors (off the page criteria).

There are basically two factors. First one is query dependent factors and another one is query independent factors. The query dependent factors include word count frequency, language of the document and geographical distance [55]. While, query independent factors tries to find out the quality of a document, irrespective of the given query. These factors are generally based on link analysis. PageRank [47], which is a measure of link popularity, belongs to the most famous query independent factors.

Various techniques such as spamming are used to prevent the manipulation on ranking of the results because many designers attempt to improve the ranking of the results in the search engine result page [163].

Results retrieved by search engines are of two types:

1. Natural Search Results
2. Paid Search Results

Natural search results are also known as organic search results and include web pages found through spiders. Paid results are also known as sponsored links like advertisement for which webmasters pay and are based on keywords. Paid results are always separated from natural results [27].

Web Directories

As indicated by Berners-Lee [175], the web is considered to be an organization of hypertext link. Web directories present a context-based framework for structured browsing. It is also known as subject directories or catalogs. Web directories are compared with table of contents in a book [36], because of hierarchical structure to provide high level outline of important topics just like table of contents. Search engines are compared with index of a book.

Search engines are also compared with telephone directory having names and address list. Web directories on the other hand compared with yellow pages structured by categories and give descriptive information [36].

Search engines use autonomous software programs while in directories, information about websites and pages is manually placed into specific categories with the help of human editors. Directories are also known as "human based search engines". The way in which data are arranged in search engine and web directories is the key difference [172].

Web directories are created by human editors by evaluating site by site and page by page and deciding whether site or page is important enough to be added to the web directory. The topics are classified on the basis of similarity of pages and a hierarchical structure is created to make things easier from a user's perspective [163]. Like search engine, web directories also support paid results and ranking of the website. The directory listings are based on the title of the web site and its description.

2.2 TYPES OF SEARCH ENGINES

Generally, the term “Search Engines” is used to illustrate both the directory based search engines and crawler based search engines [108]. These search engines are different in the way they gather the information for their listings.

1. **Crawler-based search engines:** Crawler based search engines develop their listings by using software agent known as crawler or spider. The crawler indexes the web pages by crawling the whole web periodically. Examples of crawler based search engines are Google, Altavista etc. Any change in the web pages can be identified by the crawler and will influence the listing of web pages in the search engines.

   When we want to find relevant information for a specific search query then in this condition crawler based search engine works efficiently and provides relevant results. But, when we are interested in searching the general search query, then it provides many irrelevant search results.

2. **Directory based search engines:** Directory based search engines or human powered directories develop their listings by human editors, for example, open directory and Yahoo directory.

   When we want to search for general query by human powered directory then in this condition, human directory helps us and provides refined and relevant search results but it does not provide relevant results i.e., does not work efficiently when we search a specific query. So, we can say crawler based search engines and human powered directories are opposite in their functioning.

From the Table 2.1, it is clear that some search engines provide search results from both types of search engines i.e., directory based and crawler based search engines, for example, MSN and Yahoo. These types of search engines are known as hybrid search engines.

Different search engines and their types are given in the table below:

Name of Search Engines	Types
Google	Crawler based
Yahoo	Directory based as well as crawler based search results
Lycos	Crawler based search results
HotBot	Crawler based search results powered by AllTheWeb, Google, Inktomiand Teoma
AskJeeves	Crawler based search results
AOL Search	Crawler based search results
MSN Search	Directory based search engine powered byLookSmart and Crawler based search results powered by Inktomi
Open Directory	Directory based Search engine
LookSmart	Directory based Search engine
AltaVista	Crawler based search engine
Inktomi	Crawler-based search engine
AllTheWeb	Crawler based search engine
Netscape Search	Crawler based search results

Table 2.1: Search Engines and Their Types

2.3 CHARACTERISTICS OF SEARCH ENGINES

There are certain parameters on the basis of which the results are retrieved by the search engines. The results retrieved by different search engines are different. There are some characteristics of the search engines which makes one search engine different from another search engine.

Some of the characteristics of search engines are discussed below:

1. **Web Crawling or Spidering**

   A web based crawler is a software agent or program that crawls the whole web. It tracks the list of URLs known as seeds. These URLs are recognized by the web crawler from many different sources and are stored in the local database of the web search engines. In order to keep the web database of the search engines up to date, the crawler crawls the web periodically.

There are some important parameters that may cause web crawling to pose certain limitations [33] such as ever dynamic page modification, increasing database size, huge amount of the data mined, high rate of modification etc. The following policies determine the behavior of the web crawler [129].

a. Parallelization: This policy decides the communication among distributed web crawlers.

b. Politeness policy: Congestion of the web sites is decided by this policy

c. Re-visit policy: The rate with which the crawlers check for updates and modifications [180].

d. Selection policy: Selection criteria for the pages that are to be downloaded to the web database of the search engine are determined by this policy.

2. Result Matching

The results matching technique is used to determine the all relevant pages in the database of search engine corresponding to a query. Different matching algorithms are used by different search engines to show more relevant pages in the search results. There are certain challaneges in matching the search results from the vast database of search engines [192].

a. Parsing: Parsing algorithms may pose difficulties if they encounter complex Hyper Text Markup Language (HTML) used in some of the web pages.

b. Filtering: A search engine needs to perform effective filtering in order to show the most relevant results to the searchers. It is really significant to show unique results to the user by minimizing the chances of redundancy [139].

3. Result Ranking

The order in which the search results are displayed to the user is known as result ranking. There are number of results which can be displayed to the user but the order in which the results are displayed matters. It would be better for the user if the desired results are shown to the user in first or second page of the search engine result page [166].

There are some certain factors on the basis of which the result ranking algorithms works such as:

a. **Frequency**: The algorithm looks out for how often are the search keywords recurring in the context of the search results. Occurrence of search keywords is not considered to be an ideal factor as it gets biased to content-rich pages [158].

b. **Location**: It is significant for the search engine to look out for the search keywords at the top of a webpage. For example: looking for the search keywords in the title of a webpage [56].

4. Single source search engines and Meta-search engines

Search engines are classified as either single source search engines or Meta search engines. When the search results are retrieved by only one search engine then it is known as single source search engine. But, when the results are retrieved by more than one search engines then it is known as Meta search engines.

There is clear advantages of using Meta search engine over single source search engine. It offers more coverage than single source search engine by facilitating more number of databases [169].

Google, Altavista, yahoo etcetera are the example of single source search engine while Dogpile and Mamma are example of Meta search engine.

2.4 INTRODUCTION OF GOOGLE, YAHOO AND BING

It is found from the study [53] of overlapping between first page results on Google, AskJeeves and Yahoo that only 85% links are distinct whereas 12% links were found to be common in any of two search engines and 3% of the results were found to be common across all three search engines. As there is very small amount of overlapping, search engines vary from each other in ranking and retrieval policies. We can conclude that no single search engine can retrieve all possible relevant search results and user may fail to retrieve some of the relevant results [17].

Above discussion stimulates us to implement the Meta search engine which can provide the results by combining the results from existing search engines. It also refines the results with the help of ranking and merging algorithms.

There is no need of crawling the web and indexing the database in case of Meta search engine whereas it is required in case of search engine. This is the basic difference between the search engine and Meta search engine.

Web designers want their sites to be placed and get listed in major search engines such as Bing, Yahoo and Google which in turn may cause a lot of traffic. Therefore, these search engines are of at most interest to the web developers and designers.

Search Query Report (January 2012 versus December 2011) **Source: ComScore Search**			
Core Searches	**Explicit Core Search Queries (MM)**		
	December 2011	**January 2012**	**Percent Change**
Total Core Search	18,235	17,804	-2%
Google	12,019	11,786	-2%
Microsoft	2,750	2,707	-2%
Yahoo!	2,647	2,506	-5%
Ask Network	531	527	-1%
AOL, Inc.	287	277	-3%

Table 2.2: Search Query Report (December 2011)

Google is one of the most important and largest search engines in the word. A survey conducted by ComScore shows that in the year 2013 Google had 67.5% searches (18,266 billion searches) whereas Yahoo with 2,111 billion searches and Bing ranked forth with 3,048 billion searches [44].

In January 2013, 19.484 billion searches were conducted in which Google ranked first with 13.06 billion searches. In second position, Microsoft Sites were placed with 3.208 billion searches, followed by Yahoo with 2.348 billion [43].

Search Query Report (January 2013 versus December 2012)

Source: ComScore Search

Core Search	**Explicit Core Search Queries (MM)**		
	December 2012	**January 2013**	**Percent Change**
Total Core Search	17,628	19,484	11%
Google	11,763	13,060	11%
Microsoft	2,867	3,208	12%
Yahoo!	2,154	2,348	9%
Ask Network	534	536	0%
AOL, Inc.	310	331	7%

Table 2.3: Search Query Report (December 2012)

Studies of Nielsen-Net show market share rankings of search engines. With 89% market share, Google is found at first place and Bing with 4%, Baidu with 4% and Yahoo! With only 2%, while 1% of the total market share is captured by other search engines.

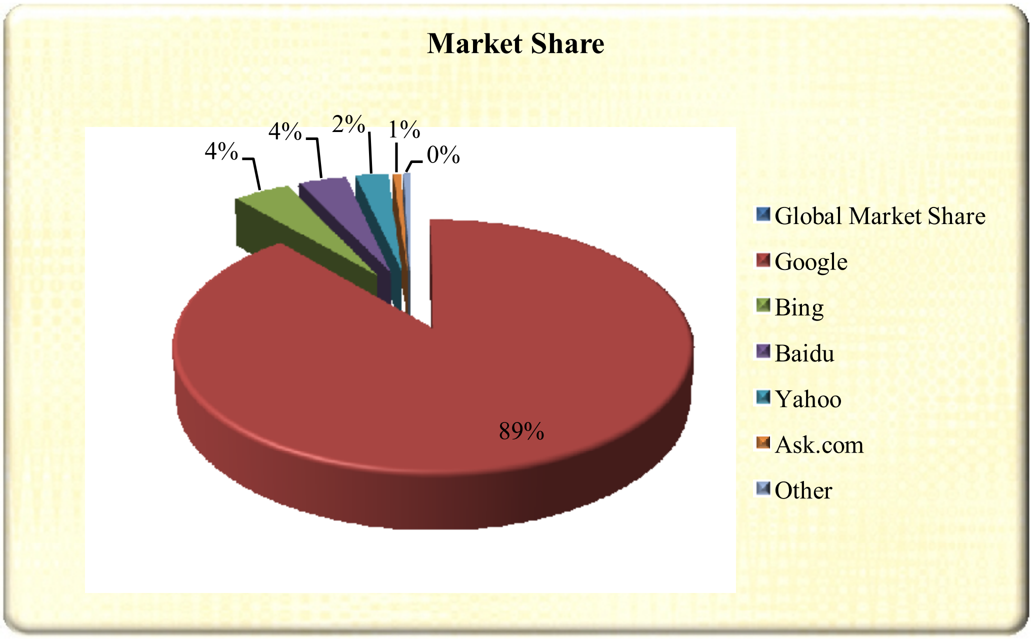

Figure 2.3: Search Engine Market Share (April, 2013)

Source [Nielsen-Net]

Best Search Engine?	
Google	76.58%
Yahoo!	16.30%
Bing (MSN/Live)	4.50%
Ask	1.11%
AOL	1.51%
Total Votes = 36606	

Table 2.4: Best Search Engine Survey

Source: www.webdevelopsnotes.com

Top 15 Search Engines

S. No.	Search Engine
1	Google
2	Bing
3	Yahoo! Search
4	Ask
5	AOL Search
6	MyWebSearch
7	Blekko
8	Lycos
9	Dogpile
10	WebCrawler
11	Info
12	InfoSpace
13	Search
14	Excite
15	Goodsearch

Table 2.5: Top 15 Search Engines

In china, Baidu is the most used search engine but not wide spread use in other parts of the world. Although, it is placed in the forth position in survey by Nielsen-Net, we will not consider it for our research purpose due to Chinese version of the Baidu search engine.

With 13,304 participants, a survey was conducted by Webdevelopersnotes.com in September 2010. The most of participants ranked Google at first position and best search engine while only around 25% of the participants selected other search engines.

Google

In 1998, Larry and Sergey developed the search engine known as “Google,” named after the term “googol,” mathematically, this term means 1^{100} terms. In 1998, the private company named Google Inc. was established. The mission of the Google was “to organize the world’s information and make it universally accessible and useful”. Google is a spider based search engine and is considered to be the vastly accepted search engine. Various search features, additional services and tools such as Google Maps, Google Earth were developed by Google known as Google products.

Yahoo!

In 1994 by Jerry Yang and David Filo developed a search engine known as Yahoo!, which is an abbreviation for "Yet Another Hierarchical Officious Oracle”. The vision of the Yahoo! was “to be the center of people's online lives by delivering personally relevant, meaningful Internet experiences” [187].

Although, Yahoo was started with crawler based search engine which was powered by Google but since 2004, Yahoo adopted its own technology with indexing and ranking algorithms [37]. It is one of the oldest directories in the world.

Today, Yahoo has both types of search engine index; a crawler based index (Yahoo! Search) as well as human powered directory (Yahoo! Directory) with sites in subject categories evaluated by editors from Yahoo.

Bing

Microsoft designed and developed its own search engine, known as Windows Live Search, MSN Search and Live Search. Later on in 2009, it was named as Bing. Bing is a crawler based search engine. The company PowerSet was acquired by the Microsoft which allows slight improvement in search technologies. It has also put great effort in visual presentations with pictures.

Bing, Yahoo and Google: At a Glance

Let us compare three major search engines, Bing, Yahoo and Google, after brief introduction of these search engines.

There are so many factors on the basis of which the search engines can be compared but the major distinction among these search engines lies in the following criteria:

1. Size of the Database
2. The Actuality
3. The Capability
4. The Technology

These characteristics are discussed in the next section.

1. Database Size

Although, it is not known the exact size of the web but some estimation are done on the size of the database of search engine. In order to estimate the rough size of search engine database, the very common word which can be found in almost every database such as the word “the” can be found in almost every page. Approximately 84 billion pages were found by Google, Yahoo found around 60 billion and Bing found around 6 billion results as on February 11th, 2012 that contained the word “the”. These numbers are only representative of the database not the exact size of the search engine database.

2. Actuality

In order to keep up to date of the Web pages, the crawlers need to crawl the web repeatedly and periodically. The “Crawling is a never ending process”. How well do the major search engines control the up-to-dateness of the database?

Lewandowsky [52] in a research tested the frequency in which the indices are updated by Google, Yahoo and MSN. During 42 days observation, it was observed that four different groups with nine or ten websites updated every day to get whether these search engines are capable to index contents on a regular basis. Google’s web crawler, known as GoogleBot updates several sites daily. The re-crawling time usually depends on the link popularity and on the frequency how often the web page changes. On the other hand, Bing updates its index with MSNbot regularly, while Yahoo updates with its crawler Yahoo Slurp [57].

3. Capabilities

Most of the search engines provide both basic and advance search features in finding the search results, although used rarely by the user. The search engines work on many operators to search engine’s query language.

The Boolean operators such as AND, OR and NOT are used. The operator AND is used to join words, for example, “New Delhi” = “New AND Delhi”). Other operators OR and NOT are commonly used for corresponding terms.

For filling in the blanks (*), wildcards can be used. They replace any unknown whole word like a placeholder for terms.

Quotation marks (“”) permit the user to check for an exact phrase in a web document. Stop words such as “a” or “the” are discarded unless they are placed within quotation marks.

The minus (-) and plus (+) signs can be used to remove or include words.

The main operators used in the Bing, Yahoo and Google are summarized in the table 2.6.

4. Technology

Two aspect of the technology such speed and ranking of the relevant results are taken into consideration while discussing the technology used by the search engines.

The findings [179] show that search engines has fewer number of searches if longer is the search time. This proves that speed is one of the important parameter to be considered by the search engine and to the users as well.

Google announced a new search technology in September 2010, the “Google Instant” which outperforms the other search engines. This feature gives the user the enhancement over other search engines and opportunity to “search-before-you-type”. The results are displayed and recommended to the user on search engine result page while typing in the search box without even typing the entire search term or pressing “Return”. It is not only faster but shows and predicts more number of results than earlier.

Comparison among Bing, Yahoo and Google

S. No.	Parameters	Google	Yahoo	Bing
1	**Database**			
	Rank	First	Second	Third
	Index Size	> 45 Billion Pages	> 10 Billion Pages	> 9 Billion Pages
	Crawler Name	GoogleBot	Yahoo Slurp	MSNBot
	Updations	Fastest, Many daily updates	No clear frequency	Frequent Updates
2	**Capability**			
	Search Operators	- to remove	- to remove	- to remove
		+ to include	+ to include	+ to include
		" " for exact words	" " for exact words	" " for exact words
		Wildcards	Wildcards	
		Stemming	Stemming	Stemming
		Meta word search		
		Advanced Search	Advanced Search	Advanced Search
	Languages	46 languages	32 languages	42 languages
3	**Technology**			
	Speed	Shown for every query		
		Google Instant "search-before-you-type	Not exactly known	Not exactly known
	Ranking	PageRank	Keywords	Keywords
		Hypertext matching analysis	Click popularity	Inbound Links

Table 2.6: Comparison among Google, Yahoo and Bing

This chapter gives overview of the three major search engines namely, Bing, Yahoo and Google. It also gives valid reasons why we selected these three search engines for research work.