A

Seminar report

On

Search Engines

Submitted in partial fulfillment of the requirement for the award of degree
Of CSE

SUBMITTED TO:
www.studymafia.org

SUBMITTED BY:
www.studymafia.org

Preface

I have made this report file on the topic **Search Engines**; I have tried my best to elucidate all the relevant detail to the topic to be included in the report. While in the beginning I have tried to give a general view about this topic.

My efforts and wholehearted co-corporation of each and everyone has ended on a successful note. I express my sincere gratitude to ……..who assisting me throughout the preparation of this topic. I thank him for providing me the reinforcement, confidence and most importantly the track for the topic whenever I needed it.

Acknowledgement

I would like to thank respected Mr…….. and Mr. ……..for giving me such a wonderful opportunity to expand my knowledge for my own branch and giving me guidelines to present a seminar report. It helped me a lot to realize of what we study for.

Secondly, I would like to thank my parents who patiently helped me as i went through my work and helped to modify and eliminate some of the irrelevant or un-necessary stuffs.

Thirdly, I would like to thank my friends who helped me to make my work more organized and well-stacked till the end.

Next, I would thank Microsoft for developing such a wonderful tool like MS Word. It helped my work a lot to remain error-free.

Last but clearly not the least, I would thank The Almighty for giving me strength to complete my report on time.

Content

Introduction

Searching is one of the most used actions on the Internet. Search engines as an instrument of searching, are very popular and frequently used sites. This is the reason why webmasters and every ordinary user on the Internet, must have good knowledge about search engines and searching.

Webmasters use major search engines for submitting their sites on it, and for searching. Ordinary users use major search engines primarily for searching, and sometimes for submitting their homepages or small sites.

Why search engines are so popular? If you are ordinary user and you want to find some texts or pictures about certain theme, first thing you will do is to visit search engine to get some good URLs about certain theme. This will happen every time you need some information or data about any theme.

If you are webmaster, you will also need some information while preparing site for the Web, and you will also use search engines. Then, when you finish with it, you must submit your URL to many search engines. After that you will check your URL ranking on every search engine... There are also hot news on every major search engine, many other interesting contents... All of this shortly describes why search engines are so popular.

As a user at novice level you must learn how to use search engines for searching the Internet. You must know that there are two ways of searching: by using user's query or by using categories. If you have keywords or phrase that best describes the theme you need, you should use user's query. But if you need some theme, and you don't have keywords or phrase, you should use categories.

If you use user's query, you should type keyword or phrase in this form and click on "search". Then you will get search results, and you can choose URL, which is the best in your opinion. If you use categories, you should click on category that best describes the theme you need. You will then get subcategories and should choose some subcategory that best describes the theme you are after. Repeat this action until you find group of URLs, which content is related with theme you want.

As a webmaster you must submit your URL to all major search engines. This is the way to promote your site. You could get many visits from major search engines, if you have a good ranking of your URL. We made page with URLs which takes you to submit forms of major search engines. You will not lose your valuable time on searching for these forms, all of them are on one page.

The History Of Search Engines

Internet technology has been a revolutionary one; there is simply no second question about this fact. Different internet tools and sources have mesmerized the world to a great extent. Today people can hardly imagine of a life without internet technology. Find a sector where internet tools and resources do not find an application and you get a million dollars for this task! Literally, there is no such field or sector devoid of these applications.

And when it comes to internet, it is not just about the websites it has. There are a variety of tools that help you to browse, find and incorporate information of your choice on different platforms. And in this regard, the use of search engines cannot be masked with any argument.

Search engines first emerged in the midst of 1990's when Google got fame as the first standardized and facilitating search engine of the entire world. Though there were some search engines working prior to Google but none of them was as user friendly and easy to use as was this one. And taking a look into Google, it was actually an academic project of a couple of university students. Later on, this idea was expanded and new technological features were incorporated to make it a better one.

And today this search engine stands out as one of the most widely used website all around the world. The owner of this search engines are literally making millions and billions of dollars each year. Following this search engine, many other search engines have also been brought to light by different software designers but none of them have yet been able to replace Google by any mean. But all these search engines are serving people a great deal and it is hoped that more improvements will be brought to these search engines in near future.

Different Types of Search Engines

When people mention the term "search engine", it is often used generically to describe both crawler-based search engines and human-powered directories. In fact, these two types of search engines gather their listings in radically different ways and therefore are inherently different.

Crawler-based search engines, such as Google, AllTheWeb and AltaVista, create their listings automatically by using a piece of software to “crawl” or “spider” the web and then index what it finds to build the search base. Web page changes can be dynamically caught by crawler-based search engines and will affect how these web pages get listed in the search results.

Crawler-based search engines are good when you have a specific search topic in mind and can be very efficient in finding relevant information in this situation. However, when the search topic is general, crawler-base search engines may return hundreds of thousands of irrelevant responses to simple search requests, including lengthy documents in which your keyword appears only once.

Human-powered directories, such as the Yahoo directory, Open Directory and LookSmart, depend on human editors to create their listings. Typically, webmasters submit a short description to the directory for their websites, or editors write one for the sites they review, and these manually edited descriptions will form the search base. Therefore, changes made to individual web pages will have no effect on how these pages get listed in the search results.

Human-powered directories are good when you are interested in a general topic of search. In this situation, a directory can guide and help you narrow your search and get refined results. Therefore, search results found in a human-powered directory are usually more relevant to the search topic and more accurate. However, this is not an efficient way to find information when a specific search topic is in mind.

Table 1 summarizes the different types of the major search engines.

Search Engines	Types
Google	Crawler-based search engine
AllTheWeb	Crawler-based search engine
Teoma	Crawler-based search engine
Inktomi	Crawler-based search engine
AltaVista	Crawler-based search engine
LookSmart	Human-Powered Directory
Open Directory	Human-Powered Directory
Yahoo	Human-Powered Directory, also provide

	crawler-based search results powered by Google
MSN Search	Human-Powered Directory powered by LookSmart, also provide crawler-based search results powered by Inktomi
AOL Search	Provide crawler-based search results powered by Google
AskJeeves	Provide crawler-based search results powered by Teoma
HotBot	Provide crawler-based search results powered by AllTheWeb, Google, Inktomi and Teoma, “4-in-1” search engine
Lycos	Provide crawler-based search results powered by AllTheWeb
Netscape Search	Provide crawler-based search results powered by Google

Table 1: Different types of the major search engines

From the table above we can see that some search engines like Yahoo and MSN Search provide both crawler-based results and human-powered listings, therefore become hybrid search engines. A hybrid search engine will still favor one type of listings over another as its type of main results.

There is another type of search engines that is called meta-search engines.

Meta-search engines, such as Dogpile, Mamma, and Metacrawler, transmit user-supplied keywords simultaneously to several individual search engines to actually carry out the search. Search results returned from all the search engines can be integrated, duplicates can be eliminated and additional features such as clustering by subjects within the search results can be implemented by meta-search engines.

Meta-search engines are good for saving time by searching only in one place and sparing the need to use and learn several separate search engines. "But since meta-search engines do not allow for input of many search variables, their best use is to find hits on obscure items or to see if something can be found using the Internet."

How Search Engines Work

For those of us in SEO (or aspiring to be), there are a lot of little details that fill our days. Server architecture, 301 redirects, 404 errors, title tags, and various other things.

Sometimes, we forget to sit back and figure out what it all means. Add to that the fact that most SEOs were never trained, but just picked things up "on the job," and it's no surprise that most SEOs don't really know how search engines work.

When's the last time you sat down and considered how search engines (like Google) really work? For me, it was last month, while writing the post about a recent Google Webmaster Hangout and the information about link disavowal that came out of it.

But before that, I think it honestly had been 8 or 10 years since I'd really thought about it. So let's fix that. Here is a high level explanation of how one search engine (Google) works. While the terminology and order of operations may change slightly, Bing and Yahoo use a similar protocol.

Crawling Vs. Indexing

What does it mean when we say Google has "indexed" a site? For SEOs, we use that colloquially, to mean that we see the site in a [site:www.site.com] search on Google. This shows the pages in Google's database that have been *added to the database* – but technically, they are not necessarily *crawled*, which is why you can see this from time to time:

A description for this result is not available because of this site's robots.txt – learn more.

Indexing is something entirely different. If you want to simplify it, think of it this way: URLs have to be discovered before they can be crawled, and they have to be crawled before they can be "indexed" or more accurately, have some of the words in them associated with the words in Google's index.

My new friend, Enrico Altavilla, described it this way, and I don't think I can do any better than he did, so I'm giving it to you word-for-word:

An (inverted) index doesn't contain documents but a list of words or phrases and, for each of them, a reference to all the documents that are related to that word or phrase.

We colloquially say "the document has been indexed" but that really means "some of the words related to the document now point to the document." Documents, in their raw format, are archived elsewhere.

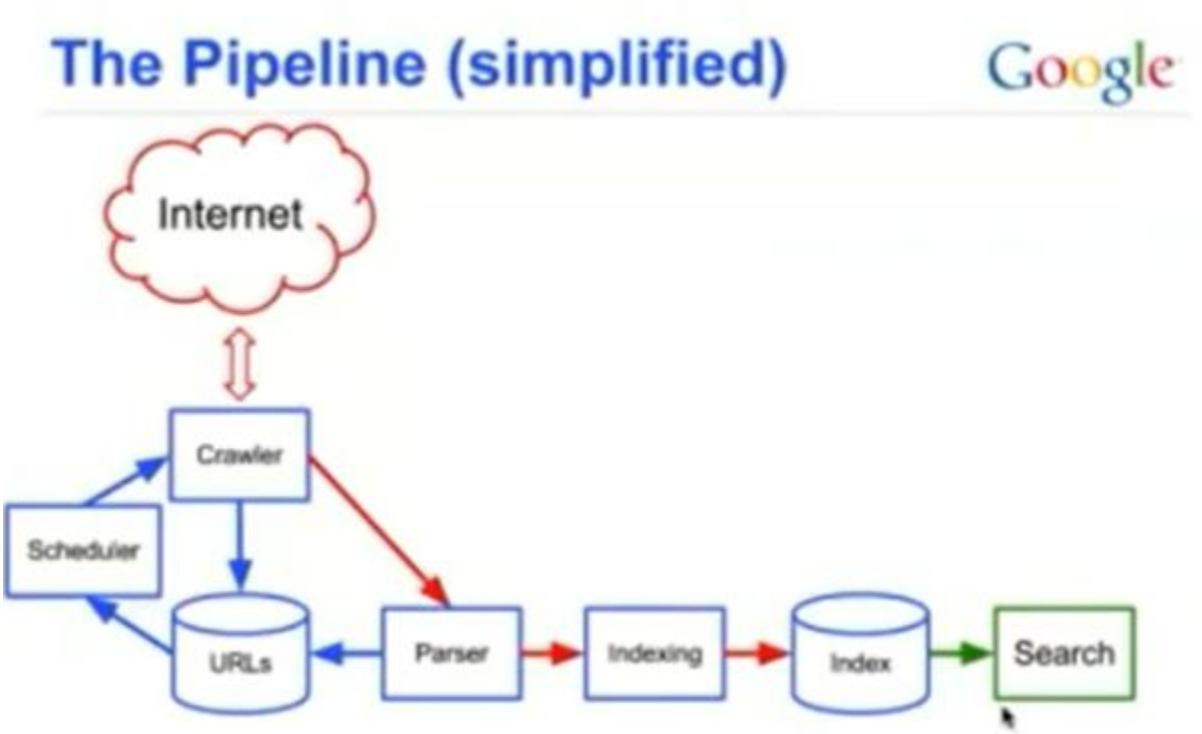
My old friend and former Googler, Vanessa Fox, had this to say on the subject:

Google learns about URLs… and then adds those URLs to its crawl scheduling system. It dedupes the list and then rearranges the list of URLs in priority order and crawls in that order.

The priority is based on all kinds of factors… Once a page is crawled, Google then goes through another algorithmic process to determine whether to store the page in their index.

What this means is that Google doesn't crawl every page they know about and doesn't index every page they crawl.

Below is a simplified version of the pipeline that was shared by Google:



A couple of other important things to note:

• Robots.txt will only block a page from being crawled. That's why Google sometimes has pages in its search results like the example above. Because, although Google was able to associate the page with words based on things like internal links, it wasn't able to actually crawl the content of the page.

• Noindex commands at the page level are not definitive. Although Google can crawl the page and associate words on the page with the index, it is not *supposed* to include that page in search results.

However, I have seen cases where Google has included a noindexed page in their publicly available records, and Google has said it may disregard the command if other signals indicate strongly enough that the page should be indexed. This is one important area where Google

differs from the rest. Yahoo and Bing will respect your noindex commands and they will not index the page or include it in search results.

One other important thing to note is that canonicals, parameter exclusion, and various other elements are also processed at some point between when Google learns about the page and when it crawls and/or indexes it.

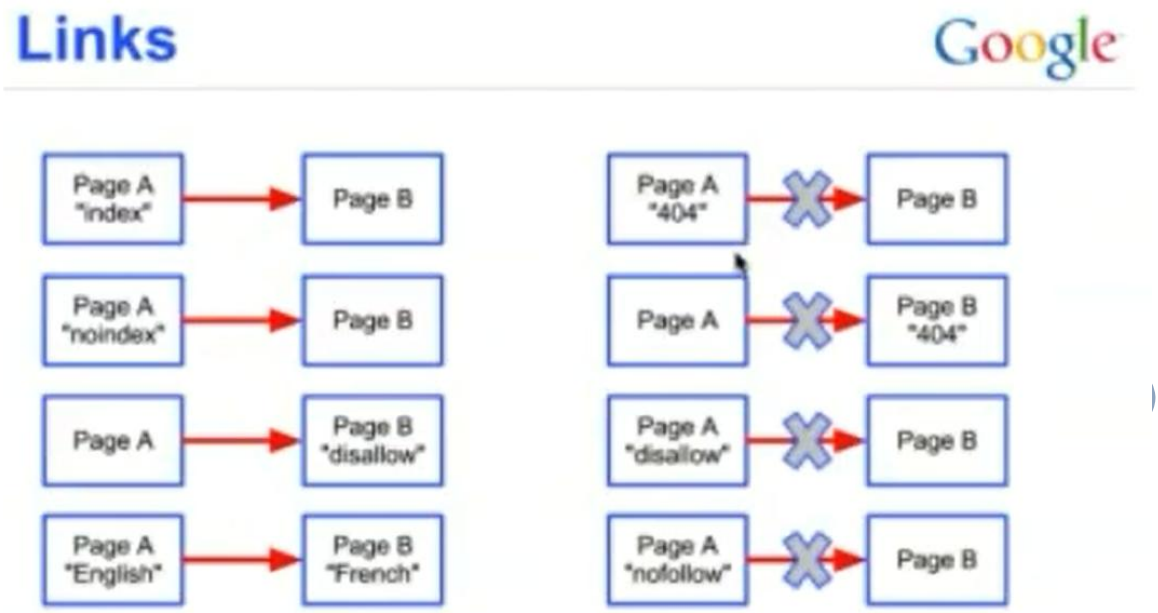
Links And The Link Graph

The next thing SEOs need to understand are links and how they are processed. The most important thing to learn from this is that links (and, by extension, PageRank) are not processed during the crawl event. In other words, Google does the crawling as indicated above, but PageRank is not considered during the crawl — it's done separately.

What does this mean?

- PageRank, despite what many may say, is a measure of the quantity and quality of links. It has no connection to the words on a page.
- Many SEOs believe that there are two elements of PageRank: a domain-level and a page-level PageRank. The belief is that the domain-level PageRank is the one that determines domain authority, a factor many believe is used in deciding how to rank sites. While I believe that Google likely uses some element of domain authority, this has never been confirmed by Google.
- Because PageRank is processed separately from the crawl, directives like "noindex," "disallow," and referrer-based blocking do not work to stop the flow of PageRank from one page to another.
- You can't control PageRank with any kind of referrer-based tracking. In other words, you can't block a referrer in .htaccess (for example) and expect it to work on Googlebot like a nofollow.
- Contrary to popular belief, a 302 redirect WILL pass PageRank.
- The only four things that work to stop the flow of PageRank are:
    1. A nofollow directive on the link at its source
    2. A disallow directive in the robots.txt on the page where the link originates. This works because the robots.txt command keeps the search engine from crawling the content of that page; therefore it never sees the link.
    3. A 404 error on the originating page.
    4. A 404 error on the destination page. The only reason 404s work is that both of these directives occur on the server side. When the link graph is processed, those pages are again inaccessible to the search engine. By contrast, a page blocked in robots.txt can continue to accrue PageRank, because it does exist; it just hasn't been added to the index.

Here is a screenshot of a slide shared by Google in a Webmaster Hangout on August 20, 2012 that describes this:



The only other way to handle bad links is to disavow the link source. This has the same technical impact as adding a “nofollow” to the source link, if Google accepts it.

The Importance of Search Engines

I once read that the average person living in a modern industrialized society is exposed to as many different pieces of information in a single day as a person living 100 years ago would have seen in a year. That includes advertisements, newspaper headlines, websites, text messages, traffic signs, T-shirt slogans, and on and on and on. It's hardly surprising that attention spans are getting shorter and that the majority of people believe themselves to be busier than ever.

With this information overload, it is next to impossible to remember everything we need to, to call up names, dates, figures, phone numbers, email addresses and all the corporate and client information we need to do business effectively. That's why we use tools to do the remembering and information retrieval for us. My company uses Salesforce.com to handle the bulk of our customer relationship management information. I use Microsoft Outlook to manage my email. When I want to find a product, service or piece of information online, I use a Search Engine.

I'm not alone in using Search Engines. Far from it. In the month of March 2006 alone, there were 6.4 billion searches. Assuming each user looks at an average of two search results pages, each of which displays 10 search results, that gives an average of 128 billion search results shown to Internet users in a single month. Search Engines are ubiquitous, and so accepted in contemporary culture that the word "Google" now appears in the dictionary as verb (as in "to Google something").

Search Engines essentially act as filters for the wealth of information available on the Internet. They allow users to quickly and easily find information that is of genuine interest or value to them, without the need to wade through numerous irrelevant web pages. There is a lot of filtering to do - three years ago in 2004 the number of pages in Google's index exceeded the number of people of the planet, reaching the staggering figure of over 8 billion. With that much content out there, the Internet would be essentially unworkable without the Search Engines, with Internet users drowning in sea of irrelevant information and shrill marketing messages.

The goal of the Search Engines is to provide users with search results that lead to relevant information on high-quality websites. The operative word here is "relevant". To attain and retain market share in online searches, Search Engines need to make sure they deliver results that are relevant to what their users search for. They do this by maintaining databases of web pages, which they develop by using automated programs known as "spiders" or "robots" to collect information. The Search Engines use complex algorithms to assess websites and web pages and assign them a ranking for relevant search phrases. These algorithms are jealously guarded and frequently updated. Google looks at over 200 different metrics when assessing websites, including copy, in-bound links, website usability and information architecture.

What this means is that the Search Engines provide users with the information they are looking for, and not necessarily the information that marketers would like them to see. Type the name of a major brand into Google, and you will most probably be served a wide range of search results that include not only the official website of the brand you searched for, but also other websites, consumer review sites, Blogs, online articles on Web 2.0 sites and press releases on news syndication channels. Of course, not all searches are for brand names. The majority of searches

are for non-brand keyphrases - for example, "Hong Kong luxury hotel" rather than "The Peninsula Hong Kong". With keyphrases that are service or product-specific rather than brand-specific, results pages will also include many competitors, which makes acquiring a prominent position at the top of the page even more crucial.

There are two major ways to make sure a website appears in a prominent location on the major Search Engines for relevant keyphrases: Paid Search (also known as Pay-Per-Click) and Organic Search Engine Optimization. Of the two, Organic Search Engine Optimization tends to yield the best long-term results and the optimum return on investment, for the simple reason that Internet users are four times as likely to click an Organic search result as they are a Pay-Per-Click ad on the same results page.

In a September 2006 poll by MarketingSherpa, 68.7% of marketers in the US identified Search Engine Optimization as yielding the best Return on Investment for product marketing. I will discuss Paid and Organic search in much more depth in a separate article. It is enough here to state that companies doing business or marketing online should look at striking a healthy balance of both techniques to make the most of the potential of marketing through the major Search Engines.

Search Engines matter because they increasingly determine the information about brands, products and services that customers access online. Being easy to find on Google, Yahoo and MSN is now as much of a marketing necessity as having a strong presence in print and broadcast media, or an effective traditional direct marketing program. And as consumers and organizations come to rely more heavily on them to find the goods, services and suppliers they need, the importance of the Search Engines to modern businesses can only increase.

Challenges faced by Search Engines

- The web is growing much faster than any present-technology search engine can possibly index (see distributed web crawling).
- Many web pages are updated frequently, which forces the search engine to revisit them periodically.
- The queries one can make are currently limited to searching for key words, which may result in many false positives.
- Dynamically generated sites may be slow or difficult to index, or may result in excessive results from a single site.
- Many dynamically generated sites are not indexable by search engines; this phenomenon is known as the invisible web.
- Some search engines do not order the results by relevance, but rather according to how much money the sites have paid them.
- Some sites use tricks to manipulate the search engine to display them as the first result returned for some keywords. This can lead to some search results being polluted, with more relevant links being pushed down in the result list.

Conclusion

The usefulness of a search engine depends on the relevance of the results it gives back. 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. The methods also change over time as Internet usage changes and new techniques evolve.

References

- www.google.com
- www.wikipedia.com
- www.studymafia.org