Searching the web

Richard Hardstone
rhardstone@gmail.com

Manuel Lorenzo Parejo
manulorenzop@gmail.com

Lieven van Velthoven
lievenvv@gmail.com

ABSTRACT

Search-engines are an essential technology for finding websites on the internet. From 1990 there has been a wide range of techniques employed to do this. Most of these consist of three elements, the web crawler, indexer and query language. There are problems however with search engines including privacy concerns and spam. Search-engines have also been used in unintended ways such as Google whacking where people use a search-engine as a game instead of to find sites.

PURPOSE, CONTEXT AND HISTORY

The internet has grown at an astonishing rate, and will continue growing in the future [0]. One of the most important developments that propelled Internet use to its mainstream status has been the introduction of 'search engines'.

A web search engine usually consists of a page where users can enter keywords to specify what information they want to find. The results are then presented as a (long) list of 'hits': links to hopefully relevant webpages. Finding these hits turns out not to be a trivial task.

The main problem, besides its sheer size, is the Internet's distributed nature. There is no central administration of the pages on the Web: anyone can add pages anywhere, anytime, and finding all these pages is quite a task. E.g., when looking for files on a hard disk, you can systematically and exhaustively search through all the files. Due to the free, interconnected structure of the Web it is not nearly as straight forward to search through and even then, the problem remains of determining which of the millions of possible results is most relevant to the user.

In this text we will start by giving a short overview of the history of web search, then explain the principles of web crawling and indexing, and highlight some unintended or unforeseen uses.

History

The history of web search engines has been full of apparitions of new search engines that lasted for a short time, big companies taking over small to mid sized companies web search engines and deals between big companies.

Here is a brief history of the main milestones in the web search engines history.

- 1990. Alan Emtage creates the very first tool for searching on the internet called *Archie.* It downloaded the directory listings of all the files located on public anonymous FTP sites, creating a searchable database of filenames.
- 1993. Due to the rise of Gopher –a TCP/IP Application layer protocol designed for the retrieval and transmission of documents over the internet- *Veronica* and *Jughead* are created. The first provided a keyword search of most Gopher menu titles in its listings, while the latter obtained the menu information from different Gopher servers.
- 1993. Oscar Nierstrasz wrote several Perl scripts that would periodically mirror the pages in the different specialized catalogues –maintained by hand- hosted all over the internet and rewrite them into a standard format and creating thus the W3Catalog, the very first primitive search engine.
- June 1993. Matthew Gray creates the first web robot, the World Wide Web Wandered with the purpose of measuring the size of the internet –a task that took two years to be fully accomplished.
- December 1993. *JumpStation* becomes the first web search engine to make use of the three essential features –crawling, indexing and searching- by means of using a web robot.
- Late 1994. WebCrawler becomes the first full-text crawler-based search engine that let it users to search for any word contained in a webpage.
- 1994-2000. Several search engines appear – Magellan, Excite, Infoseek, Inktomi, Northern Light, Altavista, Yahoo!- and remains for the next years on top of the users' preferences, in spite of the inability of a search based on full-text copies, but rather using a search function operated on web directories.
- 2000. Google raised its popularity due to the invention of the PageRank, an algorithm that ranks the web pages based on the number and PageRank of other web sites and pages that link to them, based on the principle that good pages are the ones that are linked more. Google was also the first web search engine that used a minimalist page, in contrast with the rest, that used a search engine embedded in a web portal.
- 2002. Yahoo! Was providing web search services based on Inktomi's technology, until the first company took over the second's product.

- 2004. Yahoo! switches to Google's web search engine until 2004 when they launch their own based on the combined technologies and its acquisitions.
- 2009. Microsoft launches Bing!, a web search engines based on their own technology –MSN Search, the previous Microsoft search engine used Inktomi's technology- and their own web crawler called *msnbot.*

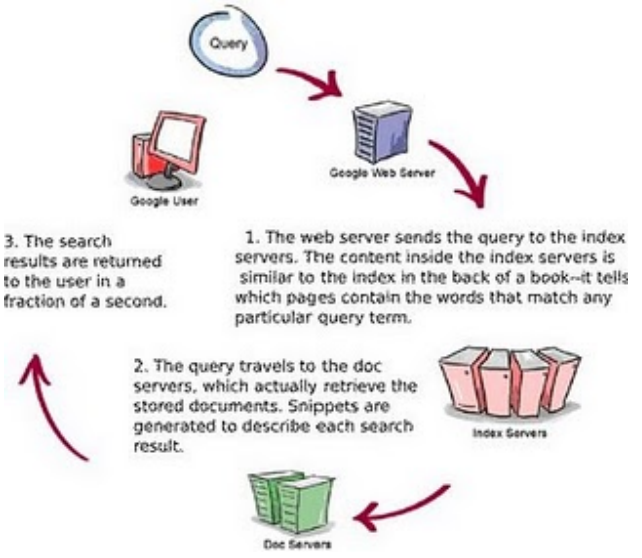
OPERATING PRINCIPLES

Web search engines can be dissected in three essential phases or steps [1,2]:

Web crawling

A web crawler or *spider* [3] is an automated browser that explores the different webpages on the internet. Usually, the starting point for a crawler to start its travel over the networks are lists of heavily used servers or very popular pages. Then, it will index the words on the different pages that are being visited and following every link that is being found on the webpage and will build indexes of the found words based on its own system of weighting, then saving the data and storing it in a database.

Web crawlers or *spiders* are automated software whose mission is to search the web looking for new pages in order to index them and make them available to the user. They are given a URL to start their search with or just a list of very popular pages. For every page they visit, the robots extract the words and record the different URL to create a table that associates words to the URL of the pages where they have been found.



Web crawling. Source of the image: codeglobe.blogspot.com

Normally the web search engines do not take very long in order to crawl a site, as they have thousands of crawlers indexing the whole internet and these do not use a huge data flow, but only process words and URLs, making the task of web crawling practically a real-time task.

The crawlers pay attention to all the words found in a web page, but they assign them different preferences or weights depending on in which part of the webpage the words are written, having special consideration with those words appearing in the title, subtitle of *meta-tags* –words used by the owner of a webpage to describe the content of its website and that are taking in special consideration by the web crawlers. For instance, Google's crawler leaves out articles such as "a", "an", etc, which allows it to operate faster and more reliably, while others such as AltaVista indexed every single word on a page, no matter their origin and semantic categorization.

Indexing

Web crawlers are constantly search for new pages and information on internet in order to offer the user the possibility of seeing in almost real-time the maximum amount of websites.

While the *spiders* are crawling the internet, relating the words and the URLs where they have been found, they need to save the information in a certain place so that the computers and *mainframes* that compose the web search engine can build the index that is where the engine will finally look for results to a user query.

There are two main concepts that need to be properly used in order to achieve the best reliability and efficiency, that are the information store and the way this information is indexed.

Taking in consideration that the process of crawling internet and creating indexes requires massive amounts of data storage space, the amount of information stored needs to be reduced to the minimum possible.

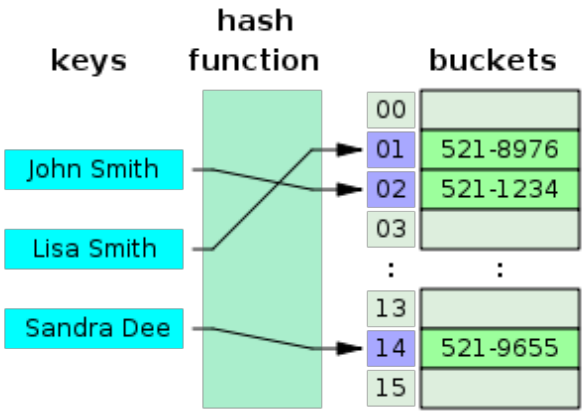
How do web search engines perform this? They need to categorize the words they find in the webpages depending on the different places they are found –title, subtitle of the web site, meta-tags, body of the web page- , whether they are written using capital letters –although this feature can vary from different web search engines-, etc. This is normally done by employing a chain of bytes that each one symbolizes the state in which the features described above are present in each of the words. This way, the information corresponding to a certain word and its characteristics can be saved using only several bits, thus saving an enormous storage space.

After being stored, the data needs to be indexed in order for the information to be found, accessed and retrieved as soon as possible.

One of the most common methods for the indexing is using a *hash table [4]*, which consists of a formula to each word (called *keys* in the hash table) in order to give each one of them a numerical value (*values*), distributing the different entries in an equal way, so that the cost of accessing each element of the hash table –and thus each entry in the database- is minimum.

For example, the most common word in English is the

personal pronoun *I*, while seldom used words could be *bulwark* or *glede*, which will certainly appear less often in webpages than the first one, and this is why hashing tables spread the values along a certain number of divisions that the hash table has, so they do not have a huge number of words –*values*- in a certain *bucket* while some others are empty.



Hash table. Source of the image: Wikipedia

Searching

Web search queries are what a user types into the web search engine in order to get a set of results, and normally these queries are unstructured and ambiguous, as well as the fact they vary from query languages, which is the language that the web search engine will use and that is strictly regulated by syntax rules.

In order to link the different words typed in a web search query by the user, Boolean operators can be used to refine the results thrown by the web search engine, and becoming thus a *structured query*. These Boolean operators are massively used in databases.

Normally –although it always depends on the nature of the query that the user types in the search box of the web search engine- the user will be able to easily find what he/she is looking for, but there are certain aspects to have in mind when searching for more complicated information when using more complex queries

There are some tips that can help the user improve the results set, by refining the search query. As search queries are based on structured query language, there are certain Boolean operators [5] that can be used. These operators include:

- AND: will show a search result containing two terms, e.g.: cars AND trucks.
- OR: will show only ONE of the two terms. For the query cars OR trucks will display either one or another, but not both of them in the same result.
- NOT: will show the results corresponding to the first term, but not to the second, e.g.: for cars NOT trucks will always show cars, but never trucks.

Although they are not Boolean operators, there are certain operators that will make easier the task of search for certain information. According to Google Search Basics [6], some of the operators are the following:

- Phrase search (“”): It is a literal search that tells the web search engine that the query we entered between quotations has to be literally searched, including the order of the words.
- Terms to exclude (-): With this operator we specify a certain word that we do not want to have included in the results page of the query.
- Search in an specific site (site:): With this command we let the engine know that we want the query to be search in a specific website.

It is always possible to combine the different Boolean operators in order to refine the search even more. For example in order to look for cars and tracks made by the brand Ford, we could use the following statement when typing the query into the web search engine textbox:

cars OR trucks AND ford

STRENGTHS AND WEAKNESSES

The strengths of internet search engines are well-known to almost anyone nowadays: in just a fraction of a second you can retrieve an immense number of links to information from all over the world.

The amount of information on the web is both a bless and a curse when searching. The sheer amount guarantees that something can be found on just about any topic. On the other hand it can also lead to an overload of results, making it hard to find actual relevant ones.

Web search engines are simply webpages that help the user to find the information this may be looking for, and as such are just intermediate products between the user and what they are looking for. However, there may be certain weak or *obscure* points, especially when big companies such as Google and Yahoo! are the owners of the biggest and most important web search engines.

Privacy

Although Google is the most used search engine, it is also the most criticized, as there are some obscure points that may appear as disadvantages when using this web search engine, mostly concerning the privacy.

When using any web search engine and introducing a query to search on the internet, the user is actually typing relevant information, be it about the user itself or not, but this information is stored in the servers –presumably, to *improve the users' searching experience by refining the result set*, according to Google's executives- and this, along with the fact that Google is known by keep a *cookie* –this is a small piece of text stored in our hard disk drive that identifies uniquely a computer by assigning it an ID- gives the company the opportunity to associate the queries the user enters in the search engine with a unique user, which

automatically shows the lack of anonymity that a search engine can offer. These cookies were set by Google to last for 32 years, until the company was severely criticized for this and they afterwards changed this expire time to 2 years –although every time a computer uses a certain Google service, this cookie is automatically renovated.

Although it is not a web search engine but an e-mail service provided by Google, every time a user writes an e-mail, Google automatically –allegedly, as stated in the Google's Terms of Service, non-human will ever read the contents of a message except the account holder itself- scans the e-mail message and provides the distinctive and important words to advertiser companies so they can offer the user their products. This way it is very common to see that if the user is writing an e-mail containing the words *car* and *buy*, several advertisements related to new and used car companies.

However, Google is not the only company to scan the users' e-mails to look for relevant information, as Yahoo! and Microsoft Hotmail also do such a thing.

The search engines can also be used to advertise certain products, as the case of Google's AdWords service, which shows some *featured* results at the very top of the results page that correspond to advertisers that have paid a certain amount of money to be announced on the users' results page.

Usability and reliability

Concerning the user experience and the usability of the web search engines, there are several differences. As previously stated, a web search engines is merely a way to look up information on the internet, but they are not perfect, and even if the user types a very concrete query in order to get the information, this might have to skim through the results page in order to find the information that the user is looking for. The web search requires experience and thus has a learning curve, which may also make the beginners to give up on web search engines –although on the long run, when the user is fairly experienced, the search engines become an indispensable part of the internet experience for them.

Google's PageRank, although it is certainly one of the best algorithms used by web search engines in term of reliability and efficiency, also has some weaknesses. PageRank measures the importance of a certain webpage judging the amount of pages where it has been linked to and the importance of these pages themselves in the PageRank system. As the reader may foresee, this system can lead to a discrimination of new pages and to a favoring of the already established sites, even though if the information in the new page is more accurate for users' query than the popular page.

Yahoo! is also known because of using a paid inclusion program for the search engine, in which commercial websites were guaranteed to be included in the listing that the Yahoo! search engine uses –although a certain position in the results page was not guaranteed after October 2006, as it proved to be very unpopular amongst both the advertisers, who were reluctant to pay, and the ordinary users, who were unhappy because of the paid inclusion of commercial websites in an indistinguishable way in their results page.

Security

Both Yahoo! and Google have been exposed to several bugs which would make the users alter the results of the search engine, especially Google, as it is the web search engine with the biggest market quota [7].

Yahoo! has been accused of providing ads via the Yahoo! search network to the users of it, causing them to experience abnormal behavior in their computers, such as pop-ups.

Until they corrected the algorithm in 2007, Google suffered several Google Bombing attacks. These attacks consisted not in exploiting a vulnerability but a *smart use* of a black point in the PageRank algorithm; the Google page of results can be altered and ranking a page higher if enough other sites are linked to that page using an *anchor text* –a simple internet link-, which made this form of attack very popular. For example there is the case of thousands of pages linking to the White House page by attaching a hyperlink to the text "Miserable failure", or another very famous case was a campaign made by the internet users in Spain referring to the SGAE –the Spanish equivalent to the American DMCA- which was appearing as the top result in Google when writing "thieves".

Spam

One of the worst weaknesses of using a web search engines is the possibility of altering the web pages in order to appear higher in the results page.

The SEO (Search Engine Optimization) is a very renowned job position, and is very useful, especially for the companies who wish to place their websites as high as they can in the users' results page. The SEO consists of improving the volume of quantity of traffic to a certain web site via the web search engines, based on the theory that the higher a site appears in a web search results page, the more visitors will come to the page. This practice normally includes modifications of the HTML code of the web site in order to be indexed and appeared higher in the results page of the users' queries, as well as other techniques.

The way the web search engines work give the SEOs the opportunity to alter their web sites in order to get a better position in a results page, by using different techniques such as the *cross linking*, which consists of linking between pages of the same website in order to increase the amount of links that the current websites has, and thus to try to increase its visibility. It is also common to see websites just full of frequently searched keywords phrases or words, so as to appear more relevant when the web crawlers travel through the website.

Unfortunately, web search engines are extremely vulnerable

to this type of *disapproved techniques*, but the plus side is that web search engines are constantly improving their algorithms in order to prevent these kinds of frauds, and the way they operate has been enhanced incredibly since the first version of the engines.

A deep analysis may be found in SEOBook [8], a website about *SEO* –Search Engine Optimization- that was written to compare the differences between the main web search engines that were operating then (the article was written in 2006 and revised in 2007): Google, Yahoo! and MS Search. Essentially, it shows very interesting results such as that Google, on one hand, is so much better than Yahoo! or MSN Search when determining if a certain link is a real citation or an artificial link -used to rank higher in the results page and used as a spamming technique-, but counteracts this weakness by determining the quality of the link, and not only the quantity, which may make a site to score lower on the PageRank algorithm or not even being indexed by the crawler if linked excessively by low quality links. Yahoo! for example, besides having been one of the first big web search engines, as it has been previously said, has a paid inclusion program and normally makes the results page biased to show commercial content, and this gives MSN Search and Google a great advantage over it.

MSN Search uses a very different system of ranking webpages, for example assigning too much weight to the page content and not to the links, and new sites –that might be potentially dangerous and unsafe- are ranked high in the search algorithm and thus show in a high position of the results page, which may result in a harmful situation for the final user.

Nevertheless, web search engines are incredibly versatile and their strengths massively outweigh their weaknesses. They give the users a huge amount of possibilities to discover internet, as well as a way to save time when surfing the net. They offer almost instant information, the chance to help the user to discover new websites, the possibility of helping the user to specify their queries –for example if the user is not able to remember a certain information or the URL of a website, most of the web search engines include the possibility of correcting what the user types in case this types it incorrectly, based on the amount of queries that have been introduced in the web search before.

These powerful tools have become indispensable in every internet user's life, not only because they save time, but also because they offer some functionalities that are extremely useful, as well as safe for a new user. For example the modern browsers such as Firefox and Chrome have incorporated a very nice feature in case the user mistypes a website address, in which case the browser automatically shows a Google results page (or whichever default web search engines the user has marked as default in the browser) let the user know that the website that is being looked for is not available or does not exist, showing also sometimes, a correction of the URL in case the web search engines finds some patterns that could match what the user intended to look for, even they might have *cached* – saved the webpage in a small but very fast memory- the website that might not be already available on internet, letting the user still read it on the web search engine's memory.

Web search engines not only let the users look for some words to be found in a web page, but some of them, like Google and Yahoo! offer different services to help the users take the maximum advantage of the search queries. For example Yahoo! offers a service called Yahoo! BOSS [9] which let the users build their own web search and that is specially designed and built for companies, so that they can offer their clients a personalized way to search, for example, combining the powerful features that Yahoo! offers plus some proprietary technology that the company has, resulting in very interesting products such as *Cluuz* [10], a web search engine that thanks to the user of Yahoo! BOSS, generates a more understandable results page using semantic analysis and image extraction in real-time.

Google for example, provides users with not only with their powerful web search engines, but also let them personalize their search queries to look for only images, or news, scholar papers or even software code.

Moreover, in the last years there has been a new generation of web search engines that promise to astonish the world with their power. One of these is the Wolfram Alpha, an answer engine capable of resolving factual queries and natural language based questions by computing the answer and presenting the results to the user by means of structured data, and not by simply showing a list of URLs and images as normal web search engines do. Wolfram Alpha is also able to answer and resolve complex mathematical and calculus questions, such as limits.

INTENDED APPLICATIONS

The intended application of search-engines are well-known. The objective is to find web sites that contain files or information that the user is looking for, and millions of people are using search engines exactly for this.

For example if you were looking for a restaurant in Amsterdam you can type the query "restaurant Amsterdam" into Google [16] and you will get over 5 million results. Unfortunately, not all of these results will be relevant, but due to the page ranking the results on the front page has 6 websites for individual restaurants in Amsterdam, 3 websites which give reviews of lots of restaurants in Amsterdam, and the location on Google maps of more than 7 restaurants. This shows that the search engines are able to bring you relevant results, quickly, out of millions of possibilities.

UNINTENDED APPLICATIONS

Besides the rise of SEO and spamming, there are luckily also some more innocent examples of search engine 'abuse'. A famous one that actually made it to news bulletins at the time is the sport of 'Google Whacking'.

In 2002, some people started playing around with Google's search function just for fun. As most of us have probably done at some times in our lives, they tried to find queries with the largest number of results. Quickly, people realized that is was more fun to try and find the smallest number of results instead, and soon formulated the 'Google Whacking' rules:

- try and find two unique words that produce together exactly one result on Google.

This simple game can be played by anyone; all you need is access to a web search engine. Valid 'whacks' can be submitted to 'The Whack Stack' (www.googlewhack.com/tally.pl) where they are verified and shared with the result of the world.

Paradoxically, when a valid whack gets recorded onto the Whack Stack, there are actually two places online where those two words can be found together. Luckily, Google no longer indexes the Whack Stack, allowing people to share their word tuple without compromising it as a Googlewhack.

Naturally, due to the ever changing content of web, a Googlewhack might become obsolete after a while. Also, while the internet keeps growing, it will be harder and harder to find new ones in the future.

You can read more about the phenomenon here [11].

Another unintended application of search engines is when the number of people searching for an item is used as the information. This has been used to track epidemics of disease by looking at the changes in the number of people searching for certain symptoms [14]. Google has made this easier to do by releasing Google trends which lets you see the change in people searching for a term over a large timespan[15].

Another different way of looking at search results, is looking at the number of results returned. This can turn your search engine into a dictionary (as correctly spelled words will produce more results).

GETTING STARTED

The standard interface to most web search engines needs no further explanation. However, many of these engines like Google and Yahoo! also expose API's to their technology, meaning that developers can do searches from within their software or websites.

Using the web search API is quite simple in both Google's and Yahoo!'s cases. From here on we will focus on Yahoo!, but Google offers roughly equivalent technologies to the ones we describe.

In a broad range of services, the most interesting ones are Boss (Build Your Own Search Service) and YQL. Boss allows access to Yahoo!'s index trough specifying queries in a URI. The server then returns the results as XML or JSON. A general query looks like this:

http://boss.yahooapis.com/ysearch/web/v1/{query}?
appid={yourBOSSappid}
[&param1=val1&param2=val2&etc]

Different parameters can be provided to in/exclude certain domains, geographic regions, adult content, etc. Also, users can define boolean combinations of queries and parameters. [12]

Even more interesting is YQL (Yahoo Query Language), a relatively new service. It affords access to a variety of categorized indexes or 'tables' through SQL-like queries. One of these tables is Yahoo's main search index, but loads of others like 'local', 'flickr', 'music' and 'weather' exist. The community can also upload and maintain data tables in the YQL platform, opening up more and more categorized information to the world.

YQL allows for example to get the 50 nearest restaurants from your current GPS location that serve Heineken beer, or get some pages about the latest news topic that just came in through RSS feeds. A query could look like this:

select * from local.search where query="sushi" and location="san francisco, ca"

[13]

Visit http://developer.yahoo.com/yql/console/ to try out the API, or http://code.google.com/apis/ajax/playground/ for a similar service from Google.

FINAL THOUGHTS

As the number of websites on the internet continues to increase, search-engines are going to have improve and adapt in order to bring information to the user as quickly and reliably as possible. They are also going to have to develop new algorithms that are able to deal with increasingly sophisticated spam methods. There is also the possibility that there will be a split between those search engines that provide a list of hits, and those that give you a single answer.

REFERENCES

0. Zhang et al, Evolution of the Internet and its cores, New Journal of Physics, vol. 10, December 2008 (http://iopscience.iop.org/1367-2630/10/12/123027)

1.How internet search engines work, http://computer.howstuffworks.com/internet/basics/search-engine.htm

2.Web search engines, http://en.wikipedia.org/wiki/Web_search_engine

3.How does a search engine work? http://www.buzzle.com/articles/how-does-a-search-engine-work.html

4.Hash table, http://en.wikipedia.org/wiki/Hash_table

5.Boolean searching on internet http://www.internettutorials.net/boolean.asp

6.Google Search Basics

http://www.google.com/support/websearch/bin/answer.py?hl=en&answer=134479, http://www.google.com/support/websearch/bin/answer.py?answer=136861

7. Top Search Engines in 2010 http://www.seoconsultants.com/search-engines/

8. SEOBook. http://www.seobook.com/relevancy/#

9. Yahoo! BOSS, http://developer.yahoo.com/search/boss/

10. Cluuz, http://www.cluuz.com/

11. Gary Stock, GoogleWhack history, http://www.unblinking.com/heh/googlewhack.htm#20020108, retrieved 2010-05-27

12. Yahoo! Inc, Boss API guide, http://developer.yahoo.com/search/boss/boss_guide/, retrieved 2010-05-29

13. YQL Guide, Yahoo! Inc, http://developer.yahoo.com/yql/guide/, retrieved 2010-05-29

14. Detecting influenza epidemics using search engine query data, http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/papers/detecting-influenza-epidemics.pdf

15. Google Trends, http://www.google.com/trends

16. Google Restaurant query, http://www.google.nl/search?q=restaurant+amsterdam