Sections

Option C - Web Science (/topic/C-web-science.html)

C.2 Searching the web

How search engines work and how websites can be optimized for searches.

C.2.1 Define the term search engine

A search engine is a software that allows a user to search for information in the web.

C.2.2 Distinguish between the surface web and the deep web

Surface web

The surface web is the part of the web that can be reached by a search engine. For this, pages need to be static and fixed, so that they can be reached through links from other sites on the surface web. They also need to be accessible without special configuration. Examples include Google, Facebook, Youtube, etc.

Summary:

- Pages that are reachable (and indexed) by a search engine
- Pages that can be reached through links from other sites in the surface web
- Pages that do not require special access configurations

Deep web

The deep web is the part of the web that is not searchable by normal search engines. Reasons for this include proprietary content that requires authentication or VPN access, e.g. private social media, emails; commercial content that is protected by paywalls, e.g. online news papers, academic research databases; personal information that is protected, e.g. bank information, health records; dynamic content. Dynamic content is usually a result of some query, where data are fetched from a database.

Summary:

- Pages not reachable by search engines
- Substantially larger than the surface web
- Common characteristics:
    - Dynamically generated pages, e.g. through queries, JavaScript, AJAX, Flash
    - Password protected pages, e.g. emails, private social media
        - Paywalls, e.g. online news papers, academic research databases

        - personal information, e.g. health records
    - Pages without any incoming links

C.2.3 Outline the principles of searching algorithms used by search engines

The most known search algorithms are PageRank and the HITS algorithm, but it is important to know that most search engines include various other factors as well, e.g.

- the time that a page has existed
- the frequency of the search keywords on the page
- other unknown factors (undisclosed)

For the following description the terms “inlinks” and “outlinks” are used. Inlinks are links that point to the page in question, i.e. if page W has an inlink, there is a page Z containing the URL of page W. Outlinks are links that point to a different page than the one in question, i.e. if page W has an outlink, it is an URL of another page, e.g. page Z.

PageRank algorithm

PageRank works by counting the number and quality of inlinks of a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

The value of an in-link from a page A is proportional to

```
P(A) / C(A)
```

Where P(A) is the PageRank score of page A, and C(A) is the number of out-links page A has.

As mentioned it is important to note that there are many other factors considered. For instance, the anchor text of a link is often far more important than its PageRank score.

Summary:

- Pages are given a score (rank)
- Rank determines the order in which pages appear
- Inlinks add value to a page
- The importance of an inlink depends on the PageRank (score) of the linking page
- PageRank counts links per page and determines which page are most important

HITS algorithm

Based on the idea that keywords are not everything that matters; there are sites that might be more relevant even if they don't contain the most keywords. It introduces the idea of different types of pages, authorities and hubs.

Authorities: A page is called an authority, if it contains valuable information and if it is truly relevant for the search query. It is assumed that such a page has a high number of in-links.

Hubs: These are pages that are relevant for finding authorities. They contain useful links towards them. It is therefore assumed that these pages have a high number of out-links.

The algorithm is based on mathematical graph theory, where a page is represented by a vertex and links between pages are represented by edges (in form of vectors).

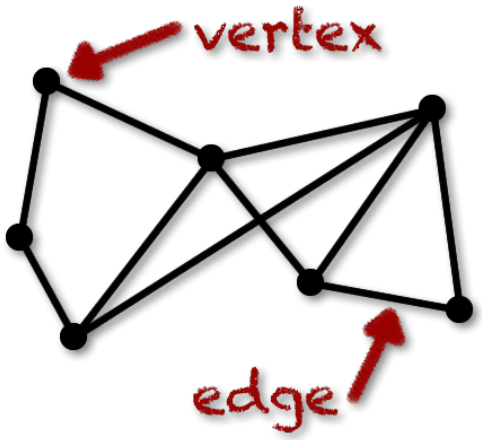

Figure 1: A simple graph [1 (http://world.mathigon.org/Graph_Theory)]

The algorithm starts by creating a graph:

- It first finds the top 200 pages based on the occurrence of keywords from the query. Let's call the set of these pages RQ
- It then finds all pages that link to the set of pages RQ and all pages which these link to (basically all pages linked in or out). Together with RQ this makes up the set SQ
- The algorithm gives each page in the set SQ a hub weight and an authority weight, based on how many pages link towards it (authority) and how many pages it links to (hub)
- The algorithm then lists the pages based on their weight

C.2.4 Describe how a web-crawler functions.

A web crawler, also known as a web spider, web robot or simply bot, is a program that browses the web in a methodical and automated manner. For each page it finds, a copy is downloaded and indexed. In this process it extracts all links from the given page and then repeats the same process for all found links. This way, it tries to find as many pages as possible.

Limitations:

- They might look at meta data contained in the head of web pages, but this depends on the crawler
- A crawler might not be able to read pages with dynamic content as they are very simple programs

Robots.txt

Issue: A crawler consumes resources and a page might not wish to be “crawled”. For this reason “robots.txt” files were created, where a page states what should be indexed and what shouldn’t.

- A file that contains components to specify pages on a website that must not be crawled by search engine bots
- File is placed in root directory of the site
- The standard for robots.txt is called “Robots Exclusion Protocol”
- Can be specific to a special web crawler, or apply to all crawlers
- Not all bots follow this standard (malicious bots, malware) -> “illegal” bots can ignore robots.txt
- Still considered to be better to include a robots.txt instead of leaving it out
- It keeps the bots from less “noteworthy” content of a website
  more time spend on indexing important/relevant content of the website

C.2.5 Discuss the relationship between data in a meta tag and how it is accessed by a web-crawler

Answer depends on different crawlers, but generally speaking:

- The title tag, not strictly a meta-tag, is what is shown in the results, through the indexer
- The description meta-tag provides the indexer with a short description of the page
- The keywords meta-tag provides…well keywords about your page

While meta-tags used to play a role in ranking, this has been overused by many pages and therefore meta-tags are not considered by most search engines anymore.

Crawlers now mostly use meta-tags to compare keywords and description to the content of the page to give it a certain weight. For this reason while meta-tags do not play the big role it used to, it’s still important to include them.

C.2.6 Discuss the use of parallel web-crawling

- Size of the web grows, increasing the time it would take to download pages
- To make this reasonable “it becomes imperative to parallelize the crawling process (Stanford (http://ilpubs.stanford.edu:8090/733/1/2002-9.pdf))

Advantages

- Scalability: as the web grows a single process can not handle everything Multithreaded processing can solve the problem
- Network load dispersion: as the web is geographically dispersed, dispersing crawlers disperses the network load
- Network load reduction

Issues of parallel web crawling

- Overlapping: parallel web crawlers might index the same page multiple times
- Quality: If a crawler wants to download 'important' pages first, this might not work in a parallel process
- Communication bandwidth: parallel crawlers need to communicate for the former reasons, which for many processes might take significant communication bandwidth
- If parallel crawlers request the same page frequently over a short time it will overload servers

C.2.7 Outline the purpose of web-indexing in search engines

Search engines index websites in order to respond to search queries with relevant information as quick as possible. For this reason, it stores information about indexed web pages, e.g. keyword, title or descriptions, in its database. This way search engines can quickly identify pages relevant to a search query.

Indexing has the additional purpose of giving a page a certain weight, as described in the search algorithms. This way search results can be ranked, after being indexed.

C.2.8-9 Suggest how developers can create pages that appear more prominently in search engine results. Describe the different metrics used by search engines.

The process of making pages appear more prominently in search engine results is called SEO. There are many different techniques, considered in section C.2.11. This field is a big aspect of web marketing, as search engines do not disclose how exactly they work, making it hard for developers to perfectly optimise pages.

In order to check the web presence of a website, there are different metrics to be used.

Metrics

- **Search Engine Share of Referring visits:** how the web page has been accessed: through direct access, referral pages or search engine results. Can indicate how

meaningful traffic is.

- **Search Engine Referral:** different search engines have different market shares; knowing which search engine traffic comes from helps to find potential improvements for certain search engines
- **Search terms and phrases:** identify the most common search keywords and optimize
- **Conversion rate by search phrase/term:** percentage of users that sign up coming from a search term
- **Number of sites receiving traffic from search engines:** As large websites have many pages, it is important to see if individual sites are being accessed through search engines
- **Time taken:** time spent by a user on a page after access through the search engine. Indicator for how relevant the page is and what resources were accessed
- **Number of hits:** a page hit is when a page is downloaded. This is a counter of the visitors of the page and gives a rough idea of the traffic to the page
- **Quality of returns:** quality of how a site gets placed in a return. Say how high it is ranked by search engines.
- **Quantity of returns:** how many pages are indexed by a search engine

Parameters Search Engines use to compare

- **Relevance:**
  - Is determined by different programs like PageRank etc. which evaluate and determine the quality of web sites and put them high on the Index
  - The bigger the index the more pages the search engine can return that have relevance to each query
- **User experience:**
  - Search engines look to find the “best” results for the searcher and part of this is the user experience a site provides. This includes ease of use, navigation; direct and relevant information; professional, modern and compatible design; high-quality, legitimate and credible content

C.2.10 Explain why the effectiveness of a search engine is determined by the assumptions made when developing it.

A search engine will return results based on the algorithms and parameters used when being developed. These algorithms and parameters are based on assumptions and therefore a search engine can only be effective as long as these assumptions are met. While assumptions can come close to reality, users search in different ways and therefore it can be hard to make universal assumptions.

C.2.11 Discuss the use of white hat and black hat search engine optimisation

Source: Cognitive SEO Infographic (http://cognitiveseo.com/infographics/blackhat-vs-whitehat-seo.jpg)

Black Hat

Definition: Black hat SEO is a technique, in simple words, to get the top positions or higher rankings in the major search engines like Google, Yahoo and Bing that breaks the rule and regulations of search engine's guidelines

Keyword stuffing

Overuse of keywords .The reason search engines don't rely on meta tags anymore.

- Not really effective anymore, because most search engines don't use meta tags anymore.

Link farming

A group of website that all hyperlink to every other site. Usually done by some program.

Hidden texts and links

Text that can't be seen by the end user, but can be found by the search engine.Considered search spam.

- Usually identified by search engines as search spam.

Blog comment spamming

Automated posting of hyperlinks for promotion on any kind of publicly accessible online discussion board. This could be blog comments, wikis, guestbooks, etc.

- Most advanced discussion boards allow to report spam, which decreases effectiveness.
- Most spam is usually easy to identify by the user, therefore decreasing effectiveness.

Content Automation

Content Automation is a process of creating content of the website in a automatic matter by using a tool or script. It means the content of the website will be automatically generate using a tool and published on the website.

Advantages:

- Website will become a very large (As per content, Not ranking of Website) at less time.
- Effort will be less, as the content is generated automatically.
- sudden growth in traffic

Disadvantages:

- Sudden dropdown of the website might occur.
- Violates the search engine guidelines
- website may be banned or blacklisted from the search engine

Scraping

Copies content from popular websites. Often to get more visits and sell advertisements

Paid Links

Paying for links on other sites to receive more visits

Doorway pages

Doorway pages are simple HTML pages that are fully optimized for search engines. Doorway pages target specific keywords or phrases for search engines, but not for users. When users visit the page, the page automatically uses JavaScript or Meta refresh property to redirect visitors to another page

Cloaking

Presenting different content to web spiders than to users, by delivering content based on IP addresses.

White Hat

Guest Blogging

The process of writing a blog post for someone else's blog is called guest blogging

- Increases backlinks to guest blogger's site and search engine rankings
- must be a lot of Guest blogging to get lots of backlinks
- and also highly depend on "authoritative" blog

Link Baiting

Link baiting to incentives people to click on a link, usually done by writing sensational or controversial content or title.

- It is highly inciting to read content that is sensational or controversial
- But it does not increase ranking directly, only through the viewership it might get a higher ranking; and not anyone falls for "link-baity" content

Quality Content

Search engines evaluate the content of a web page, thus a web page might get higher ranking with more information. This will make it more valuable on the index and other web pages might link to your web page if it has a high standard in content.

- It is time consuming to create good content
- But in long run it will be worth the effort

Site optimization

Through manipulation of content wording and site structure; tweaking content, and meta tags maximizes search engine efficiency.

robots.txt

Getting indexed by crawler and prevent duplication of content preventing algorithm to index it as redundant information

- Prevents crawler from indexing irrelevant or redundant information

C.2.12 Outline future challenges to search engines as the web continues to grow

As the web grows, it becomes harder to filter out the most relevant information, and paid results (ads) play an important role. Some data become more semantic as well and search engines will need to adapt to this.

Sources:

[1] Mathigon.org (http://world.mathigon.org/Graph_Theory)