Search Engines
Charles Severance

Google Architecture
- Web Crawling
- Index Building
- Searching

http://infolab.stanford.edu/~backrub/google.html

Google Search
- Google I/O '08 Keynote by Marissa Mayer
- Usability / User Experience / User Testing / Architecture / Philosophy
- Required Viewing

http://www.youtube.com/watch?v=6x0cAzQ7PVs

Search-Friendly Web Development
- Google I/O
- Maile Ohye (Google) - June 10, 2008
- Mission: Organize the world's information and make it universally accessible and useful
- Required Viewing

http://www.youtube.com/watch?v=NIWcZPI4Nk
Search-Friendly Web Sites

- What should you do to insure your site works well for Google Search (alt tags, title, description, url design)
- How can your site get in trouble?
- Google's focus on “User Experience” and Usability and how they feel the when your site is clicked from a search that is reflects on them

http://www.youtube.com/watch?v=NIWtZPIHNk

Web Crawler

A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.

http://en.wikipedia.org/wiki/Web_crawler

Web Crawler

- Retrieve a page
- Look through the page for links
- Add the links to a list of “to be retrieved” sites
- Repeat...

http://en.wikipedia.org/wiki/Web_crawler

Web Crawling Policy

- a selection policy that states which pages to download,
- a re-visit policy that states when to check for changes to the pages,
- a politeness policy that states how to avoid overloading Web sites, and
- a parallelization policy that states how to coordinate distributed Web crawlers

http://en.wikipedia.org/wiki/Web_crawler
robots.txt

- A way for a web site to communicate with web crawlers
- An informal and voluntary standard
- Sometimes folks make a "Spider Trap" to catch "bad" spiders


Search Indexing

Search engine indexing collects, parses, and stores data to facilitate fast and accurate information retrieval. The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan every document in the corpus, which would require considerable time and computing power.

http://en.wikipedia.org/wiki/Index_(search_engine)

Inverted Index

- An Inverted Index lists all of the documents which contain a particular word
- Allows us to quickly produce a list of documents given one or a few search terms
- The problem with the web is that we have too many documents

http://en.wikipedia.org/wiki/Inverted_index

PageRank

- Basic Idea: Incoming links signal "value" or "interest"
- Incoming links from other high-ranking sites have greater value
- Computed by giving all sites some "value" and letting value flow out the outbound links and in the inbound links until value stabilizes

http://en.wikipedia.org/wiki/PageRank
PageRank Spamming

- The computation is simple on its face and fraught with ways to spam and adjust it
- Google Bombing
- Reciprocal Links
- Google is very watchful and takes care of these things by “adjustments”

Gaming PageRank

- The real ranking mechanism has many subtle tuning parameters which are kept secret as well as human intervention
- Once the web site builders “know” the rules - they can game the system
- A busy little consultancy - Search Engine Optimization (SEO)

Search Engine Optimization

- Google’s “organic” results are free and can be very lucrative for companies “diamonds”
- Enterprising web site owners “guess” how the Google rules work
- They make changes to their web sites to take advantage of the rules
Google Supplemental Index

- Not a good place to be - crawling happens less frequently and seldom appear in search results
- Causes: duplicate content, low page rank, link manipulation, page freshness, etc.

http://en.wikipedia.org/wiki/Supplemental_Result
http://video.google.com/videoplay?docid=49816-90513029456 (SEO Funny)

Google's Webmaster Central

- Lets you work with Google's crawler and index with regards to your site
- You establish ownership of a site by adding a meta-tag
- You can look at crawling activity, page rank, set up a site map, etc.

http://www.google.com/webmasters/

PageRank Story

http://www.google.com/webmasters/
Webmaster Guidelines

- Content design
- Search Engine Optimization
- Technical Issues

http://google.com/support/webmasters/bin/answer.py?answer=35769

- Make a site with a clear hierarchy and text links. Every page should be reachable from at least one static text link.
- Offer a site map to your users with links that point to the important parts of your site. If the site map is larger than 100 or so links, you may want to break the site map into separate pages.
- Create a useful, information-rich site, and write pages that clearly and accurately describe your content.
- Think about the words users would type to find your pages, and make sure that your site actually includes those words within it.

http://www.google.com/support/webmasters/bin/answer.py?answer=35769

- Try to use text instead of images to display important names, content, or links. The Google crawler doesn’t recognize text contained in images. If you must use images for textual content, consider using the "ALT" attribute to include a few words of descriptive text.
- Make sure that your <title> elements and ALT attributes are descriptive and accurate.
- Check for broken links and correct HTML. Keep the links on a given page to a reasonable number (fewer than 100).
- If you decide to use dynamic pages (i.e., the URL contains a "?" character), be aware that not every search engine spider crawls dynamic pages as well as static pages. It helps to keep the parameters short and the number of them few.

http://www.google.com/support/webmasters/bin/answer.py?answer=35769

- Google Keyword Tool

- Allows you to explore different keywords and see approximate prices

https://adwords.google.com/select/KeywordToolExternal?defaultView=2
Search Queries

A web search query is a query that a user enters into web search engine to satisfy his or her information needs. Web search queries are distinctive in that they are unstructured and often ambiguous; they vary greatly from standard query languages which are governed by strict syntax rules.

Categories of Search Queries

- **Informational queries** – Queries that cover a broad topic (e.g., colorado or trucks) for which there may be thousands of relevant results.
- **Navigational queries** – Queries that seek a single website or web page of a single entity (e.g., youtube or delta airlines).
- **Transactional queries** – Queries that reflect the intent of the user to perform a particular action, like purchasing a car or downloading a screen saver.

---

<table>
<thead>
<tr>
<th>Informational Query</th>
<th>Navigational Query</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Informational Query" /></td>
<td><img src="image2.png" alt="Navigational Query" /></td>
</tr>
</tbody>
</table>
Google Search Architecture

- Very fast
- Many servers see your query

http://research.google.com/archive/googlecluster.html

Required Viewing

http://www.youtube.com/watch?v=6x0cAaQ7PVs
Maile Ohye

http://www.youtube.com/watch?v=NIWzZPl4Nk
Marissa Mayer

Advanced Topics (not required)

http://highscalability.com/google-architecture
http://video.google.com/videoplay?docid=7278544053668715642 Big Table
http://infolab.stanford.edu/~backrub/google.html
Search Summary

- Web Crawling
- Index Building
- Searching

http://infolab.stanford.edu/~backrub/google.html