Search engines

patrick.j.rice@gmail.com

What is a Search engines

- A web search engine is designed to search for information on the World Wide Web and FTP servers.

In the beginning

- Gopher (“go for”)
  - Text search of an index (like searching through a filing cabinet)
- Magellan, Excite, Infoseek, Inktomi, Northern Light, AltaVista and Yahoo!

Google

Patrick.j.rice@gmail.com

What is google

- Internet search
- founded by Larry Page and Sergey Brin while they were students at Stanford University
- Google began in January 1996, as a research project by Larry Page, who was soon joined by Sergey Brin,

What is Google

- The analyzed the relationships between websites (links)
- pages with the most links to them from other highly relevant web pages must be the most relevant pages associated with the search.
- 99% of Google's revenue is derived from its advertising programs.

Google

- developing the "perfect search engine,"
- co-founder Larry Page as something that, "understands exactly what you mean and gives you back exactly what you want."
- most search engines ran off a handful of large servers that often slowed under peak loads
- Google employed linked PCs to quickly find each query's answer.

The googolplex

Google Search

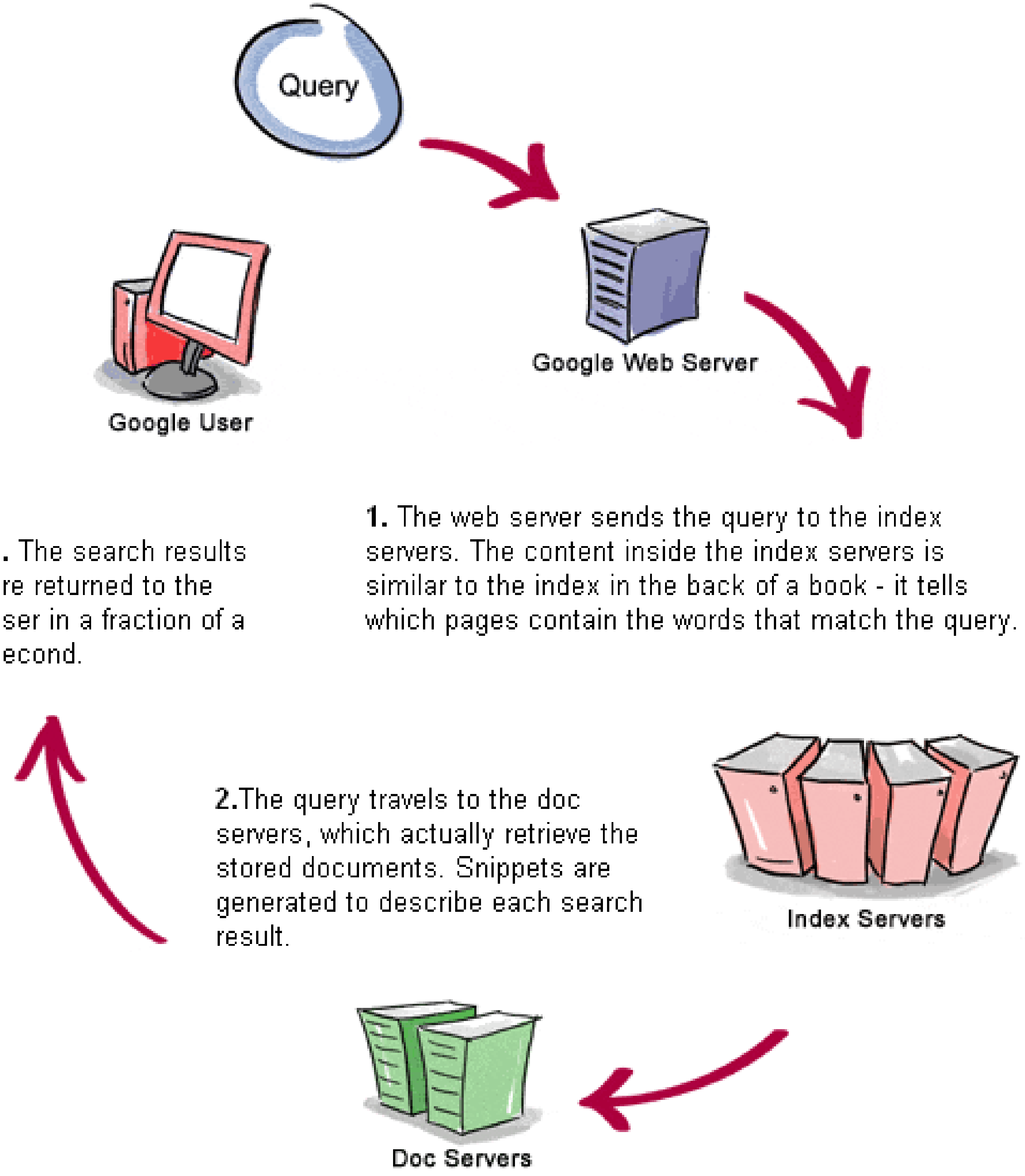
- 200 signals, including a patented PageRank™ algorithm
- entire link structure of the web
- determine which pages are most important
- conduct hypertext-matching analysis to determine which pages are relevant to the specific search being conducted.

PageRank Technology

- reflects Google's view of the importance of web pages by considering more than 500 million variables and 2 billion terms.
- Pages that Google believe are important pages receive a higher PageRank
- are more likely to appear at the top of the search results.

Hypertext-Matching Analysis

- analyzes page content
- analyzes the full content of a page
- factors in fonts, subdivisions and the precise location of each word
- analyze the content of neighboring web pages



1. The web server sends the query to the index servers. The content inside the index servers is similar to the index in the back of a book - it tells which pages contain the words that match the query.

2.The query travels to the doc servers, which actually retrieve the stored documents. Snippets are generated to describe each search result.

3. The search results are returned to the user in a fraction of a second.

Getting googled

- Submit it to Google at http://www.google.com/addurl.html.
- Get linked, ask people to link to your site.
- Create a Sitemap.

But I must get to the top of Google!!!!!!

http://www.flickr.com/photos/karanj/31469695/sizes/z/in/photostream/

Search engine optimization (SEO)

- is the process of improving the volume or quality of traffic to a web site from search engines via "natural" or un-paid ("organic" or "algorithmic") search results.
- Getting indexed
  - Google and Yahoo!, use crawlers to find pages for their algorithmic search results.
- Cross linking between pages of the same website. This increases PageRank used by search engines
- Keyword rich text in the webpage and key phrases.

Getting Googled Site Design

- Make a site with a clear hierarchy and text links.
- Offer a site map to your users
  - links that point to the important parts of your site.
- information-rich site, and write pages that clearly and accurately describe your content.

Getting Googled Site Design

- Try to use text instead of images to display important names, content, or links.
- The Google crawler doesn't recognize text contained in images.
- In pictures use the "ALT" attribute to include a few words of descriptive text.

Getting Googled Site Design

- Check for broken links and correct HTML.
- links on a given page to a reasonable number (fewer than 100).
- Use a text browser such as Lynx to examine your site
- Allow search bots to crawl your sites
- Make use of the robots.txt file on your web server
    - This file tells crawlers which directories can or cannot be crawled.

Google design don'ts

- Avoid hidden text or hidden links.
- Don't use cloaking or sneaky redirects.
- Don't send automated queries to Google.
- Don't load pages with irrelevant keywords.
- Don't create multiple pages, subdomains, or domains with substantially duplicate content.
- Don't create pages with malicious behavior, such as phishing or installing viruses, trojans, or other badware.

More Search engines

- Bing
  - Microsoft latest incarnation of search
  - Is a standard installed “search” on all computers, so it should be considered
- Yahoo
  - Is based on directory's
  - Look up cats, you get a list of cat products

More Search Engines

- Ask.com (known as Ask Jeeves in the UK)
- Baidu (Chinese, Japanese)
- Bing (formerly MSN Search and Live Search)
- Blekko
- Duck Duck Go
- Google
- Kosmix
- Sogou (Chinese)
- Yodao (Chinese)
- Yahoo! Search
- Yandex (Russian)
- Yebol

Even More Search Engines

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