Search Engine

A BRIEF HISTORY

- Search engines were born out of the idea that the Internet is big – very BIG! In the beginning, there was really no easy way to find information on the web, and many directories that did exist were maintained by hand.

- Finding information on the Internet was very inefficient and manual, creating the need for a program that could act as a WWW resource-discovery tool to combine the three essential features of a search engine (crawling, indexing, and searching).
- Originally, because of the limited resources available, indexing and hence searching were limited to the titles and headings found in the web pages the crawler encountered.

- One of the first “full text” crawler-based search engines was WebCrawler, which came out in 1994. Unlike its predecessors, it let users search for any word in any webpage, which has become the standard for all major search engines since.
- At that time, several other search engines came out and vied for popularity including: Magellan, Excite, Infoseek, Inktomi, Northern Light, AltaVista, and Yahoo.

- Around 2000, Google's search engine rose to power most through their innovative new way to achieve better results, called "PageRank." After that, the rest is history.

HOW SEARCH ENGINES WORK

- When you sit down at your computer and perform a search query in Google, Yahoo, or Bing you're presented with a list of results from all over the web. The big questions are:

1. How do search engines find web pages that match your query?
2. How do they determine the order in which to display the search results?

- In it’s most basic sense, you can think of searching the web as looking through the index of a really really large book which tells you right where to find everything. Similarly, when you perform a search query, search engines use their intuitive technology to check their in index and return (or “serve”) the most relevant search results to your query.

- There are 3 key processes in delivery search results to a user:

1. **Crawling** – Do search engines know about your site? Can they find it?
2. **Indexing** – Is your site indexable?
3. **Serving** – Does the site have useful content that is highly relevant to user search queries?

#1. Crawling

- Crawling is the process by which search engines discover new and updated pages to be added to the search index.
- Search engines use a vast network of computers and servers across the world to fetch (or “crawl”) billions of pages on the web. The program that does the fetching is called a “spider” – or robot/bot (Google calls theirs “Googlebot”).

- A **web spider** is an automated Web browser whose job is to follow every link on a site, scan all the information on a webpage, and bring it back to the search engine's servers to index.
- Each search engine uses an algorithmic process to determine which sites to crawl, how often, and how many pages to fetch from each site.

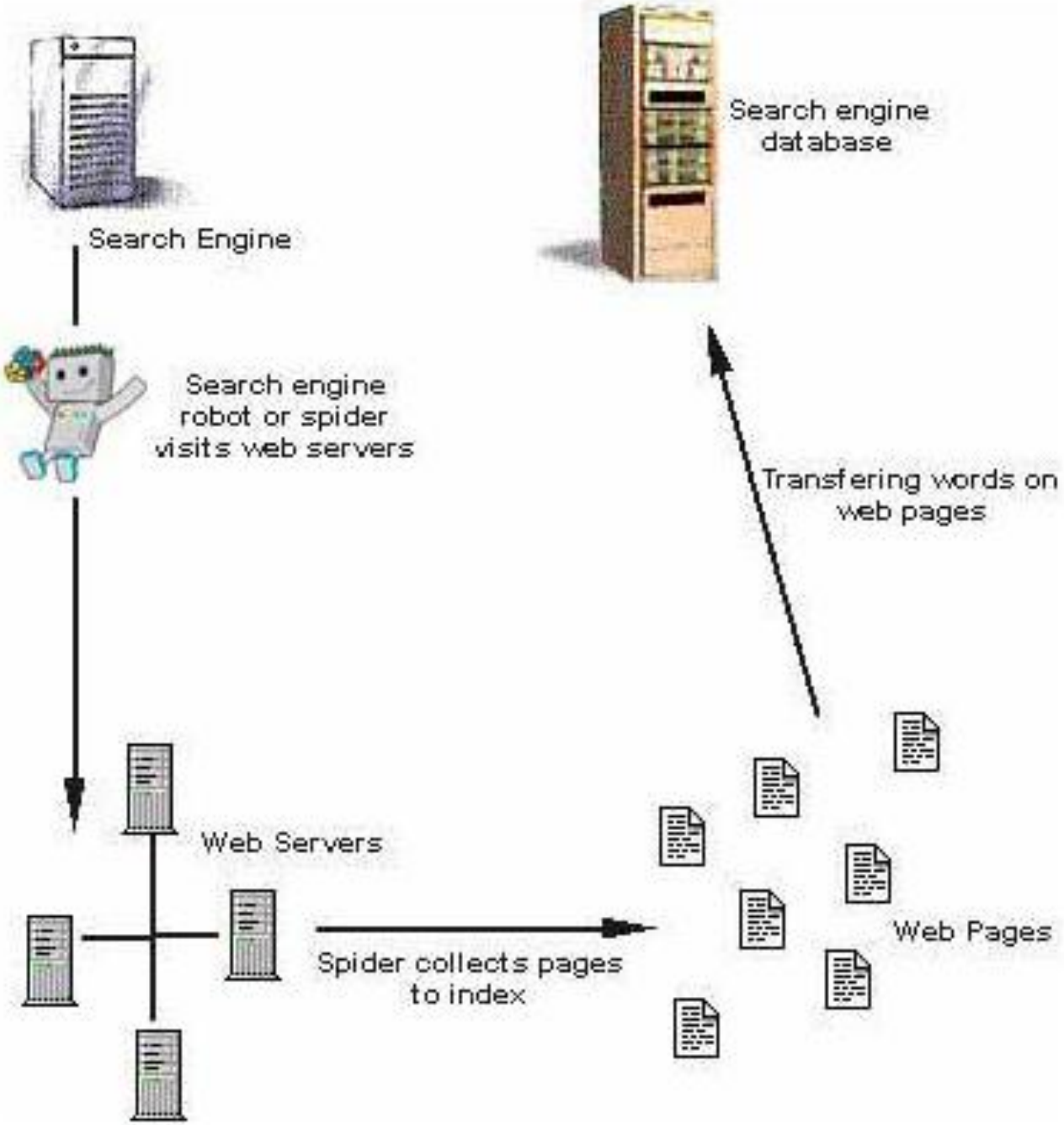
- A search engine's crawl process typically begins with a list of web page URLs, which is usually generated from previous crawl processes. It is then augmented with sitemap data provided by webmasters. Webmasters can use services such as Google Webmaster Tools to submit their sitemaps specifically to Google.

- As a web spider visits each of these websites it detects links on each page and adds them to its list of pages to crawl. New websites, changes to existing websites, as well as dead links are notated and used to update the search engine's index.
- Major search engines do not accept payment to crawl a site more frequently (or rank it higher), but they do offer services such as Pay-Per-Click advertising as a supplement to their natural search engine.

#2. Indexing

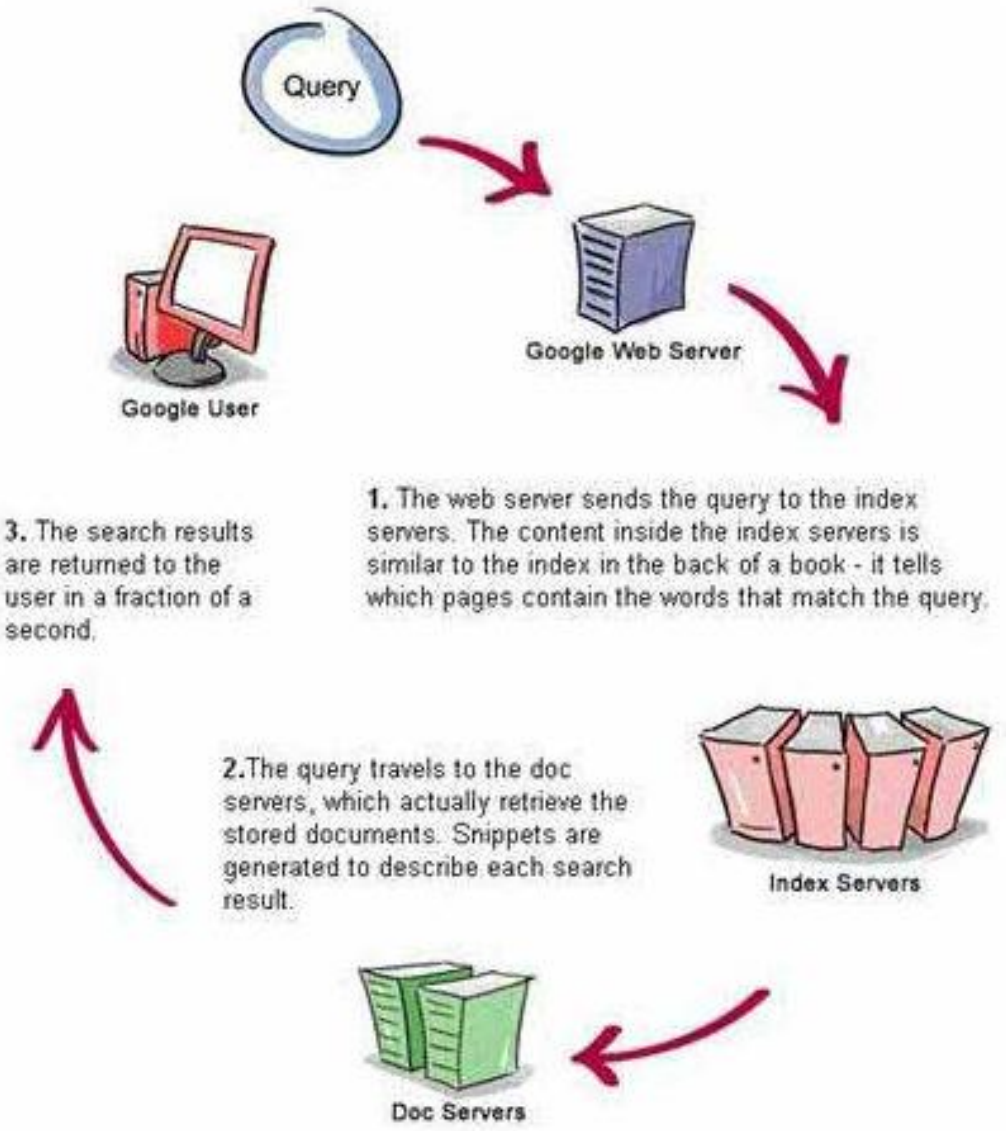
- Web crawlers systematically process each of the web pages they crawl in order to compile a massive index of all the words it sees and their location on each given web page.
- In addition, search engines process information included in key content tags and attributes, such as Title tags, Meta Tags, and ALT attributes.

- Something to keep in mind: Search engines can process most, but not all, content types. For example, they cannot process the content of some rich media files or dynamic pages. Typically, search engines have trouble with things like Flash or JavaScript-based pages.


Search engine database
Search Engine
Search engine robot or spider visits web servers
Transfering words on web pages
Web Servers
Web Pages
Spider collects pages to index

#3. Serving Results

- This is the part where you, the user comes in. When you visit a site such as Google and enter a search query, Google's machines search their index for web pages that match your search query.
- They then return (or "serve") up the results in the natural section of their Search Engine Results Page (SERP) ordered from **most relevant to least relevant**among results that match the given search query.


Query
Google User
Google Web Server
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.
Index Servers
2.The query travels to the doc servers, which actually retrieve the stored documents. Snippets are generated to describe each search result.
Doc Servers
3. The search results are returned to the user in a fraction of a second.

#4. Determining Relevancy

- So how does a search engine determine the relevancy of a group of web pages when serving up the pages in its SERPs? Most search engines have what is called an algorithm, which is a set of rules (or unique formula) for determining the significance and relevancy of web pages.

- The algorithms of each search engine are completely unique, and every search engine handles web pages a little differently. However, the premise of each algorithm is the same – to find information on the web that somebody might find interesting or relevant to their given search query.
- Google's algorithm specifically includes over 200 different factors, including Google PageRank – which was named after one of Google's founders Larry Page.

- PageRank is the measure of the importance of a page based on the incoming links from other pages. In simple terms, each link to a page on your site from another site adds to your site's PageRank.
- However, not all links are equal. Google works hard to improve the user experience by identifying spam links and other practices that negatively impact search results. The best types of links are those that are given based on the quality of your content.

- In order for your site to rank well in SERPs, it's important to make sure that Google can crawl and index your site correctly. Check out Google's Webmaster Guidelines, which outline some best practices that can help you avoid common SEO pitfalls and which should improve your site's ranking.

Basic principles

- Make pages primarily for users, not for search engines.
- Don't deceive your users.
- Avoid tricks intended to improve search engine rankings. A good rule of thumb is whether you'd feel comfortable explaining what you've done to a website that competes with you, or to a Google employee. Another useful test is to ask, "Does this help my users? Would I do this if search engines didn't exist?"
- Think about what makes your website unique, valuable, or engaging. Make your website stand out from others in your field.