Reading Discussion

Blown to Bits

Chapter 4

Needles in the Haystack

Google and Other Brokers in the Bits Bazaar

Notes for CSC 100 - The Beauty and Joy of Computing
The University of North Carolina at Greensboro

Question 1....

Describe the heart of the chapter in a few words

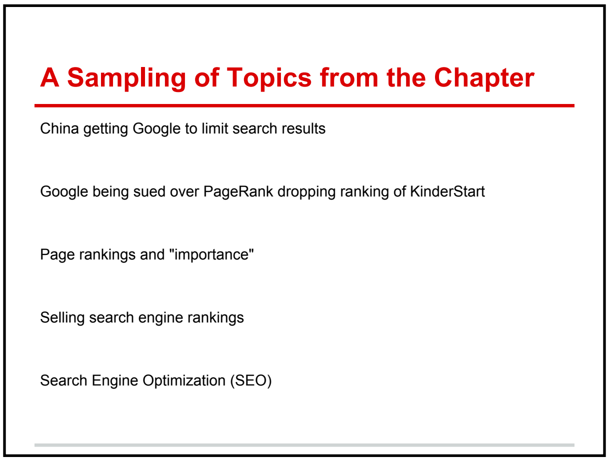
What's the Point?

From a student reading reflection - good one-sentence answer to "what was the main point of the chapter?"

> *The main idea of the chapter is to explain how search engines control how users obtain information on the web, how they work, and how they can be potentially exploited.*

There's something in this answer that most people missed.

What is it?



A Sampling of Topics from the Chapter

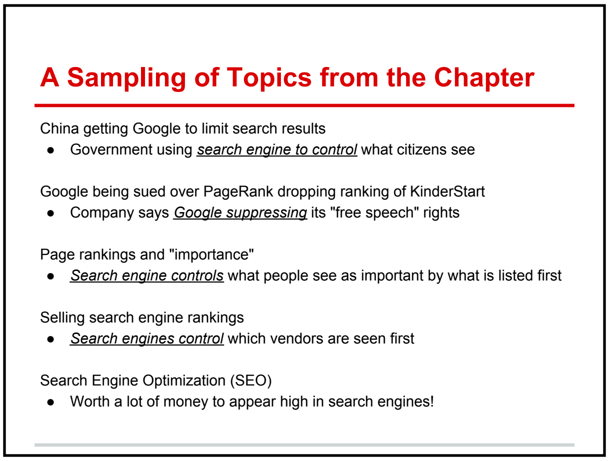
China getting Google to limit search results

Google being sued over PageRank dropping ranking of KinderStart

Page rankings and "importance"

Selling search engine rankings

Search Engine Optimization (SEO)



A Sampling of Topics from the Chapter

China getting Google to limit search results

- Government using *search engine to control* what citizens see

Google being sued over PageRank dropping ranking of KinderStart

- Company says *Google suppressing* its "free speech" rights

Page rankings and "importance"

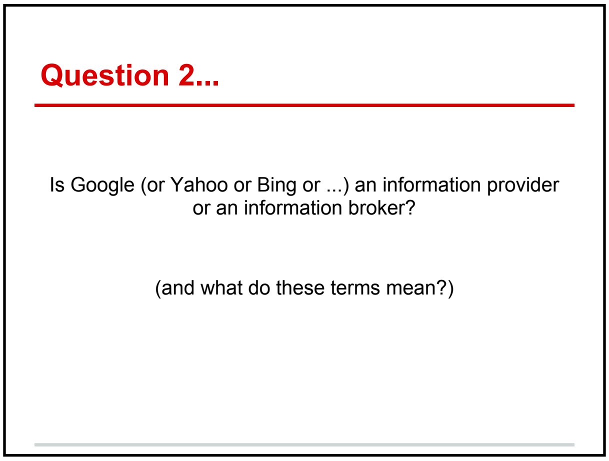
- *Search engine controls* what people see as important by what is listed first

Selling search engine rankings

- *Search engines control* which vendors are seen first

Search Engine Optimization (SEO)

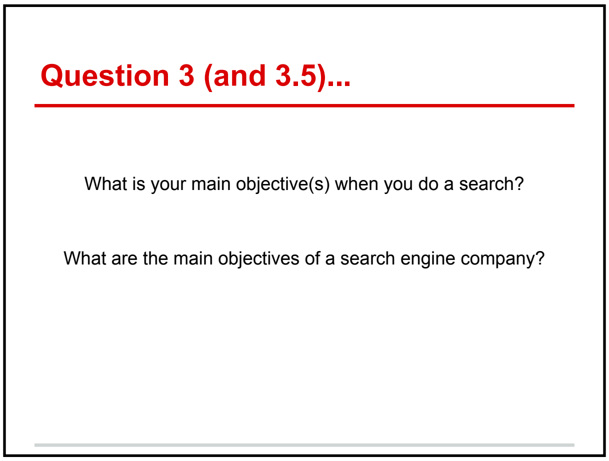
- Worth a lot of money to appear high in search engines!



Question 2...

Is Google (or Yahoo or Bing or ...) an information provider or an information broker?

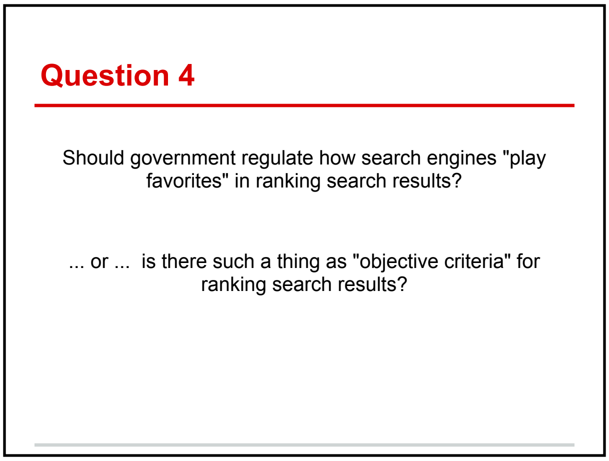
(and what do these terms mean?)



Question 3 (and 3.5)...

What is your main objective(s) when you do a search?

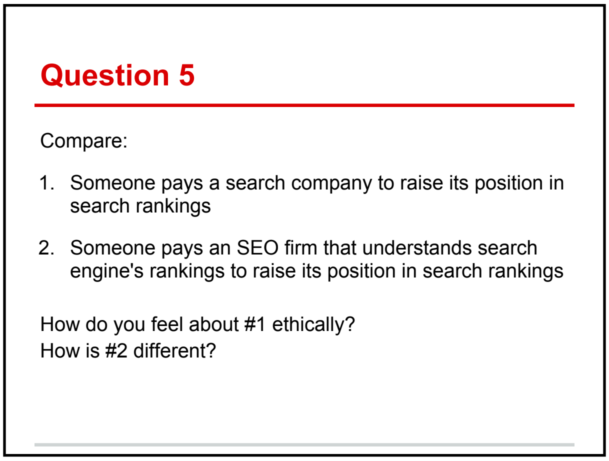
What are the main objectives of a search engine company?



Question 4

Should government regulate how search engines "play favorites" in ranking search results?

... or ... is there such a thing as "objective criteria" for ranking search results?



Question 5

Compare:

1. Someone pays a search company to raise its position in search rankings
2. Someone pays an SEO firm that understands search engine's rankings to raise its position in search rankings

How do you feel about #1 ethically?
How is #2 different?

Understanding How Search Engines Work

Background Processing (server initiated)

- Collecting web pages (crawling the web: spiders)
- Indexing information - must understand data representations!
  - Text and HTML: Easy to extract words
  - Doc and PDF: Not as easy, but doing better now
  - Images (scanned documents, pictures, etc.): Hard to make sense out of!

Foreground Processing (user initiated)

- Understanding query and finding relevant pages
- Ranking relevant pages
  - This is key! Who determines what is "most relevant"? Can make or break web-based businesses!
- Presenting to the user

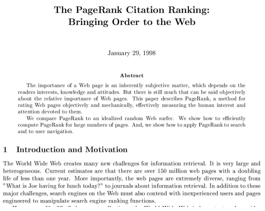
Google PageRank

This algorithm is what made Google what it is

- Invented by Larry Page and Sergey Brin when graduate students at Stanford
- Now *each* worth about $17.5 billion
- That's a valuable algorithm!

Information on basics is public:

Some specific tuning is not public (SEO companies would love to know this!)



The PageRank Citation Ranking:
Bringing Order to the Web

January 29, 1998

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

1 Introduction and Motivation

The World Wide Web creates many new challenges for information retrieval. It is very large and heterogeneous. Current estimates are that there are over 150 million web pages with a doubling life of less than one year. More importantly, the web pages are extremely diverse, ranging from "What is Joe having for lunch today?" to journals about information retrieval. In addition to these major challenges, search engines on the Web must also contend with inexperienced users and pages engineered to manipulate search engine ranking functions.

How PageRank Works

Nice picture/scenario taken from Wikipedia
http://en.wikipedia.org/wiki/PageRank

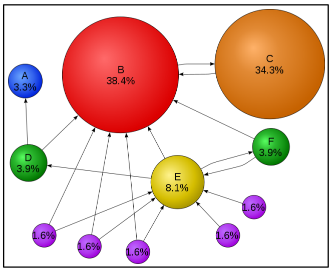
Basic idea:

Link structure indicates what is important (at least to web page creators)

Underlying idea: What's the probability that a random web surfer hits a page by randomly clicking links?

Key points:

- More likely to hit a node if linked by a likely node
- Don't know what's "likely" until after computing
- Iterative process



Manipulating PageRank: "Since December 2007, when it started actively penalizing sites selling paid text links, Google has combated link farms and other schemes designed to artificially inflate PageRank. How Google identifies link farms and other PageRank manipulation tools is among Google's trade secrets."

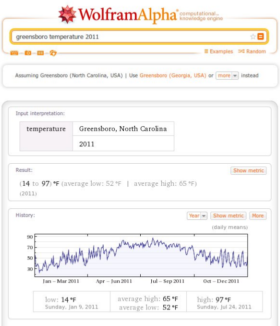
A different kind of search engine

WolframAlpha bills itself as a "computational knowledge engine"

Extracts ***information***, not just page copies

Can integrate information from different sources

http://www.wolframalpha.com



greensboro temperature 2011

Assuming Greensboro (North Carolina, USA) | Use Greensboro (Georgia, USA) or more instead

Input interpretation:

temperature	Greensboro, North Carolina
	2011

Result:

(14 to 97) °F (average low: 52 °F | average high: 65 °F)
(2011)

History:

(daily means)

low: 14 °F Sunday, Jan 9, 2011	average high: 65 °F average low: 52 °F	high: 97 °F Sunday, Jul 24, 2011