IT452 Advanced Web and Internet

Set 10
Search Engines & SEO

Outline

• How do search engines work?
  – Basic operation
  – What makes a good one?
  – What makes it difficult?

• Web Design with search engines in mind
Search Engines – Basic Operation

• Crawler

• Indexer

• Query Engine

Crawler

• How does it find the pages?

• Does it crawl everything?

• How fast does it crawl?
The Web is a Bow-Tie

- Early study of 200 million web pages and links
  - Broder et al. 2000
- Structure of the web: a bow-tie shape

Indexer

- Parse document
- Remember
  - Whole text
  - Words
  - Phrases
  - Link text
- Builds an “inverted index”

barista  531235, 4324, 6981, 125793, 41009, ...
burrito  344, 7173, 574527, 14513, 2451245, ...
burro    8375, 75346, 345231, 5123523, 52388, ...

Query Engine

- Process text query from user
- Inverse index merges document IDs
- Return *ranked* set of hopefully relevant pages
- Ranking factors
  - 1. Query-specific
  - 2. Page-specific
  - 3. Page Genre
  - 4.

PageRank

- Original basis of Google – still important
  - Developed in 1998.

- Basic Model
  \[
  R(w) = c \sum_{v \in B} \frac{R(v)}{|F_v|}
  \]

- Two interpretations:
  - Random walk
  - Pages voting
PageRank

- Two interpretations:
  - Random walk
  \[ R(w) = c \sum_{v \in B_w} \frac{R(v)}{|F_v|} \]
  - Pages voting

PageRank

- Who owns the PageRank patent?
  - (hint: not Google)
SEO

• Goal

• What does it consider?

• Types

SE0 0.1

• Early search engines heavily dependent on meta tags

• What to do?
  – White hat:
  – Black hat:

• Key issue: easy to _____________________
SEO 1.0

• Modern search engines depend heavily on links

• What to do?
  – White hat:
    – Black hat:

SEO 2.0

• Machine Learning
  – You search for “cats”, which result do you click first?
  – Learn from user clicks which they prefer
  – Smarter algorithms cluster words that “mean” the same thing

• What to do?
  – White hat:
    – Black hat:
Good principles

• Clear hierarchy
• Links to all pages (static), not as images
• Useful content
• Links from relevant sites
• Good title / alt / meta
• Limit dynamically generated pages (or # args)
• No broken links, < 100 links
• Use robots.txt – exclude internal search results
• Fresh content

Bad principles

• Stuff with lots of irrelevant content
• Show different version of content to crawler
• Link schemes, farms
• Hidden text and links
• Pages designed just for search engines, not users
• Automated querying
• Deception in general