The Four Dimensions of Search Engine Quality

Jan Pedersen
Chief Scientist, Yahoo! Search
19 September 2005
Outline

• The Search Landscape
• A Framework for Quality
  – RCFP
• Search Engine Architecture
• Detailed Issues
Search Landscape 2005

- Four major “Mainframes”
  - Google, Yahoo, MSN, and ASK
- >450M searches daily
  - 60% international
  - Thousands of machines
- $8+B in Paid Search Revenues
- Large indices
  - Billions of documents
  - Terrabytes of data
- Excellent relevance
  - For some tasks
What’s the Goal?

• User Satisfaction
  – Understand user intent
    • Problems: Ambiguity and Context
  – Generate Relevant matches
    • Problems: Scale and accuracy
  – Present Useful information
    • Problems: Ranking and Presentation
Quality Dimensions

- Ranking
  - Ability to rank hits by relevance
- Comprehensiveness
  - Index size and composition
- Freshness
  - Recency of indexed data
- Presentation
  - Titles and Abstracts
Search Engine Architecture

WWW

Crawl

Web Map

Snapshot

Indexer

Query

Serving

Web Index

Comprehensiveness and Freshness

Ranking and Presentation
Comprehensiveness

- **Problem:**
  - Make accessible all useful Web pages

- **Issues:**
  - Web has an infinite number of pages
  - Finite resources available
    - Bandwidth
    - Disk capacity

- **Selection Problem**
  - Which pages to visit
    - Crawl Policy
  - Which pages to index
    - Index Selection Policy
Crawl Policy

• Pages found by following links
  – From an initial root set

• Basic iteration:
  – Visit pages and extract links
  – Prioritize next pages to visit (or revisit)

• Framework
  – Visit pages
    • most likely to be viewed
    • most likely to contain links to pages that will be viewed
  – Prioritization by Query-independent Quality
• Problem:
  – Ensure that what is indexed correctly reflects current state of the web

• Impossible to achieve exactly
  – Revisit vs Discovery

• Divide and Conquer
  – A few pages change continually
  – Most pages are relatively static
Changing documents in daily crawl for 32-day period

![Graph showing changes in daily document crawling for 32 days. The x-axis represents the number of changes, ranging from 0 to 35. The y-axis represents the number of changes, ranging from 1 to 10,000. Two lines are plotted: one for HTML files and the other for out-links. The graph shows a trend with a peak at around the 32-day mark.](image-url)
Freshness on 5/17/2003

Source:
Search Engine Showdown
Ranking

• Problem:
  – Given a well-formed query, place the most relevant pages in the first few positions

• Issues:
  – Scale: Many candidate matches
    • Response in < 100 msecs
  – Evaluation:
    • Editorial
    • User Behavior
Query Serving Architecture

- Rectangular Array
  - Each row is a replicate
  - Each column is an index segment
- Results are merged across segments
  - Each node evaluates the query against its segment.
- Latency is determined by the performance of a single node
Editorial Relevance

- Users grade relevance
- Search Engines are scored in aggregate over a query sample
Clickrate Relevance Metric

Average highest rank clicked perceptibly increased with the release of a new rank function.
Ranking Framework

• Categorization problem
  – Estimate the probability of relevance given ranking features

• Query Dependent features
  – Term overlap between query and
    • Meta-data
    • Content

• Query Independent Features
  – Quality (e.g. Page Rank)
  – Spamminess
Handling Ambiguity

Results for query: Cobra
- Spelling Correction
- Also Try
- Short cuts
- Titles and Abstracts
Conclusions

• Search is a hard problem
  – Solutions are approximate
  – Measurement is difficult

• Search quality can be decomposed in separate but related problems
  – Ranking
  – Comprehensiveness
  – Freshness
  – Presentation