The Four Dimensions of Search Engine Quality

Jan Pedersen

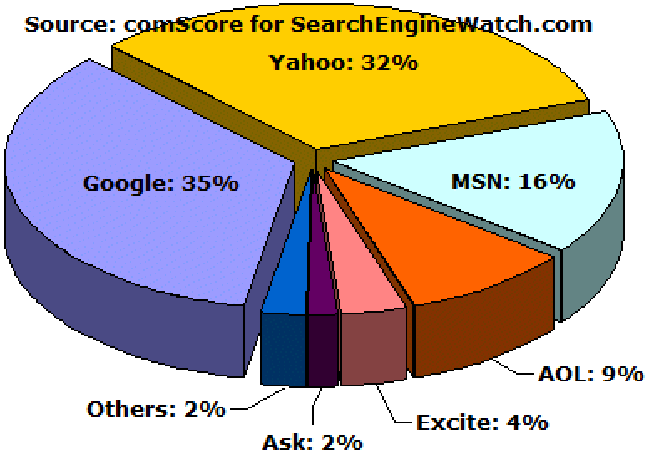
Chief Scientist, Yahoo! Search

19 September 2005

Outline

- The Search Landscape
- A Framework for Quality
  - RCFP
- Search Engine Architecture
- Detailed Issues

Search Landscape 2005



Source: Search Engine Watch

- 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

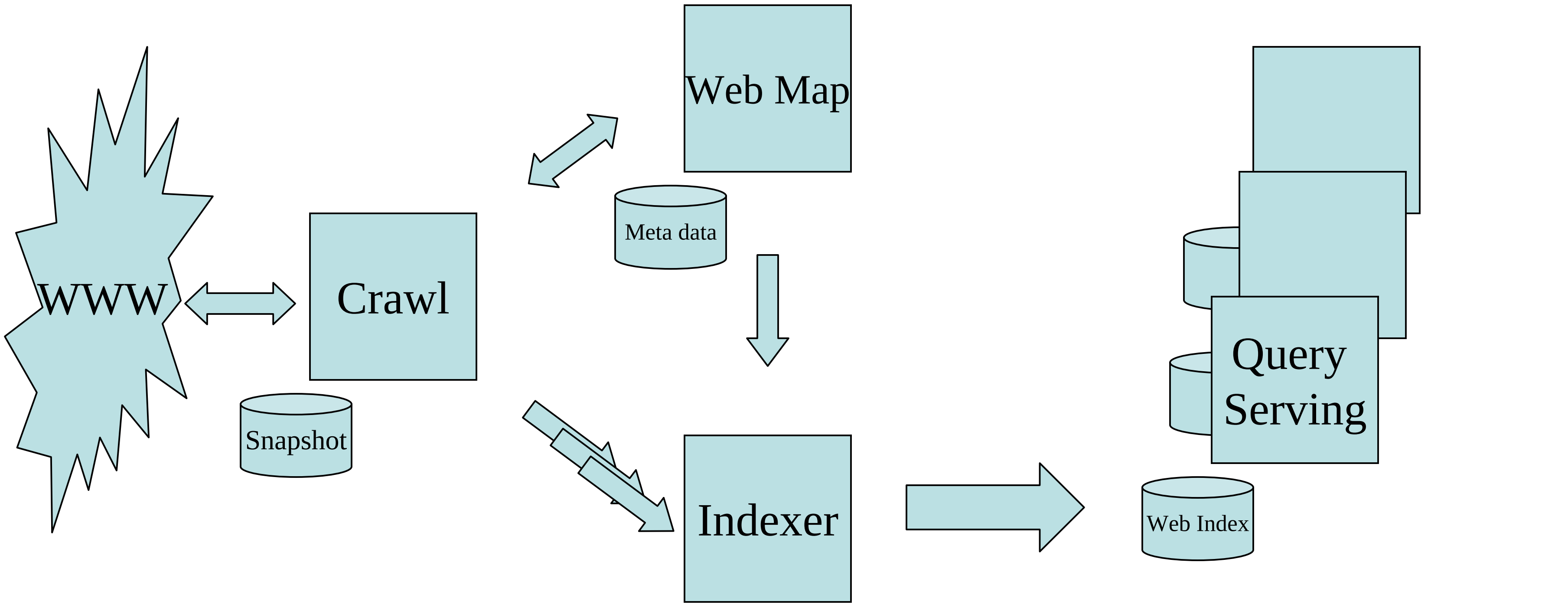
What’s the Goal?

- 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



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

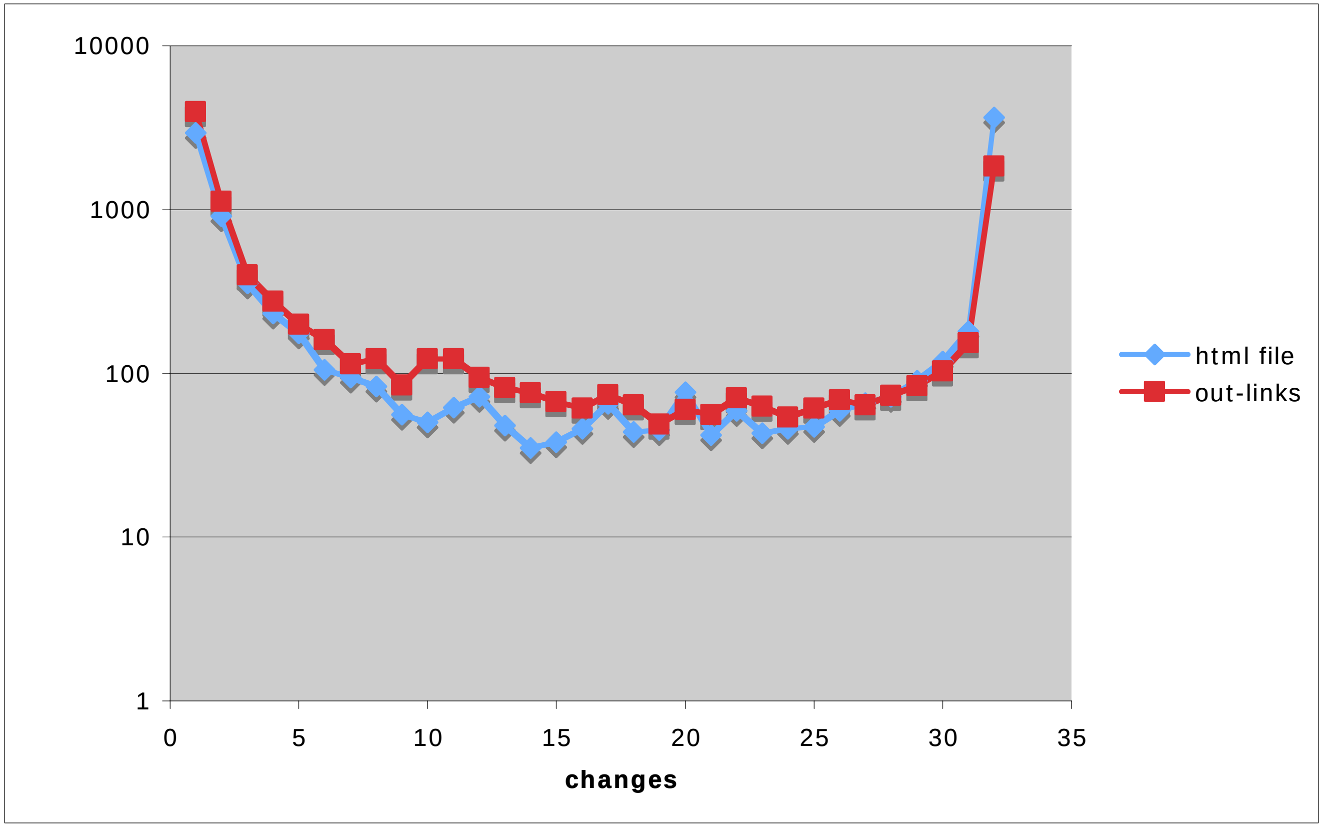
Crawl Policy

- 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

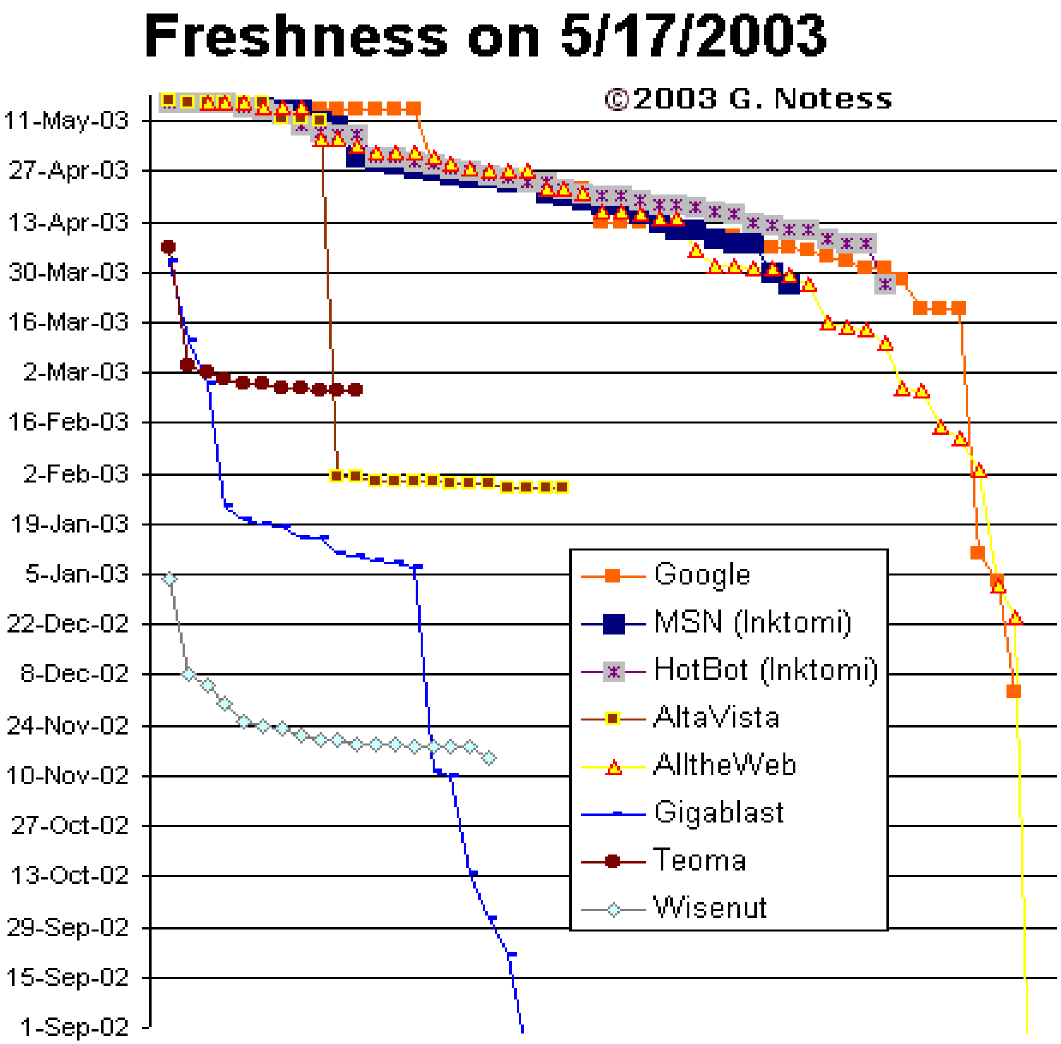
Freshness

- 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



Freshness



Source:

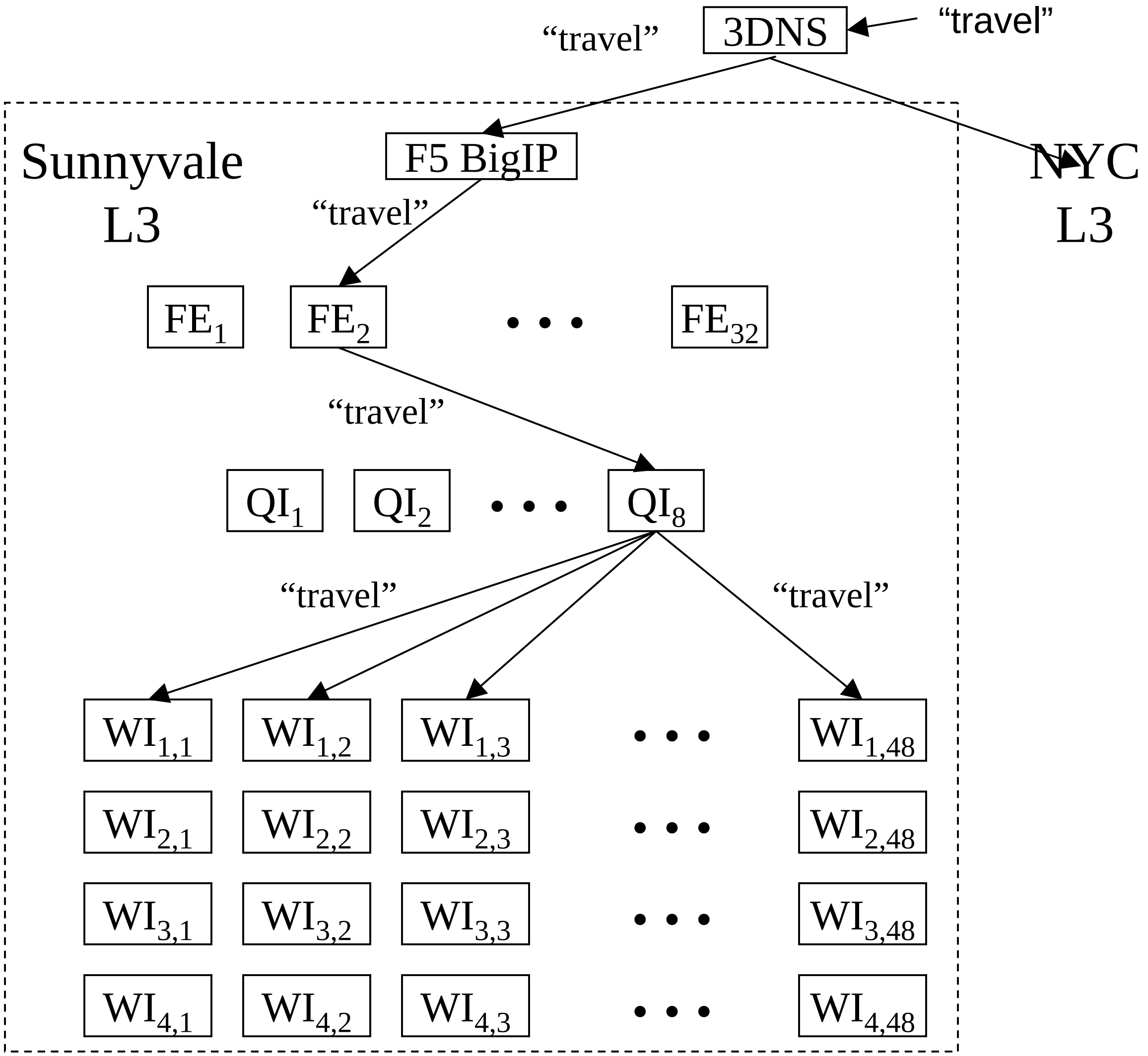
Search Engine Showdown

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



“travel”

3DNS

“travel”

Sunnyvale L3

F5 BigIP

NYC L3

“travel”

FE_1 FE_2 • • • FE_{32}

“travel”

QI_1 QI_2 • • • QI_8

“travel”

“travel”

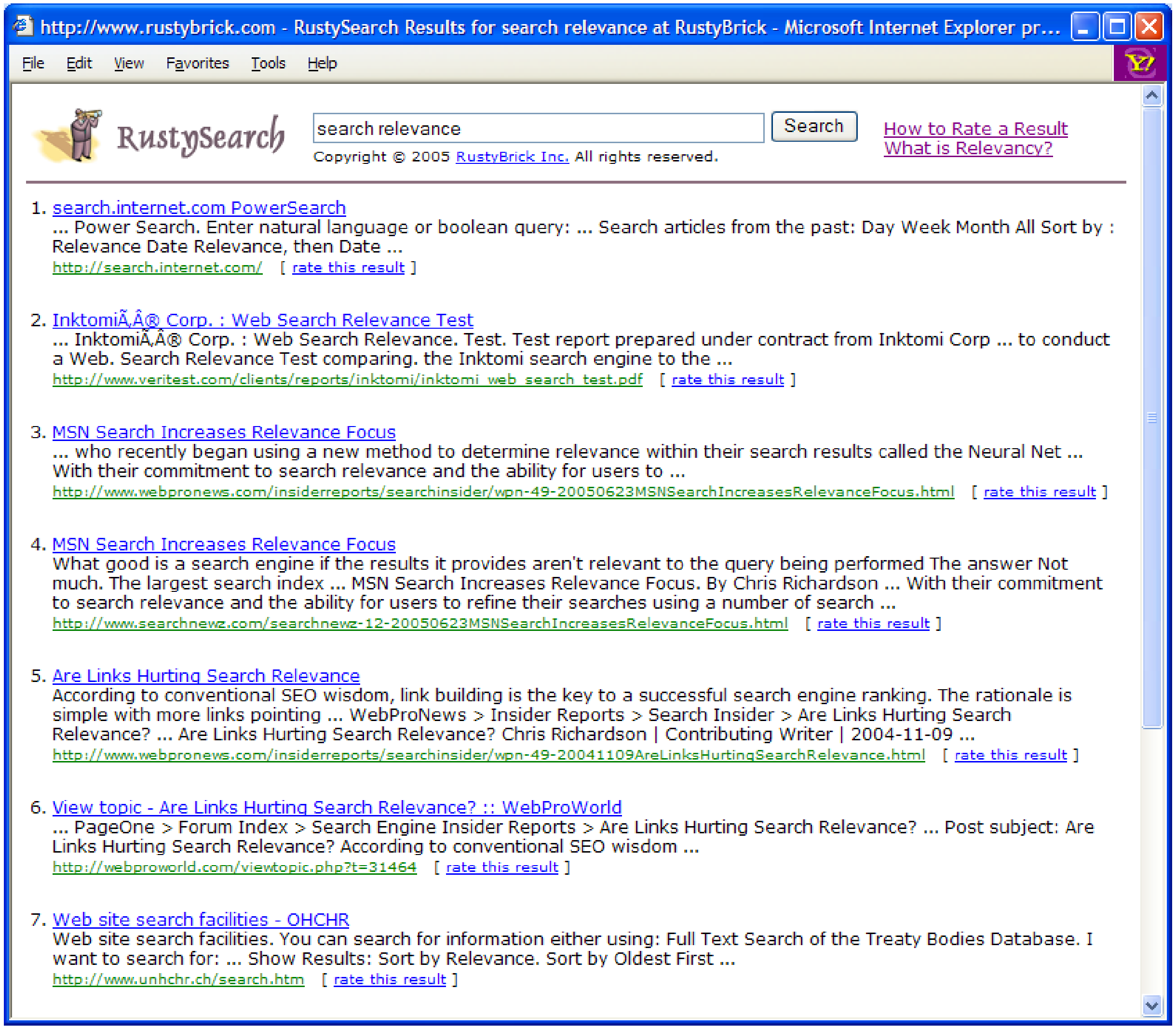
$WI_{1,1}$ $WI_{1,2}$ $WI_{1,3}$ • • • $WI_{1,48}$

$WI_{2,1}$ $WI_{2,2}$ $WI_{2,3}$ • • • $WI_{2,48}$

$WI_{3,1}$ $WI_{3,2}$ $WI_{3,3}$ • • • $WI_{3,48}$

$WI_{4,1}$ $WI_{4,2}$ $WI_{4,3}$ • • • $WI_{4,48}$

Editorial Relevance


http://www.rustybrick.com - RustySearch Results for search relevance at RustyBrick - Microsoft Internet Explorer pr...
File
Edit
View
Favorites
Tools
Help
RustySearch
Search
How to Rate a Result
What is Relevancy?

search relevance

Copyright © 2005 RustyBrick Inc. All rights reserved.

1. search.internet.com PowerSearch
   ... Power Search. Enter natural language or boolean query: ... Search articles from the past: Day Week Month All Sort by : Relevance Date Relevance, then Date ...
   http://search.internet.com/ [rate this result]

2. InktomiÃ‚Â® Corp. : Web Search Relevance Test
   ... InktomiÃ‚Â® Corp. : Web Search Relevance. Test. Test report prepared under contract from Inktomi Corp ... to conduct a Web. Search Relevance Test comparing. the Inktomi search engine to the ...
   http://www.veritest.com/clients/reports/inktomi/inktomi_web_search_test.pdf [rate this result]

3. MSN Search Increases Relevance Focus
   ... who recently began using a new method to determine relevance within their search results called the Neural Net ... With their commitment to search relevance and the ability for users to ...
   http://www.webpronews.com/insiderreports/searchinsider/wpn-49-20050623MSNSearchIncreasesRelevanceFocus.html [rate this result]

4. MSN Search Increases Relevance Focus
   What good is a search engine if the results it provides aren't relevant to the query being performed The answer Not much. The largest search index ... MSN Search Increases Relevance Focus. By Chris Richardson ... With their commitment to search relevance and the ability for users to refine their searches using a number of search ...
   http://www.searchnewz.com/searchnewz-12-20050623MSNSearchIncreasesRelevanceFocus.html [rate this result]

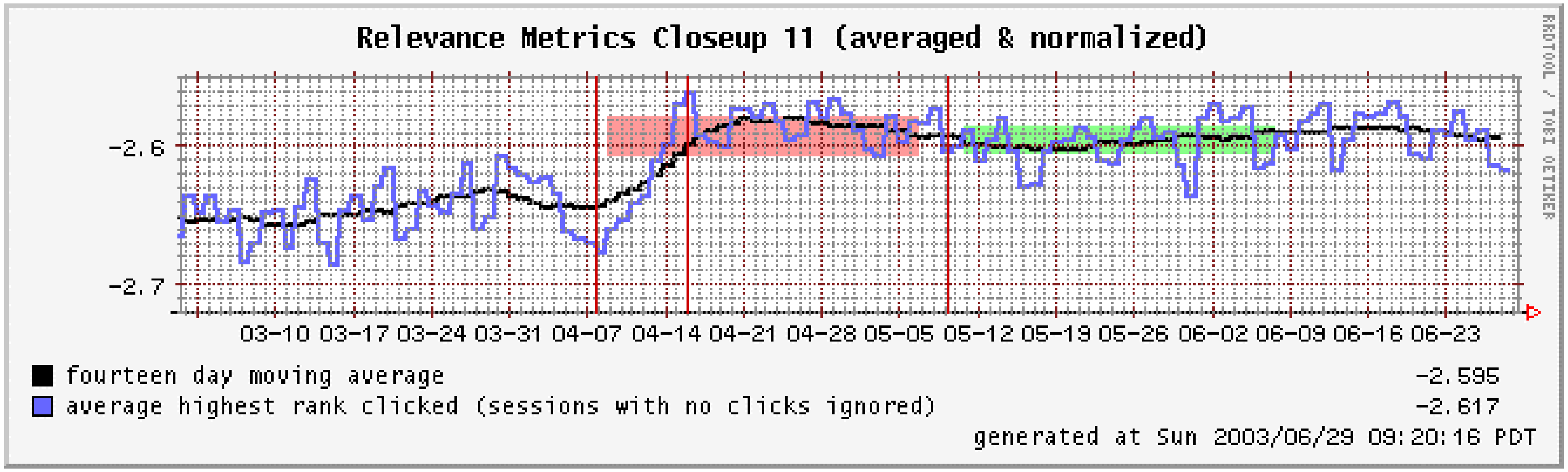
5. Are Links Hurting Search Relevance
   According to conventional SEO wisdom, link building is the key to a successful search engine ranking. The rationale is simple with more links pointing ... WebProNews > Insider Reports > Search Insider > Are Links Hurting Search Relevance? ... Are Links Hurting Search Relevance? Chris Richardson | Contributing Writer | 2004-11-09 ...
   http://www.webpronews.com/insiderreports/searchinsider/wpn-49-20041109AreLinksHurtingSearchRelevance.html [rate this result]

6. View topic - Are Links Hurting Search Relevance? :: WebProWorld
   ... PageOne > Forum Index > Search Engine Insider Reports > Are Links Hurting Search Relevance? ... Post subject: Are Links Hurting Search Relevance? According to conventional SEO wisdom ...
   http://webproworld.com/viewtopic.php?t=31464 [rate this result]

7. Web site search facilities - OHCHR
   Web site search facilities. You can search for information either using: Full Text Search of the Treaty Bodies Database. I want to search for: ... Show Results: Sort by Relevance. Sort by Oldest First ...
   http://www.unhchr.ch/search.htm [rate this result]

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

Clickrate Relevance Metric

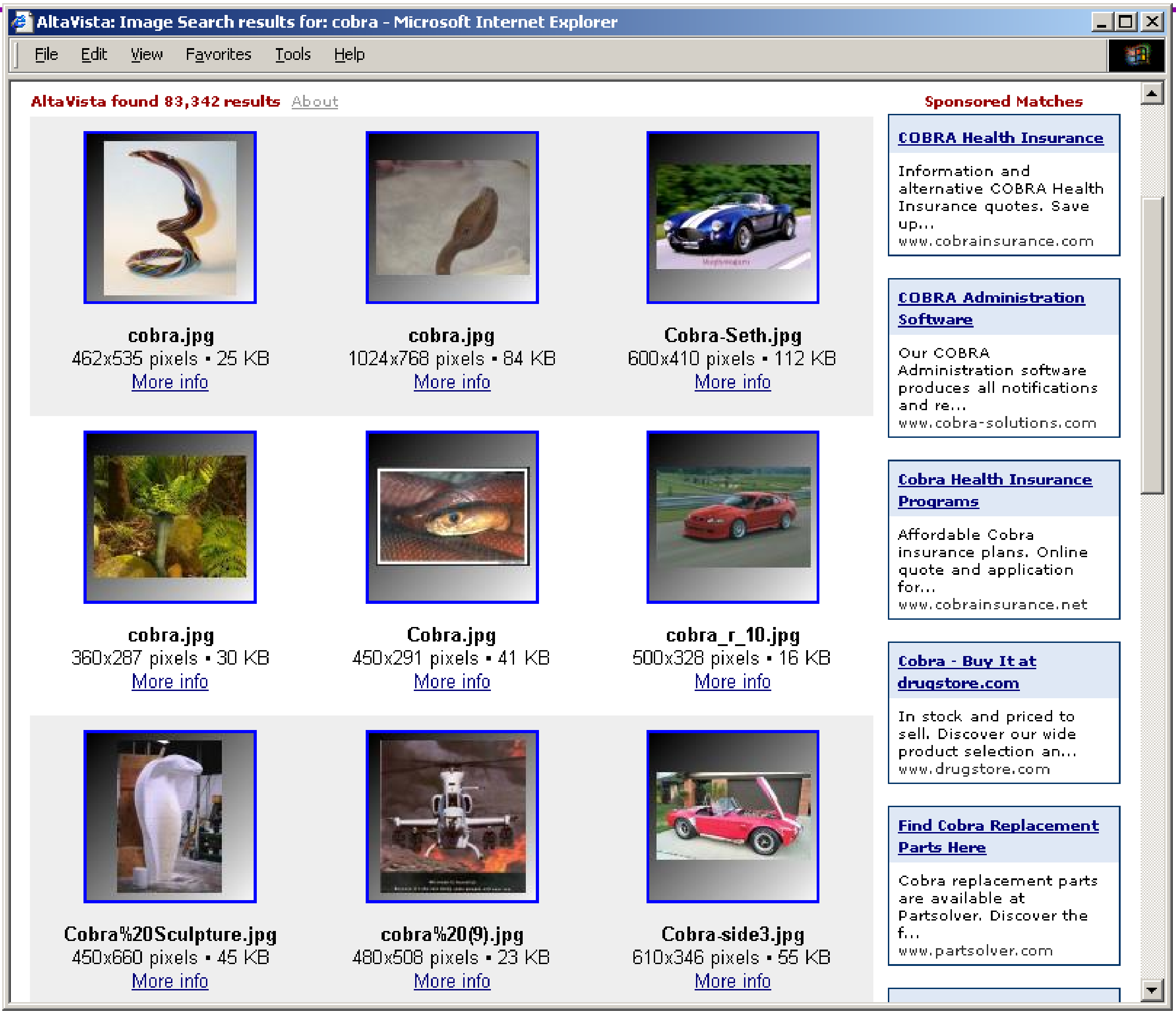


Average highest rank clicked perceptibly increased with the release of a new rank function.

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


AltaVista: Image Search results for: cobra - Microsoft Internet Explorer
File Edit View Favorites Tools Help

AltaVista found 83,342 results About

cobra.jpg
462x535 pixels • 25 KB
More info

cobra.jpg
1024x768 pixels • 84 KB
More info

Cobra-Seth.jpg
600x410 pixels • 112 KB
More info

cobra.jpg
360x287 pixels • 30 KB
More info

Cobra.jpg
450x291 pixels • 41 KB
More info

cobra_r_10.jpg
500x328 pixels • 16 KB
More info

Cobra%20Sculpture.jpg
450x660 pixels • 45 KB
More info

cobra%20(9).jpg
480x508 pixels • 23 KB
More info

Cobra-side3.jpg
610x346 pixels • 55 KB
More info

Sponsored Matches

COBRA Health Insurance
Information and alternative COBRA Health Insurance quotes. Save up...
www.cobrainsurance.com

COBRA Administration Software
Our COBRA Administration software produces all notifications and re...
www.cobra-solutions.com

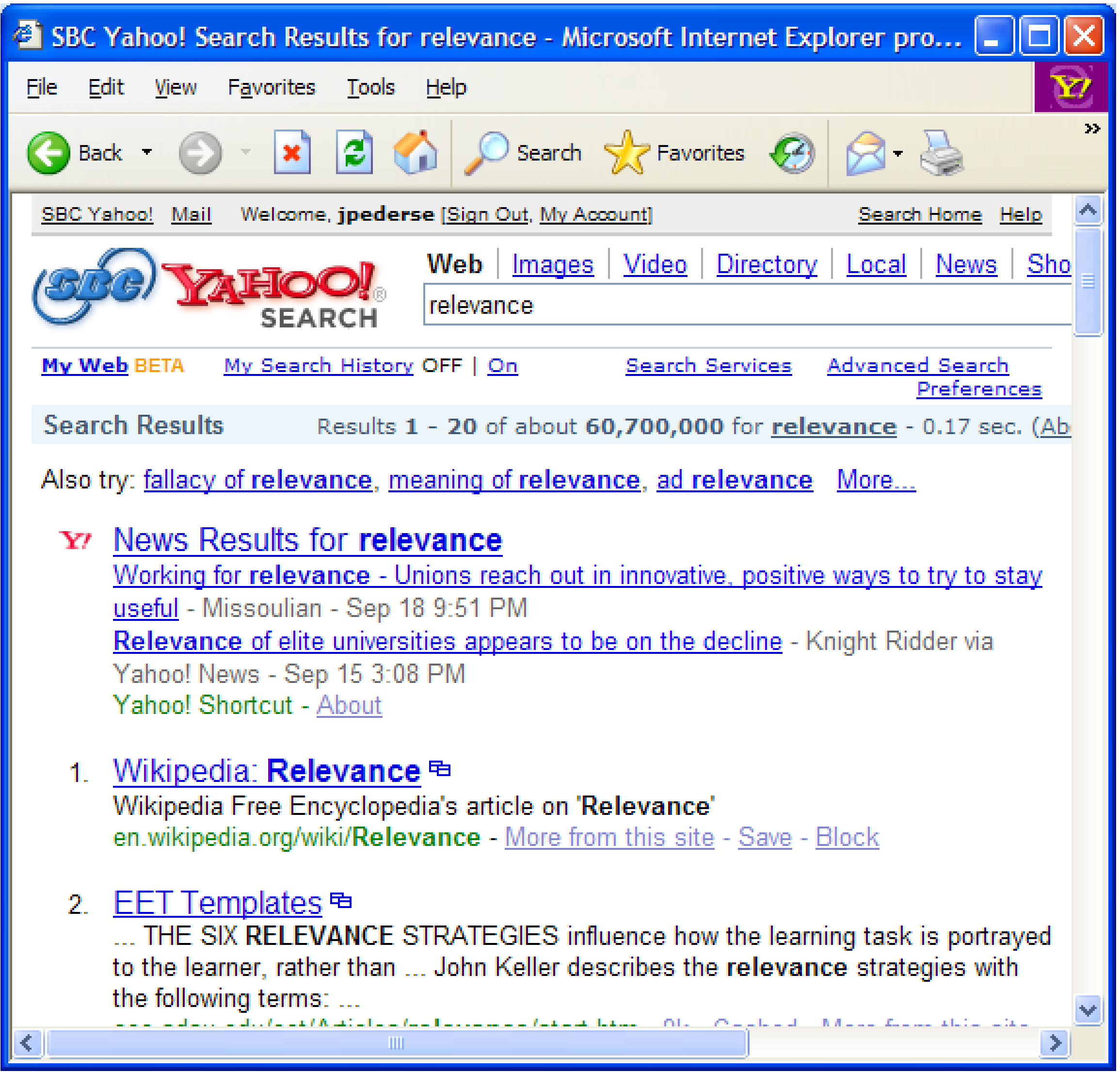
Cobra Health Insurance Programs
Affordable Cobra insurance plans. Online quote and application for...
www.cobrainsurance.net

Cobra - Buy It at drugstore.com
In stock and priced to sell. Discover our wide product selection an...
www.drugstore.com

Find Cobra Replacement Parts Here
Cobra replacement parts are available at Partsolver. Discover the f...
www.partsolver.com

Results for query: Cobra

Presentation


SBC Yahoo! Search Results for relevance - Microsoft Internet Explorer pro...
File Edit View Favorites Tools Help
Back
Search
Favorites
SBC Yahoo! Mail
Search Home Help
Web | Images | Video | Directory | Local | News | Sho
SEARCH
My Web BETA
My Search History OFF | On
Search Services
Advanced Search
Preferences

Welcome, **jpederse** [Sign Out, My Account]

relevance

Search Results Results **1 - 20** of about **60,700,000** for **relevance** - 0.17 sec. (Ab

Also try: fallacy of **relevance**, meaning of **relevance**, ad **relevance** More...

News Results for **relevance**

Working for **relevance** - Unions reach out in innovative, positive ways to try to stay useful - Missoulian - Sep 18 9:51 PM

Relevance of elite universities appears to be on the decline - Knight Ridder via Yahoo! News - Sep 15 3:08 PM

Yahoo! Shortcut - About

1. Wikipedia: **Relevance**
   Wikipedia Free Encyclopedia's article on '**Relevance**'
   en.wikipedia.org/wiki/**Relevance** - More from this site - Save - Block

2. EET Templates
   ... THE SIX **RELEVANCE** STRATEGIES influence how the learning task is portrayed to the learner, rather than ... John Keller describes the **relevance** strategies with the following terms: ...

- 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