



Search Quality

Jan Pedersen

10 September 2007

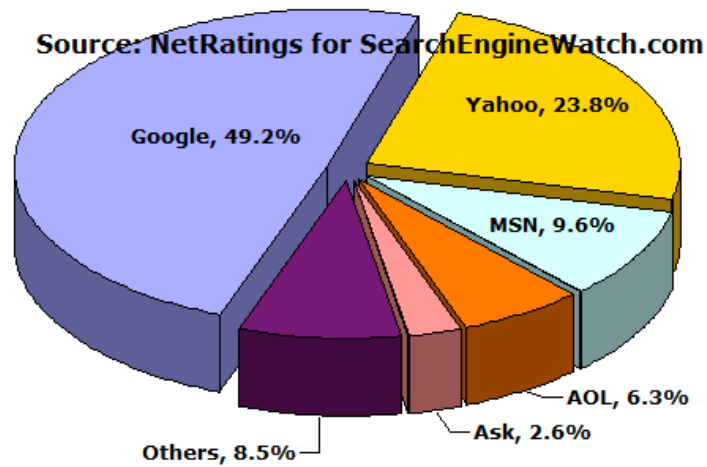


Outline

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



Search Landscape 2007



Source: Search Engine Watch:
US web search share, July 2006

- Three major “Mainframes”
 - Google, Yahoo, and MSN
- >800M searches daily
 - 60% international
 - 10^6 machines
- \$20B in Paid Search Revenues
- Large indices
 - Billions of documents
 - Petabytes of data

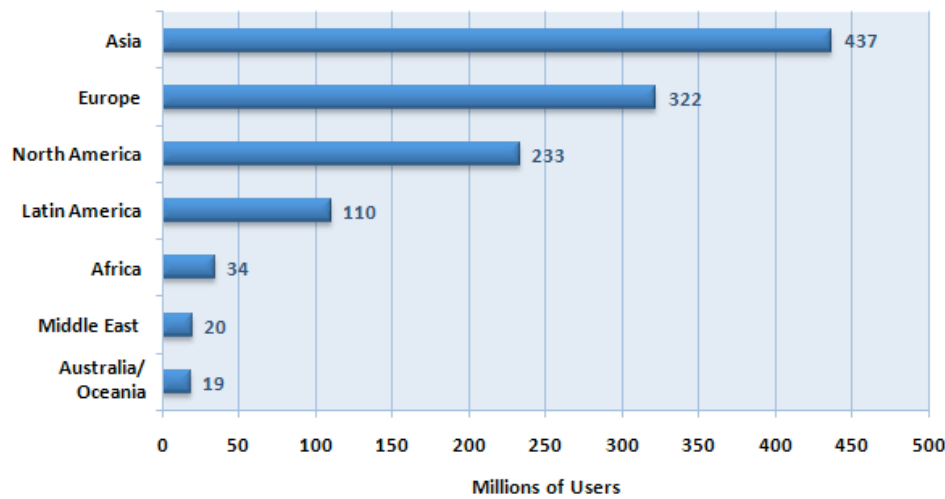
- **Key Drivers:**

Scale, Quality, Distribution



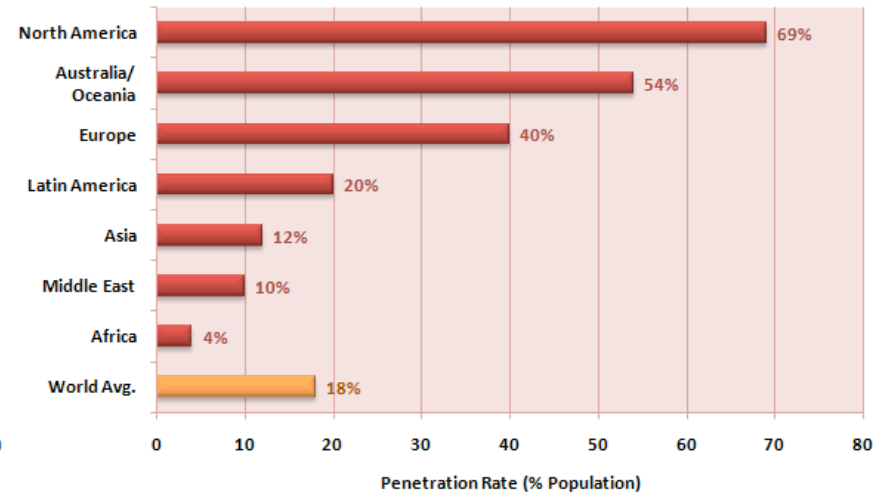
World Internet Usage

Internet Usage by World Region



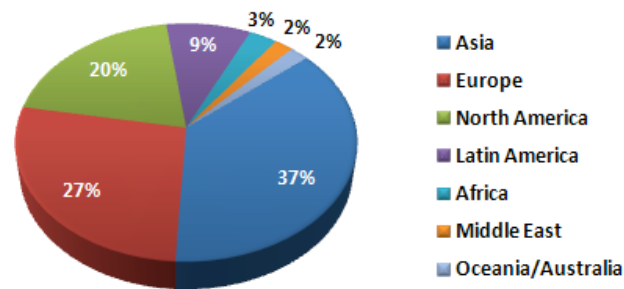
Copyright © 2007, www.internetworldstats.com

Internet Penetration by World Region



Copyright © 2007, www.internetworldstats.com

World Internet Users



Copyright © June 2007, www.internetworldstats.com



Search Results page

The screenshot shows a Yahoo! search results page for the query "great wall". The browser window title is "great wall - AT&T Yahoo! Search Results - M...". The search bar contains "great wall" and the search button is labeled "Search". The page displays search results for "great wall" with 1-10 of about 225,000,000 results in 0.22 seconds. The results are categorized into "Search Assists" (lightbulb icon), "Related Results" (Flickr photos), and "Ranked Results" (numbered list). The "Search Assists" section suggests "great wall of china" and "history of great wall of china". The "Related Results" section shows four Flickr photos of the Great Wall and a link to "More Flickr photos". The "Ranked Results" section lists five search results, including "Great Wall of China, Beijing Tour Packages, Great Wall Hiking, China ...", "The Great Wall of China - Wikipedia", "Crystalinks.com: The Great Wall of China", "Discovering the Great Wall and Ming Tombs", and "Great Wall of China - Enchanted Learning Software". The "Search Ads" section on the right contains several sponsored results, including "Great Wall Hotel", "Great Wall Travel Guide", "The Great Wall at Amazon.com", and "Great Wall Sheraton Hotel".

Text Input

Search Assists

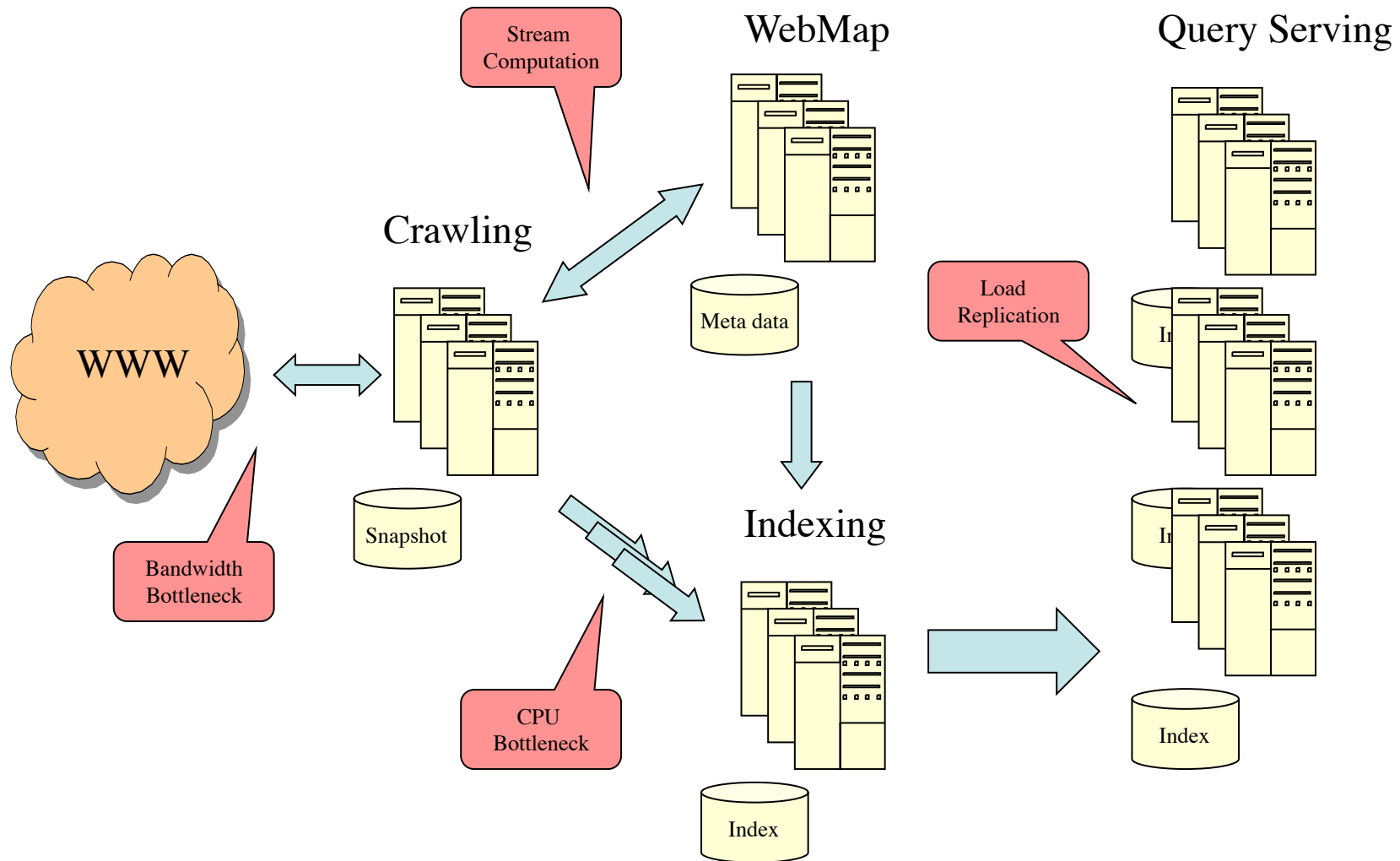
Related Results

Ranked Results

Search Ads



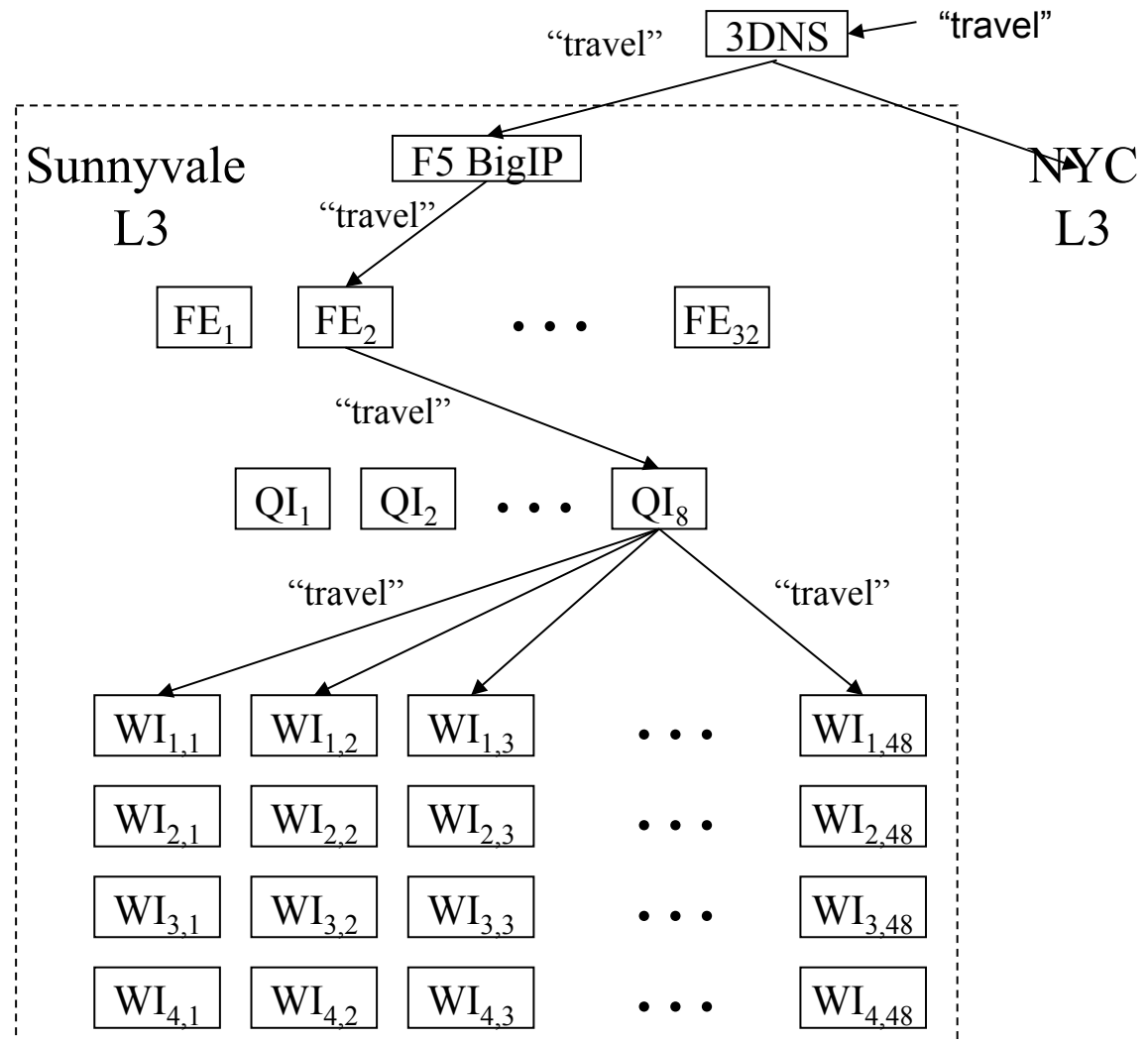
Search Engine Architecture





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





The Factory Floor

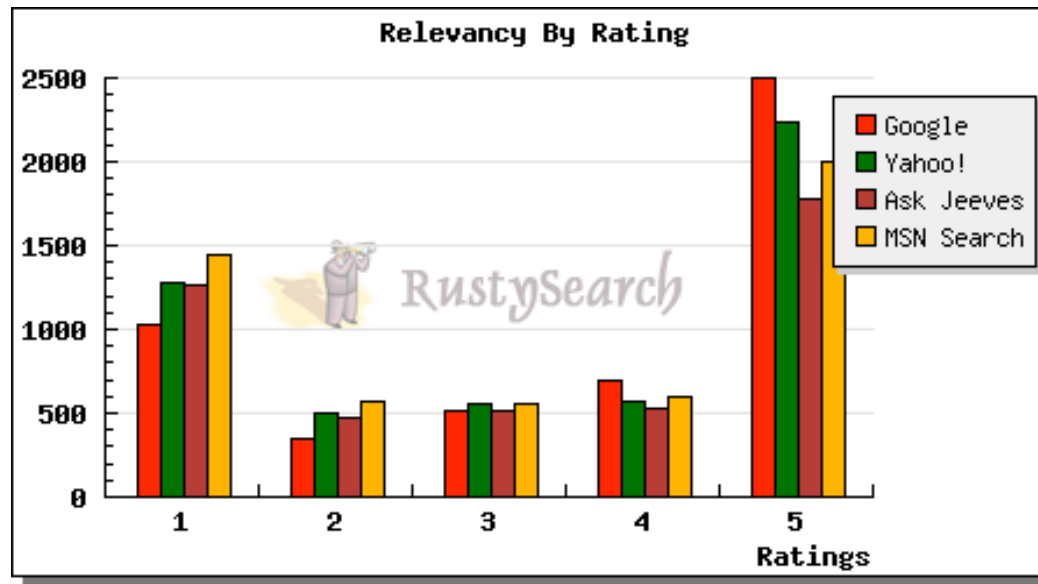


What's the Goal?

- User Satisfaction
 - Understand user intent
 - Problems: Ambiguity and Context
 - Generate relevant matches
 - Problems: Scale and accuracy
 - Present useful information
 - Problems: Ranking and Presentation



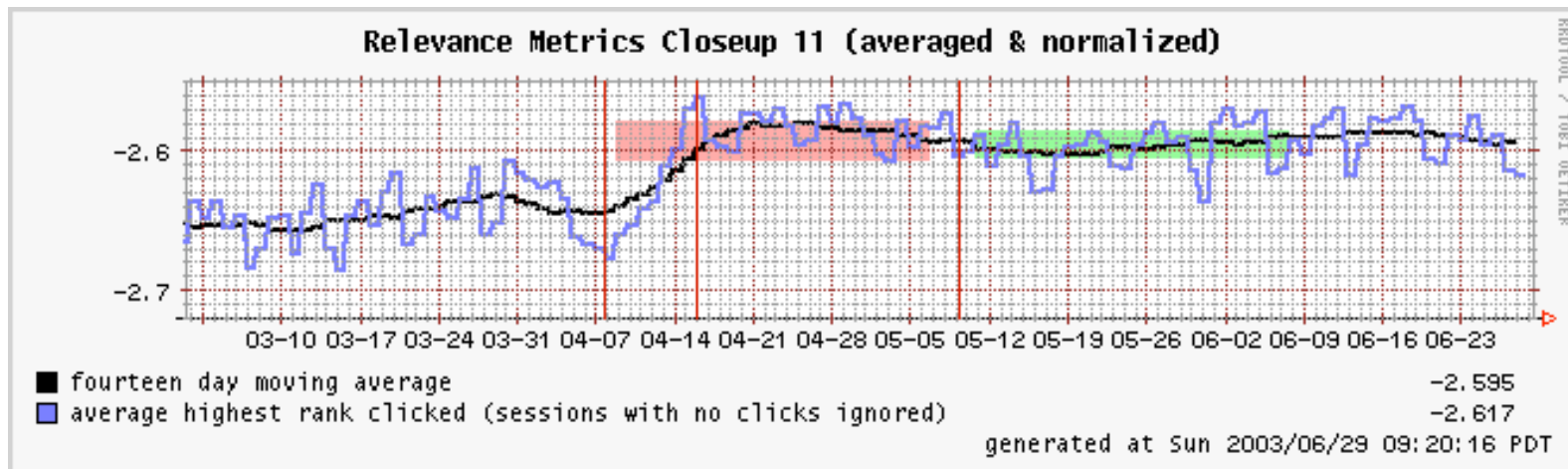
Evaluation



- Graded Relevance score
- Editorial Assessment
- Session/Task fulfillment?
- Behavioral measures?



Clickrate Relevance Metric



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



Quality Dimensions

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

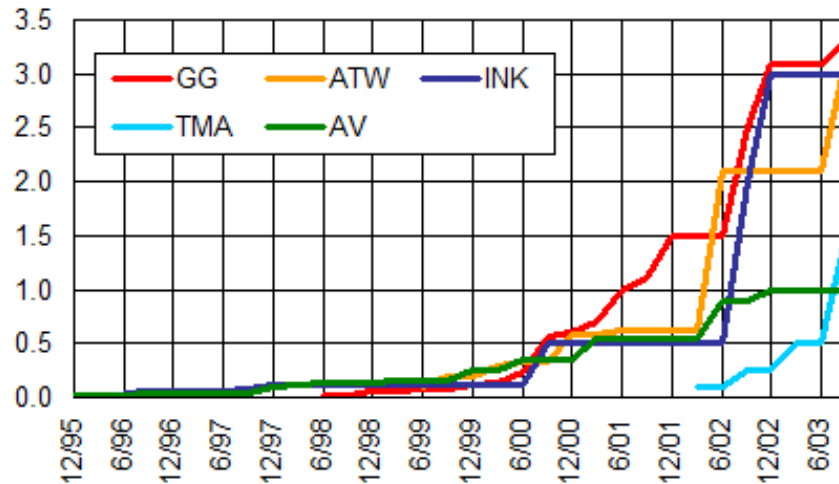


Comprehensiveness

- 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



Moore's Law and Index Size



Source: Search Engine Watch

- ~150M in 1998
- ~5B in 2005
 - 33x increase
 - Moore would predict 25x
- What about 2010?
 - 40B?

- 1994 Yahoo (directory) and Lycos (index) go public
- 1995 Infoseek and Excite go public
- 1997 Alta Vista launches 100M index
- 1998 Inktomi and Google launched
- 1999 All The Web launched
- 2003 Yahoo purchases Inktomi and Overture
- 2004 Google goes public
- 2005 Msft launches MS Live

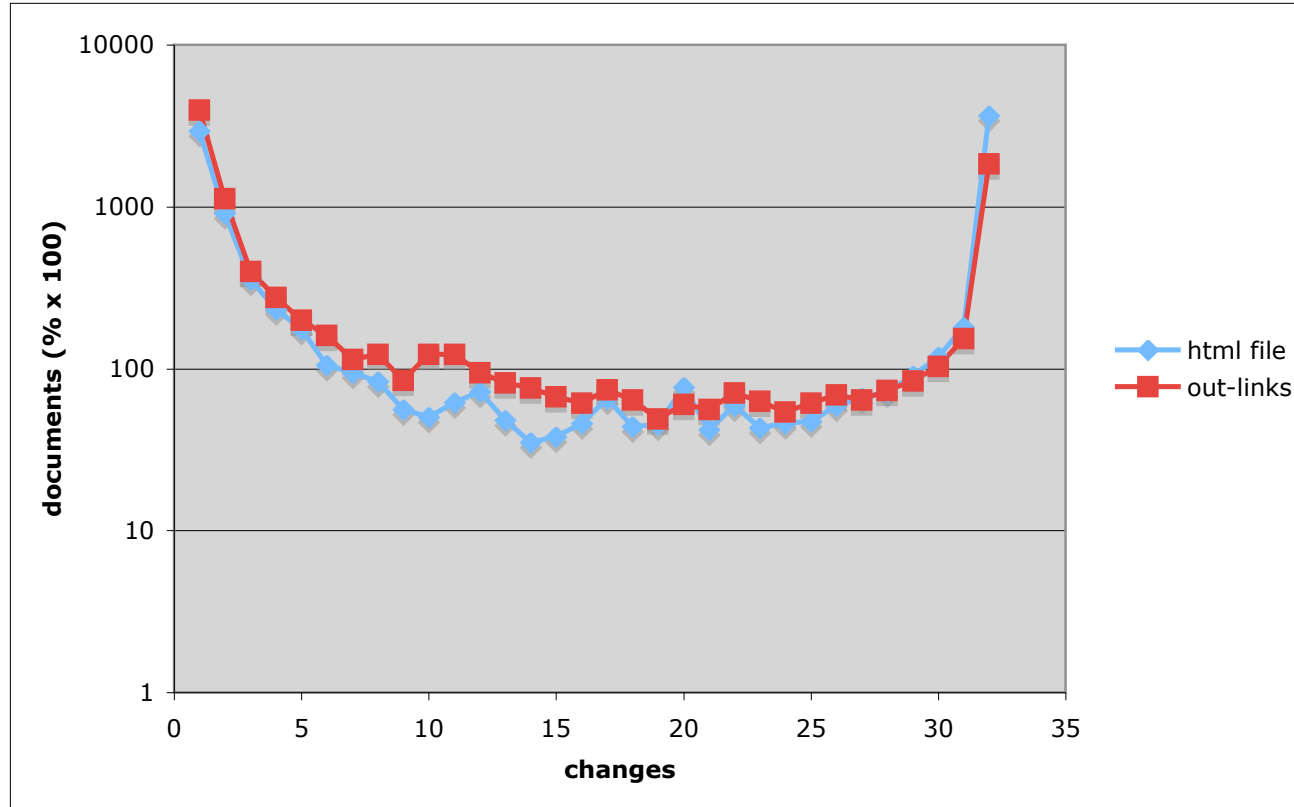


Freshness

- 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



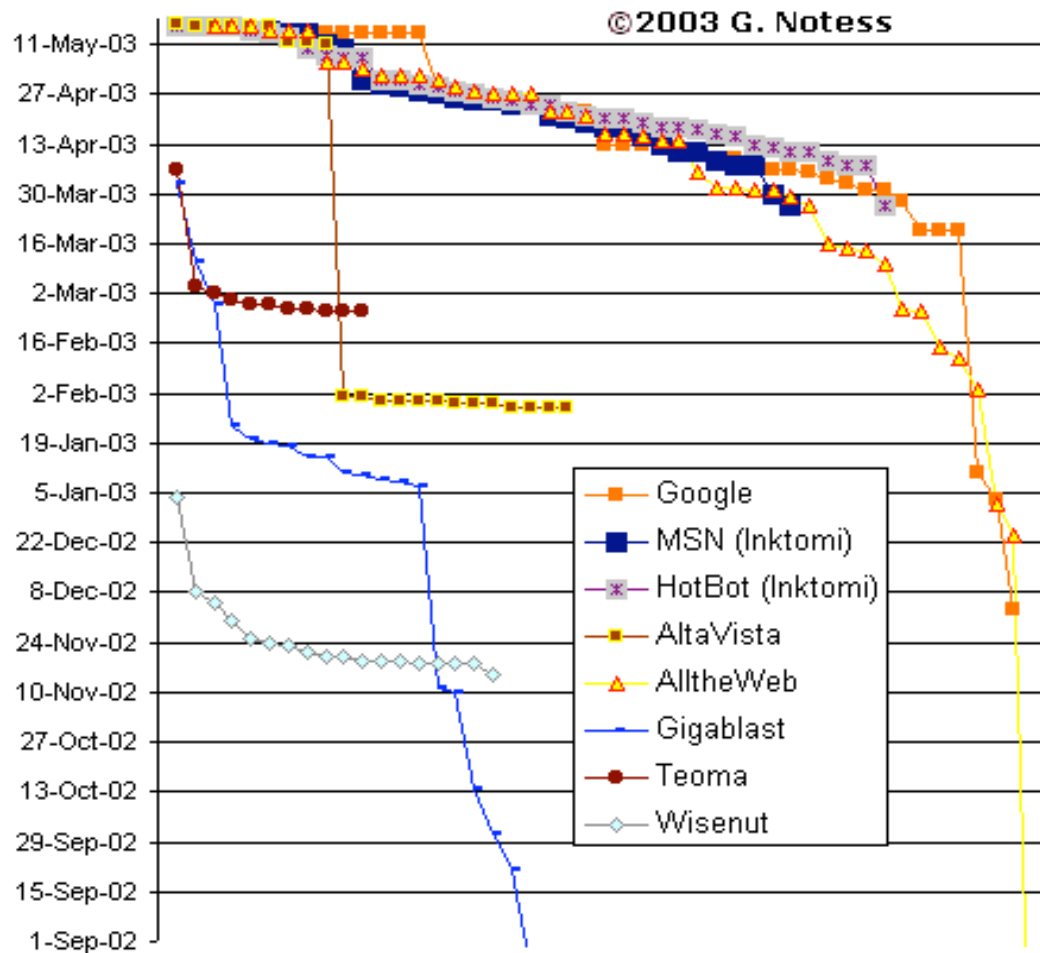
Changing documents in daily crawl for 32-day period





Freshness

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



Ranking Framework

- Regression problem
 - Estimate editorial relevance given ranking features
- Query Dependent features
 - Term overlap between query and
 - Meta-data
 - Content
- Query Independent Features
 - Quality (e.g. Page Rank)
 - Spamminess



Machine Learned Ranking

- **Goal: Automatically construct a ranking function**
 - Input:
 - Large number training examples
 - Features that predict relevance
 - Relevance metrics
 - Output:
 - Ranking function

- **Enables rapid experimental cycle**
 - Scientific investigation of
 - Modifications to existing features
 - New feature

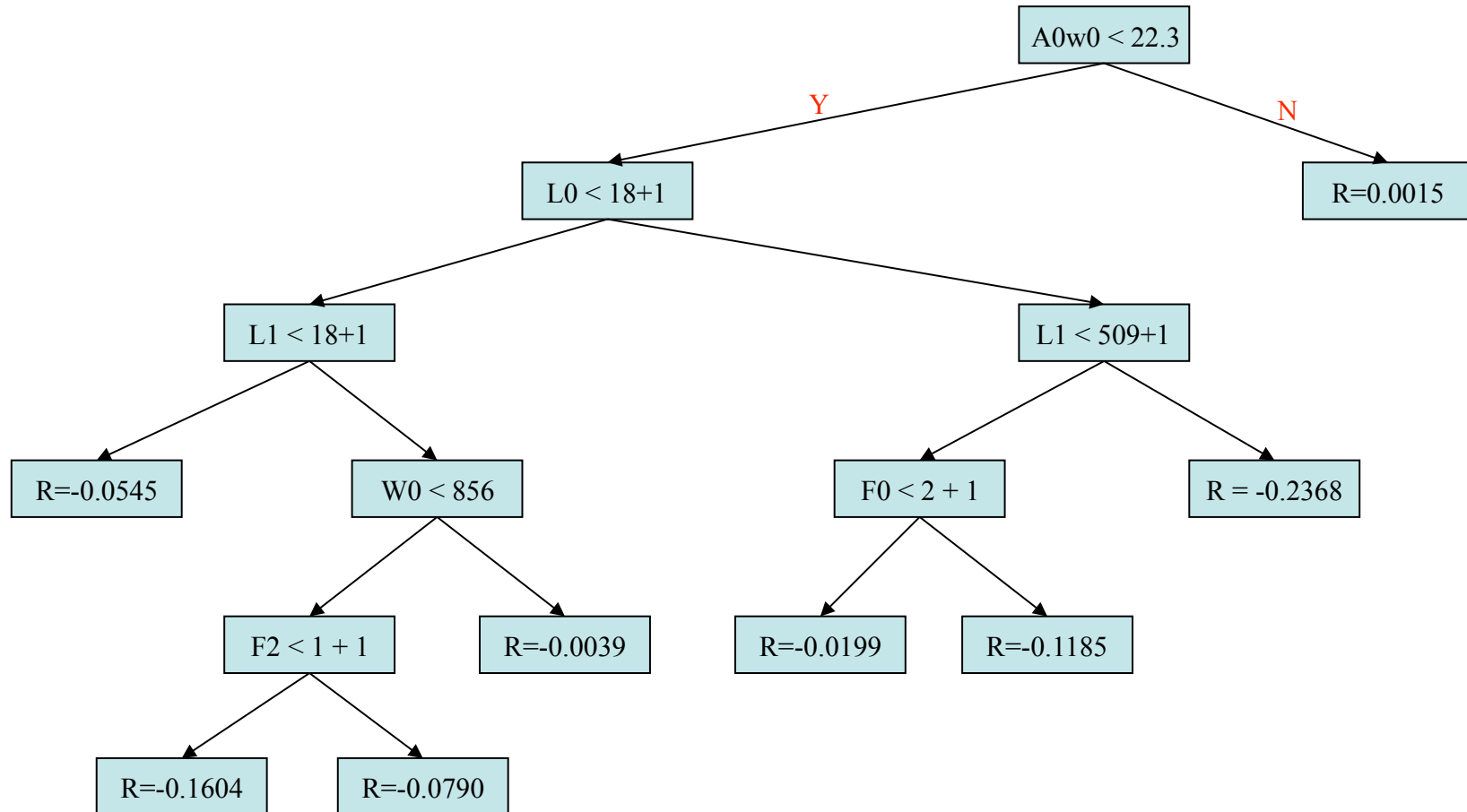


Ranking Features

- A0 - A4 anchor text score per term
- W0 - W4 term weights
- L0 - L4 first occurrence location
(encodes hostname and title match)
- SP spam index: logistic regression of 85 spam filter variables
(against relevance scores)
- F0 - F4 term occurrence frequency within document
- DCLN document length (tokens)
- ER Eigenrank
- HB Extra-host unique inlink count
- ERHB $ER \cdot HB$
- A0W0 etc. $A0 \cdot W0$
- QA Site factor –
logistic regression of 5 site link and url count ratios
- SPN Proximity
- FF family friendly rating
- UD url depth

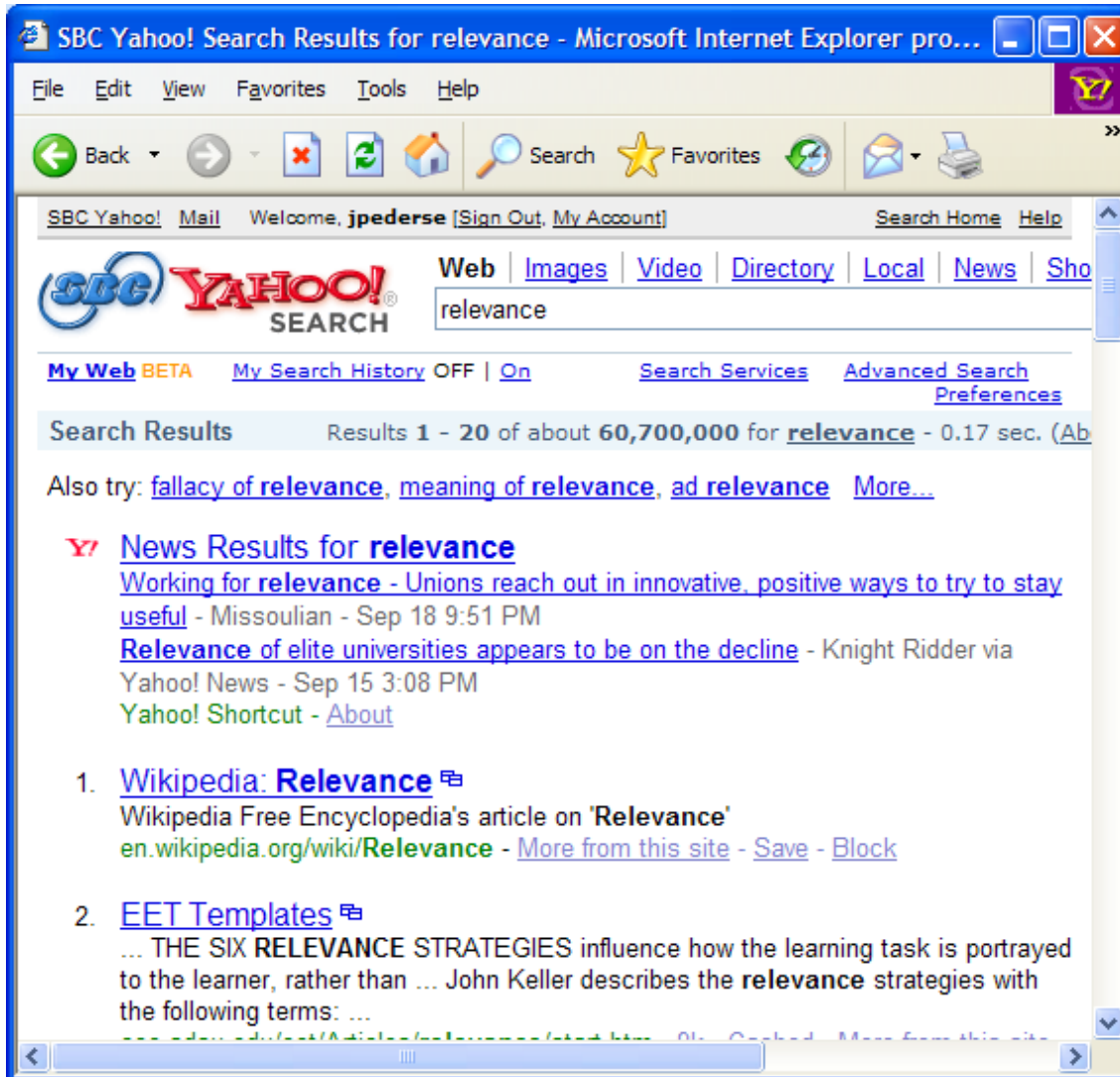


Implements (Tree 0)





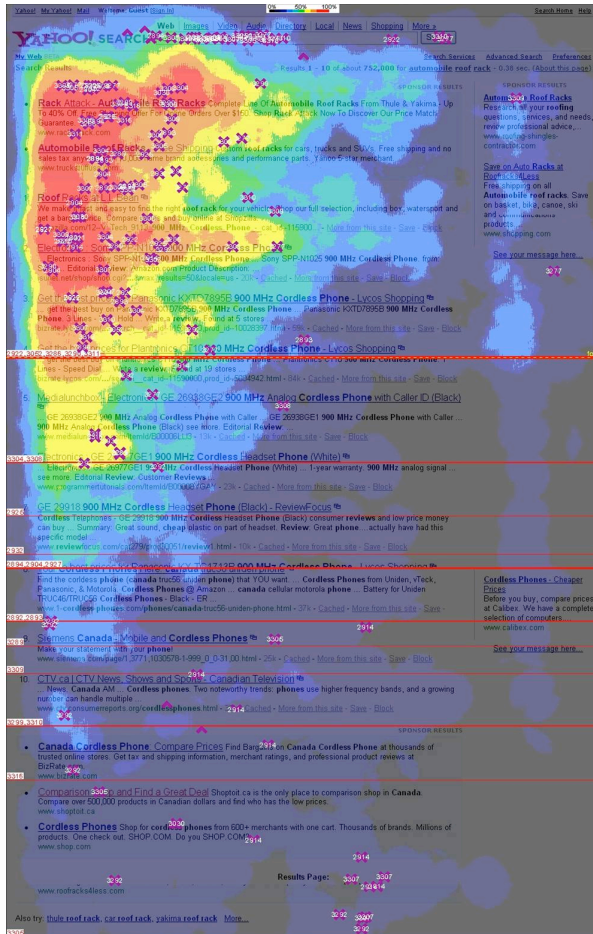
Presentation



- Spelling Correction
- Also Try
- Short cuts
- Titles and Abstracts



Eye Tracking Studies



- Golden Triangle
 - Top left corner
- Quick scan
 - For candidate
- Longer scan
 - For relevance



Comparison to State-of-the-art

Web search results 1 - 10 of 309 results most relevant to +darter +habitat

[Next 10 >](#) | [Hide summaries](#) | [Sort by date](#) | [Ungroup results](#)

[Microhabitat Use in a Diverse Assemblage of Darters ...](#)

Microhabitat Use In A Diverse Assemblage Of **Darters** In The Elk River Drainage, West Virginia. A. Welsh, Ph.D. Sue A. Perry Rita Villella 1996 - 1997 May 1997 National Biological Service, U.S. Science Center; W.V. Division of Natural Resources 1. Quantify microhabitat use for ... 92% Date: 9 Jan 1998, Size 3.9K, <http://www.caf.wvu.edu/coop/elkneck.html>
[Find similar pages](#) | [Grouped results from www.caf.wvu.edu](#)

[BAYOU DARTER, *Etheostoma \(Nothonotus\) rubrum* U.S. Fish & Wildlife Service](#)

Source: FWS Region 4 -- As of 2/91 Percidae Threatened throughout its range, , September 2. A diminutive species, the Bayou **darter** reaches a maximum length of about 1.8 ... 91% Date: 13 Apr 1998, Size 5.1K, <http://www.fws.gov/r9endspp/i/e/sae13.html>
[Find similar pages](#) | [Grouped results from www.fws.gov](#)

[CWT White-faced Darter Biodiversity Action Plan \(*Leucorrhinia dubia* \(Van der Linden\)\)](#)

Cheshire Wildlife Trust White-faced **Darter** Biodiversity Action Plan (*Leucorrhinia dubia* (Van der Linden)) Web Pages 1997. 84% Date: 5 Jan 1998, Size 5.5K, <http://www.talk-101.com/users/cwt/WfDBAP.htm>
[Find similar pages](#) | [Grouped results from www.talk-101.com](#)

[MDA Pesticide Information Sheet](#)

Address Maryland Department of Agriculture Pesticide Regulation Section 50 Harry S. Truman Telephone: (410) 841-5710 Fax: (410) 841-2765 Send E-mail to Dennis Howard ... 83% Date: 15 Nov 1998, Size 6.1K, <http://www.mda.state.md.us/plant/species.htm>
[Find similar pages](#)

[WRCF - Sand Darter](#)

Photo Credit: Rob Criswell IDENTIFYING CHARACTERISTICS: The sand **darter** is a small member averaging 2 1/2 inches in length. Adults are pale yellow above and silvery below, with a row of

Google Search: darter habitat - Microsoft Internet Explorer provided by Yahoo!

File Edit View Favorites Tools Help

Web Results 1 - 10 of about 43,200 for [darter habitat](#) (0.40 seconds)

[NIANGUA DARTER](#)
... Measures taken to stabilize and improve Niangua **darter habitat** will also benefit ... from fertilizers and pesticides threaten Niangua **darter habitat** ...
www.conservation.state.mo.us/nathis/endangered/enganger/darter/ - 12k - [Cached](#) - [Similar pages](#)

[ARKANSAS DARTER](#)
... and general development resulted in major losses of Arkansas **darter habitat**. Since the late 19th century, the Arkansas **darter's habitat** has been reduced ...
www.conservation.state.mo.us/nathis/endangered/enganger/arkdart/ - 12k - [Cached](#) - [Similar pages](#)

[RELICT DARTER, *Etheostoma chienense* U.S. Fish & Wildlife Service](#)
... of the Bayou du Chien that has significantly altered the **darter's habitat**. ... This massive alteration of the relict **darter's habitat** reduced both ...
endangered.fws.gov/i/e/sae38.html - 10k - [Cached](#) - [Similar pages](#)

[SLACKWATER DARTER, *Etheostoma boschungii* U.S. Fish & Wildlife Service](#)
... of the slackwater **darter** varies with the temperature of the breeding **habitat** and ... slackwater **darter habitat** in the Cypress Creek drainage, Tennessee. ...
endangered.fws.gov/i/e/sae1a.html - 9k - [Cached](#) - [Similar pages](#)
[[More results from endangered.fws.gov](#)]

[Characteristics of Channel Darter habitat in the Winooski River ...](#)
File Format: Microsoft Powerpoint 97 - [View as HTML](#)
Characteristics of Channel **Darter Habitat** in the Winooski River, Vermont. Douglas E. Facey and Shannon M. O'Brien. Department of Biology ...
academics.smcvet.edu/facdev/Scholarship/AFS%202003%20poster.ppt - [Similar pages](#)

[Natural Heritage Program - The Maryland Darter](#)
... We will probably never know because the Maryland **Darter's habitat** is the point ... This change in the Maryland **Darter's habitat** would have been a major ...
www.dnr.state.md.us/wildlife/mddarter.asp - 16k - Apr 7, 2005 - [Cached](#) - [Similar pages](#)

[UA AFS's Position on Ark. Darter Habitat Preservation](#)



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