CHI 2003 Tutorial

Web Search Engines: Algorithms and User Interfaces

Krishna Bharat & Bay-Wei Chang

Google Inc.
2400 Bayshore Parkway
Mountain View, CA 94043
krishna@google.com, bay@google.com
http://www.google.com/

Table of Contents

Agenda

2:00 - 2:10	Introduction, Tutorial Objectives
2:10 - 2:20	Section Outline, Role of HCI Specialists
2:20 - 3:00	Search Algorithms: Classic IR to Web IR
3:00 - 3:30	Evaluation and Measurement
3:30 - 4:00	Coffee Break
4:00 – 4:10	Web Search vs. Traditional IR
4:10 – 4:30	Interfaces for Forming Queries
4:30 – 4:50	Interfaces for Evaluating Results
4:50 – 5:10	Interfaces for Search Refinement
5:10 – 5:20	Client-side tools, 2D and 3D interfaces
5:20 - 5:30	Closing Comments, Q&A

Instructors

Krishna Bharat

Krishna is a Senior Research Scientist at Google Inc. He was previously at DEC/Compaq Systems Research Center, where he worked on interfaces and algorithms for web information retrieval. He received his Ph.D. from the GVU Center, Georgia Tech in 1996, where he worked on algorithm and infrastructure support for building distributed GUI applications.

Bay-Wei Chang

Bay-Wei is a Senior Research Scientist at Google Inc. He was previously at Xerox PARC, where his research revolved around user interface issues in web editing, portable document readers, and hyptertext annotations. He received his Ph.D. from Stanford University, where he worked on object-oriented languages, programming environments, and cartoon-inspired animation in user interfaces.

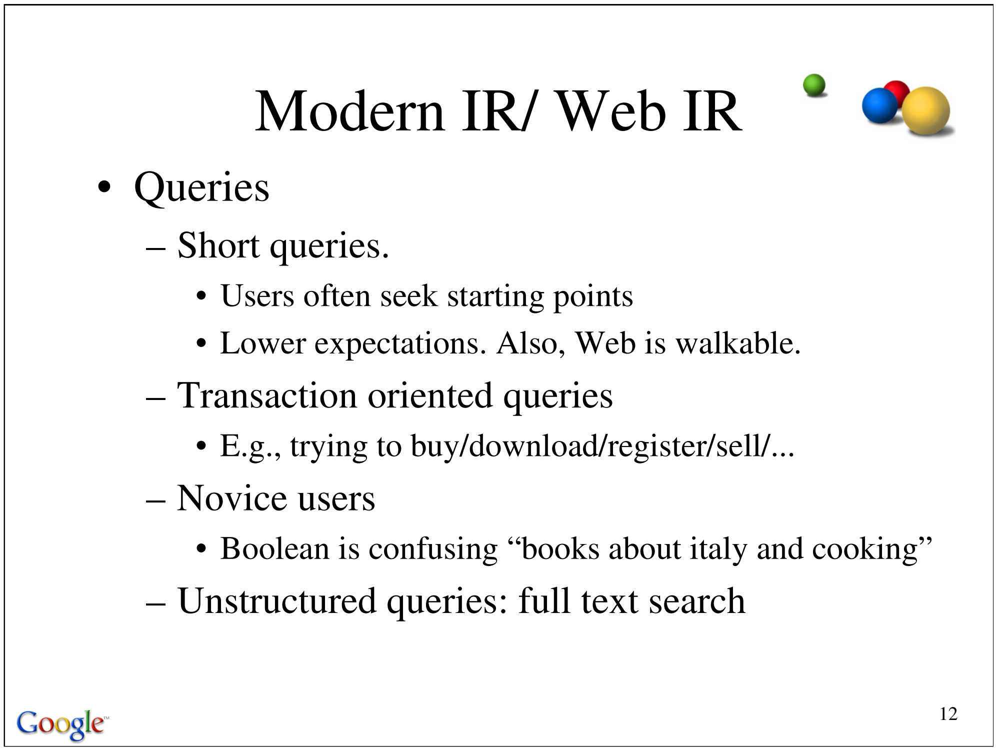
Objectives

- An introduction to the architecture, algorithms, and processes of modern search engines
- Structure and properties of the world wide web, in particular, attributes that affect the performance and quality of web search
- Search interface design, including client-side tools

Introduction

Search engines are one of the most familiar sights on the World Wide Web. As the web keeps getting larger and more unmanageable, search engines and directories become more valuable in helping people get where they want to go. Text retrieval systems, once the domain of librarians, have now moved onto the desktop, and are starting to be used on PDAs and cell phones as well.

The aim of this tutorial is to introduce HCI professionals to the user interface issues associated with search on the web. To more fully understand the interface possibilities, participants are first introduced to the architecture and algorithms of modern search engines. With this background, we will discuss prior work in user interface design for search engine front-ends and client-side search tools and opportunities for interface innovation.. We will discuss the differences between web search and traditional information retrieval in terms of audience, scope, and technologies.

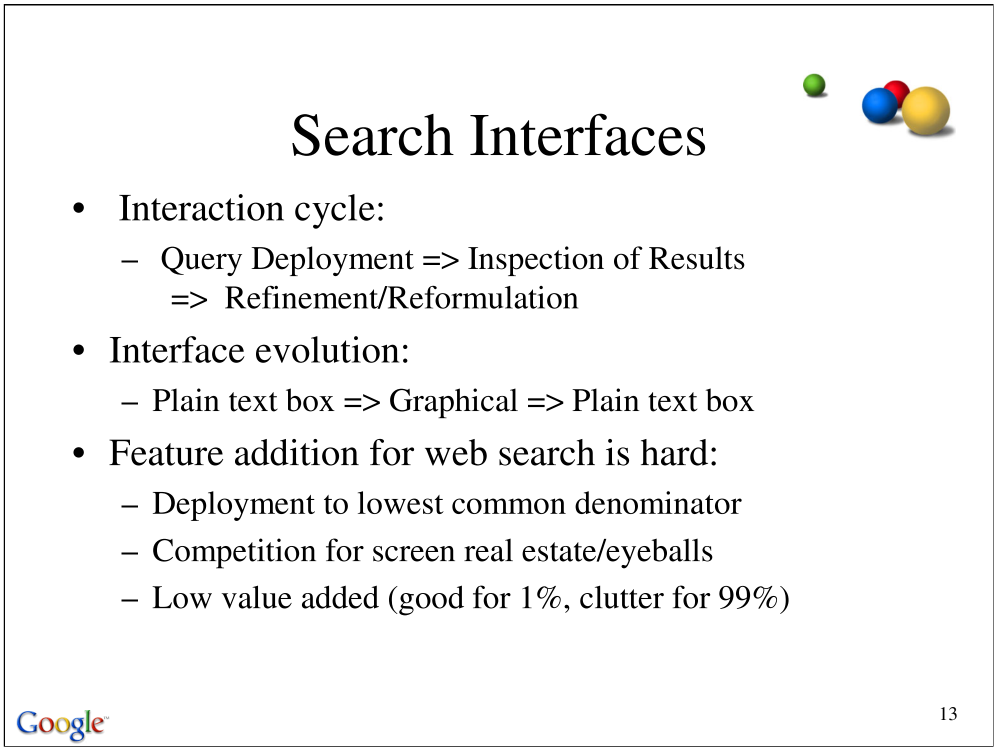


Modern IR/ Web IR

- Queries
  - – Short queries.
    - Users often seek starting points
    - Lower expectations. Also, Web is walkable.
  - – Transaction oriented queries
    - E.g., trying to buy/download/register/sell/...
  - – Novice users
    - Boolean is confusing “books about italy and cooking”
  - – Unstructured queries: full text search

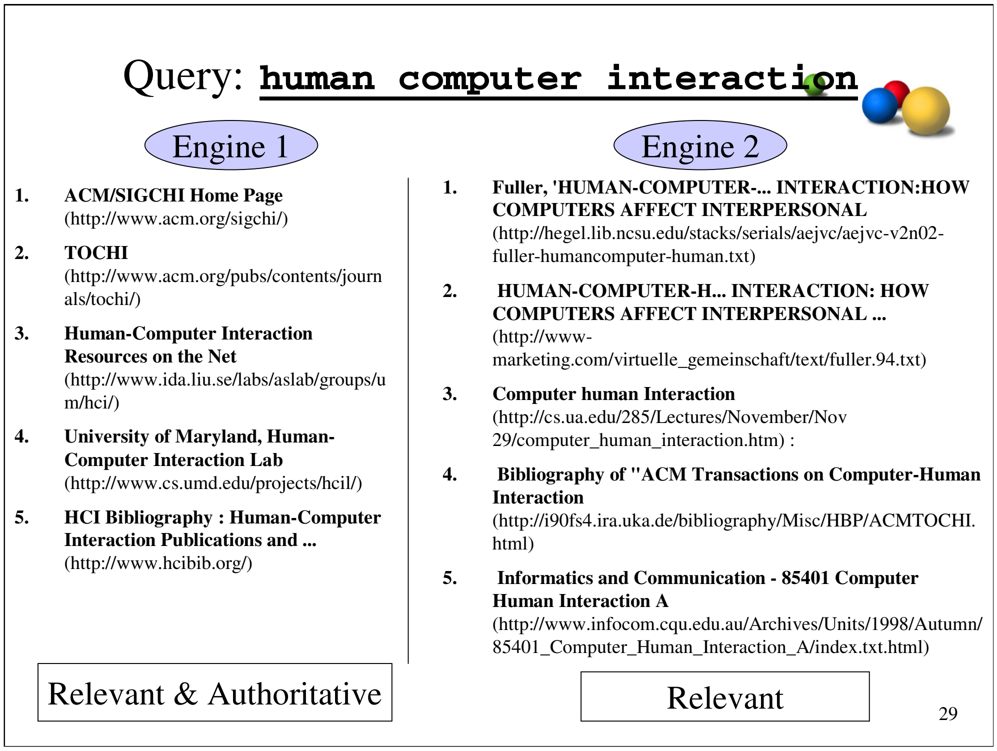
Google

12



Search Interfaces

- Interaction cycle:
  - Query Deployment => Inspection of Results => Refinement/Reformulation
- Interface evolution:
  - Plain text box => Graphical => Plain text box
- Feature addition for web search is hard:
  - Deployment to lowest common denominator
  - Competition for screen real estate/eyeballs
  - Low value added (good for 1%, clutter for 99%)



Query: **human computer interaction**

Engine 1

1. **ACM/SIGCHI Home Page** (http://www.acm.org/sigchi/)
2. **TOCHI** (http://www.acm.org/pubs/contents/journals/tochi/)
3. **Human-Computer Interaction Resources on the Net** (http://www.ida.liu.se/labs/aslab/groups/um/hci/)
4. **University of Maryland, Human-Computer Interaction Lab** (http://www.cs.umd.edu/projects/hcil/)
5. **HCI Bibliography : Human-Computer Interaction Publications and ...** (http://www.hcibib.org/)

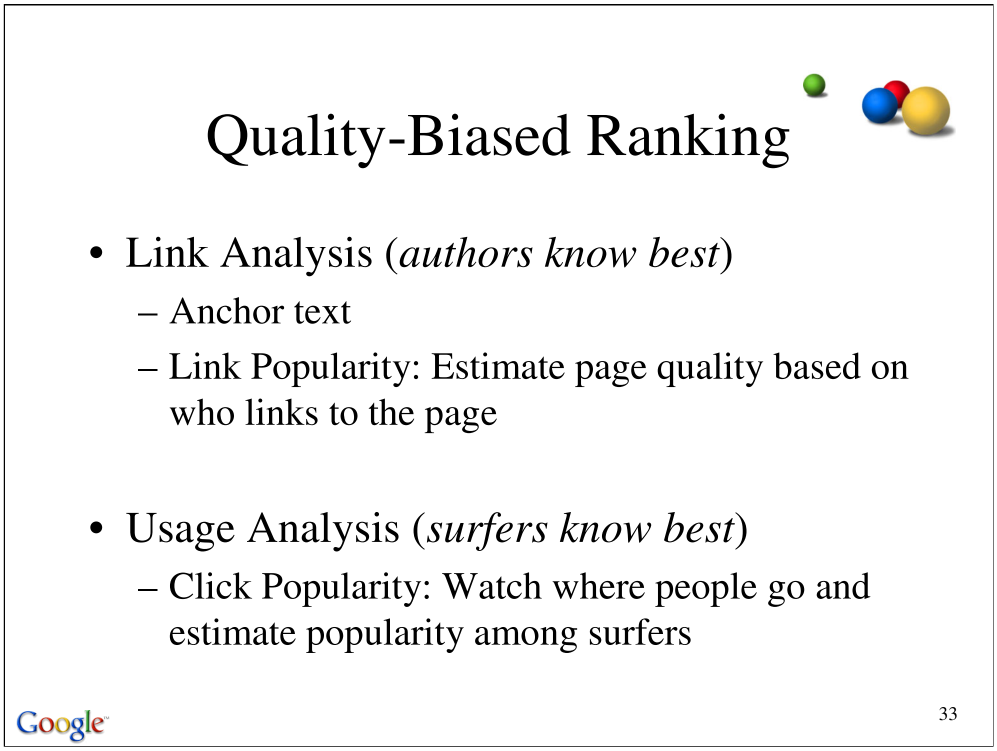
Relevant & Authoritative

Engine 2

1. **Fuller, 'HUMAN-COMPUTER-... INTERACTION:HOW COMPUTERS AFFECT INTERPERSONAL** (http://hegel.lib.ncsu.edu/stacks/serials/aejvc/aejvc-v2n02-fuller-humancomputer-human.txt)
2. **HUMAN-COMPUTER-H... INTERACTION: HOW COMPUTERS AFFECT INTERPERSONAL ...** (http://www-marketing.com/virtuelle_gemeinschaft/text/fuller.94.txt)
3. **Computer human Interaction** (http://cs.ua.edu/285/Lectures/November/Nov 29/computer_human_interaction.htm) :
4. **Bibliography of "ACM Transactions on Computer-Human Interaction** (http://i90fs4.ira.uka.de/bibliography/Misc/HBP/ACMTOCHI.html)
5. **Informatics and Communication - 85401 Computer Human Interaction A** (http://www.infocom.cqu.edu.au/Archives/Units/1998/Autumn/85401_Computer_Human_Interaction_A/index.txt.html)

Relevant

29

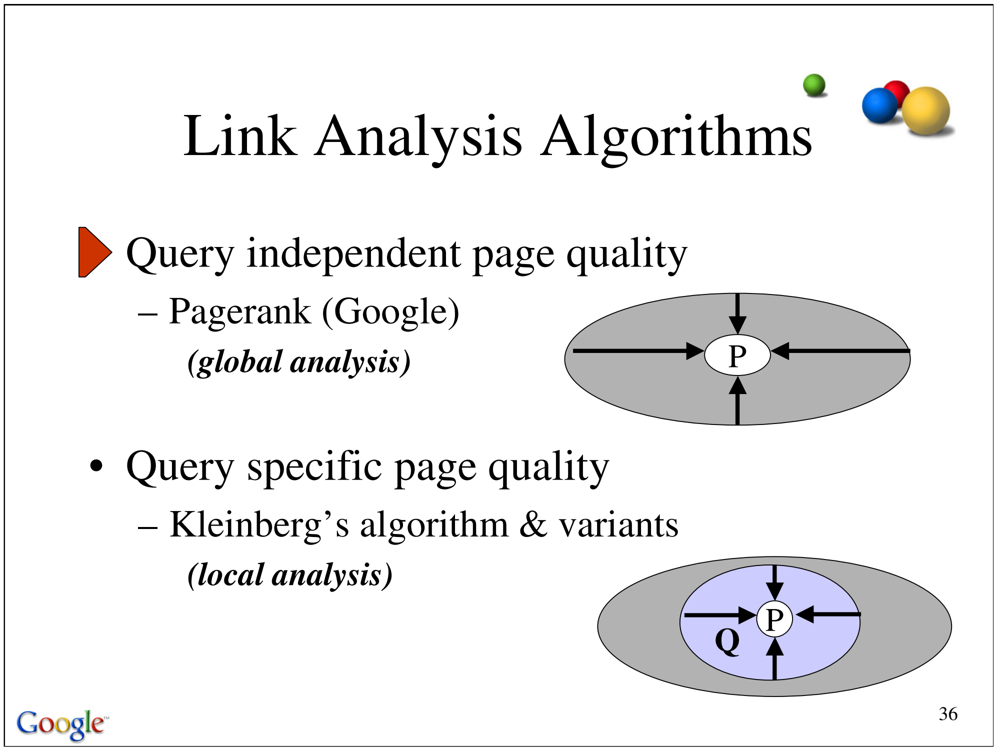


Quality-Biased Ranking

- Link Analysis (*authors know best*)
  - – Anchor text
  - – Link Popularity: Estimate page quality based on who links to the page
- Usage Analysis (*surfers know best*)
  - – Click Popularity: Watch where people go and estimate popularity among surfers

Google

33



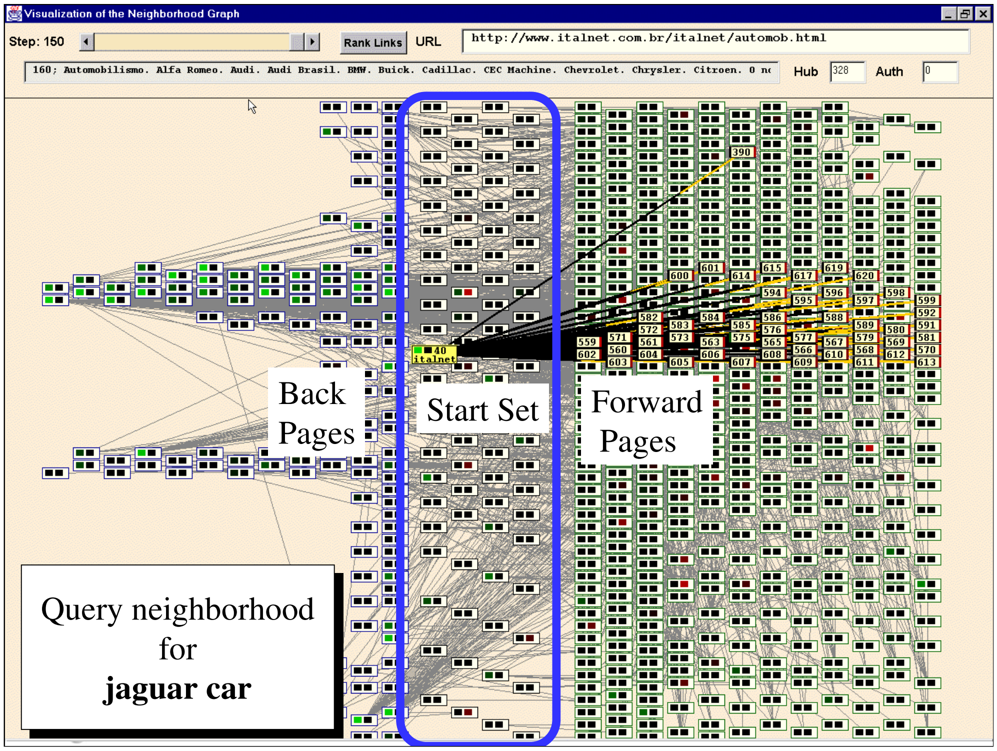
Link Analysis Algorithms

- Query independent page quality
  - – Pagerank (Google)

    (global analysis)

- Query specific page quality
  - – Kleinberg’s algorithm & variants

    (local analysis)



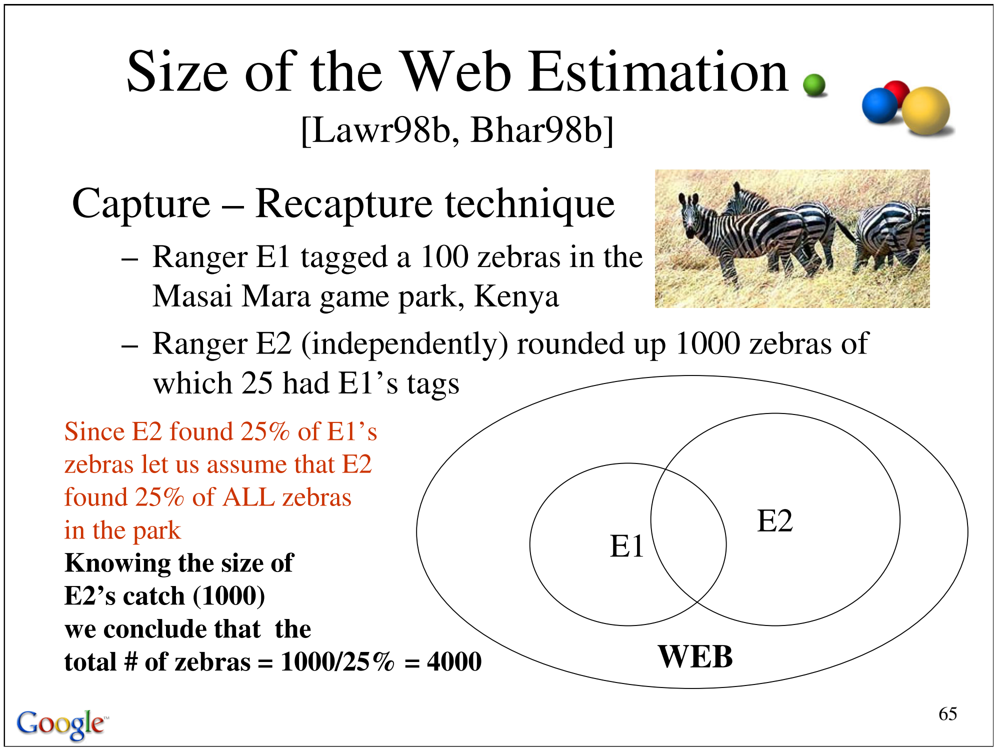
Step: 150

URL http://www.italnet.com.br/italnet/automob.html

160; Automobilismo. Alfa Romeo. Audi. Audi Brasil. BMW. Buick. Cadillac. CEC Machine. Chevrolet. Chrysler. Citroen. 0 n

Hub 328 Auth 0

Query neighborhood for **jaguar car**



Size of the Web Estimation

[Lawr98b, Bhar98b]

Capture – Recapture technique

- Ranger E1 tagged a 100 zebras in the Masai Mara game park, Kenya
- Ranger E2 (independently) rounded up 1000 zebras of which 25 had E1's tags

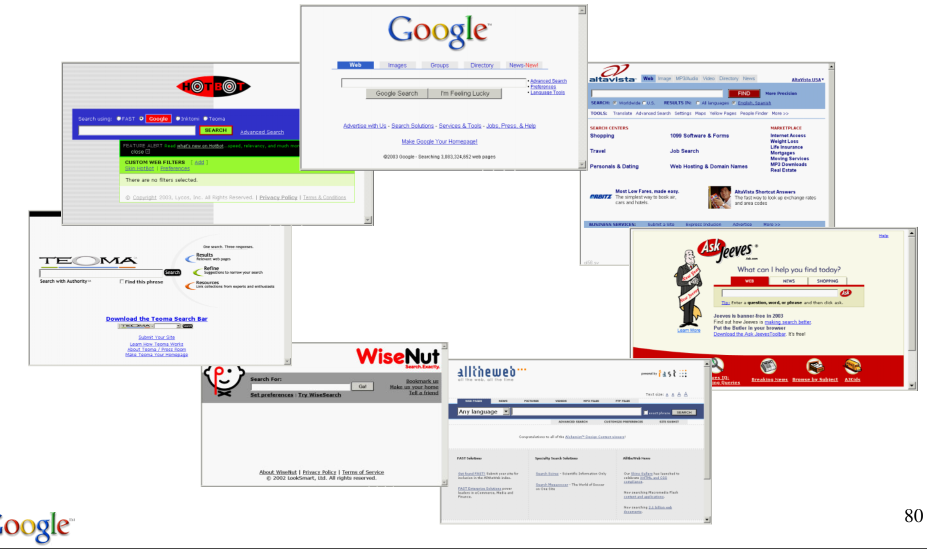
Since E2 found 25% of E1's zebras let us assume that E2 found 25% of ALL zebras in the park

Knowing the size of E2's catch (1000) we conclude that the total # of zebras = 1000/25% = 4000

Google

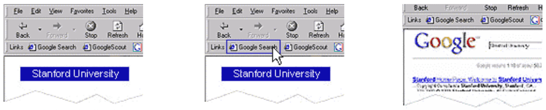
65

The search box now



Initiate queries by selection

- "Bookmarklets" [www.bookmarklets.com]
  - Search on highlighted selections in page context

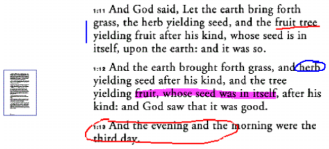

File Edit View Favorites Tools Help
Back Forward Stop Refresh
Links Google Search GoogleScout
Stanford University
Google

Highlight Text

Click Button

View Results

- Search term extraction from a wide area selection [Lowd98]
- XLibris document reader [Pric98]

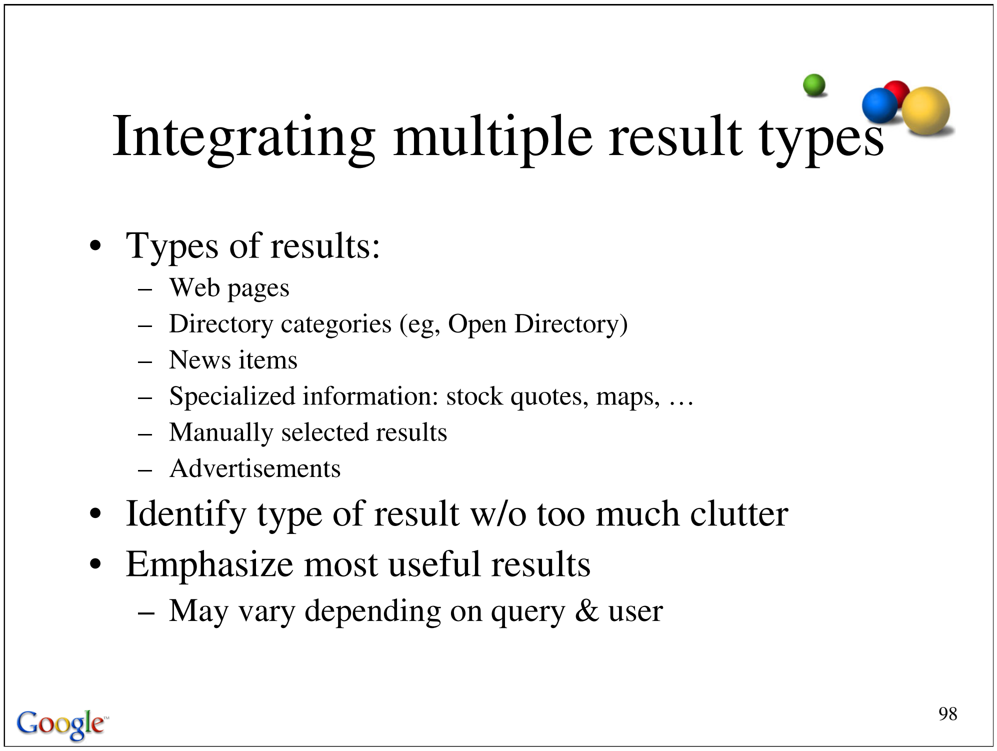


1:11 And God said, Let the earth bring forth grass, the herb yielding seed, and the fruit tree yielding fruit after his kind, whose seed is in itself, upon the earth: and it was so.

1:12 And the earth brought forth grass, and herb yielding seed after his kind, and the tree yielding fruit, whose seed was in itself, after his kind: and God saw that it was good.

1:13 And the evening and the morning were the third day.

89



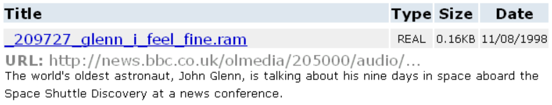
Integrating multiple result types

- Types of results:
  - Web pages
  - Directory categories (eg, Open Directory)
  - News items
  - Specialized information: stock quotes, maps, …
  - Manually selected results
  - Advertisements
- Identify type of result w/o too much clutter
- Emphasize most useful results
  - May vary depending on query & user

98

Scanning results

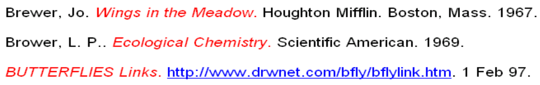
- Differentiate attributes to allow for scanning
- Tables allow easy comparison of attributes



Title	Type	Size	Date
_209727_glenn_i_feel_fine.ram	REAL	0.16KB	11/08/1998

URL: http://news.bbc.co.uk/olmedia/205000/audio/...
The world's oldest astronaut, John Glenn, is talking about his nine days in space aboard the Space Shuttle Discovery at a news conference.

  - But can be slow to load; consume space
  - TableLens [Rao94]: visualize many rows of info
- Hi-cites [Bald98]
  - Highlights similar features when moused over



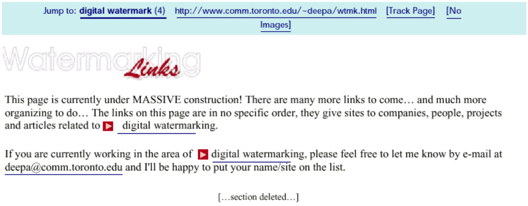
Brewer, Jo. *Wings in the Meadow.* Houghton Mifflin. Boston, Mass. 1967.

Brower, L. P.. *Ecological Chemistry.* Scientific American. 1969.

BUTTERFLIES Links. http://www.drwnet.com/bfly/bflylink.htm. 1 Feb 97.

Proxying results

- Allows enhancement of result pages
  - Highlighting search terms
  - Navigating to terms (Inquirus [Lawr98a])



Jump to: **digital watermark** (4) http://www.comm.toronto.edu/~deepa/wtmk.html [Track Page] [No Images]

This page is currently under MASSIVE construction! There are many more links to come… and much more organizing to do… The links on this page are in no specific order, they give sites to companies, people, projects and articles related to ▶ digital watermarking.

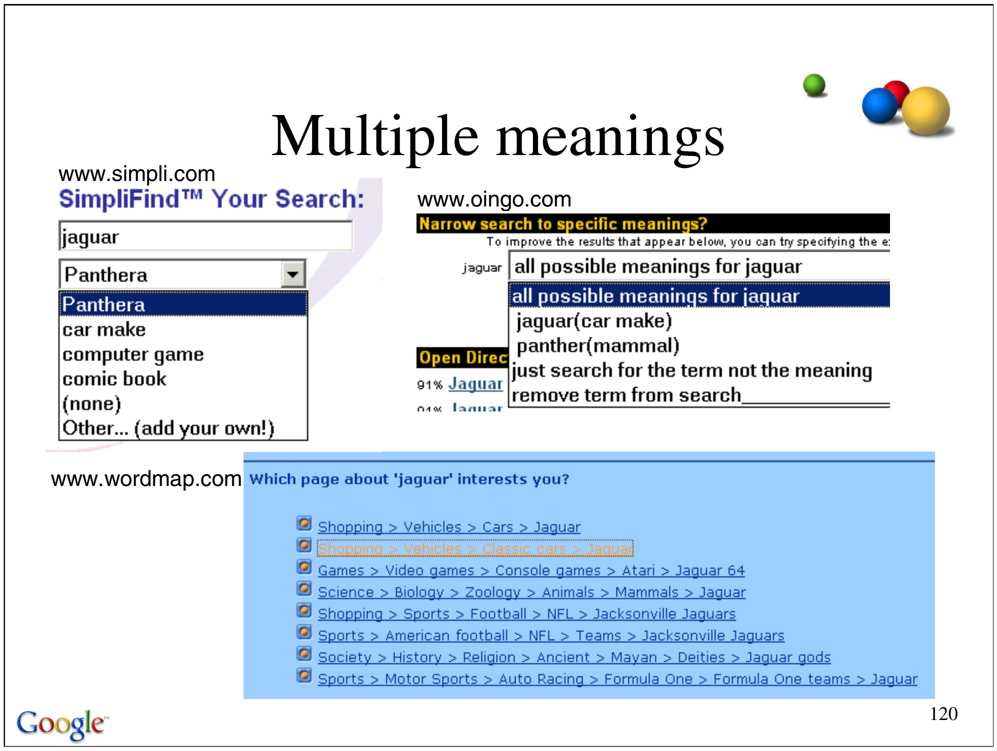
If you are currently working in the area of ▶ digital watermarking, please feel free to let me know by e-mail at deepa@comm.toronto.edu and I'll be happy to put your name/site on the list.

[…section deleted…]

  - Annotating links, adding info [Barr97, Barr98]

Google™

112



Multiple meanings

www.simpli.com

SimpliFind™ Your Search:

jaguar

Panthera ▾

- Panthera
- car make
- computer game
- comic book
- (none)
- Other... (add your own!)

www.oingo.com

Narrow search to specific meanings?

To improve the results that appear below, you can try specifying the e

jaguar: all possible meanings for jaguar

- all possible meanings for jaguar
- jaguar(car make)
- panther(mammal)
- just search for the term not the meaning
- remove term from search

Open Direc

91% Jaguar

www.wordmap.com

Which page about 'jaguar' interests you?

- Shopping > Vehicles > Cars > Jaguar
- Shopping > Vehicles > Classic cars > Jaguar
- Games > Video games > Console games > Atari > Jaguar 64
- Science > Biology > Zoology > Animals > Mammals > Jaguar
- Shopping > Sports > Football > NFL > Jacksonville Jaguars
- Sports > American football > NFL > Teams > Jacksonville Jaguars
- Society > History > Religion > Ancient > Mayan > Deities > Jaguar gods
- Sports > Motor Sports > Auto Racing > Formula One > Formula One teams > Jaguar

Google

120

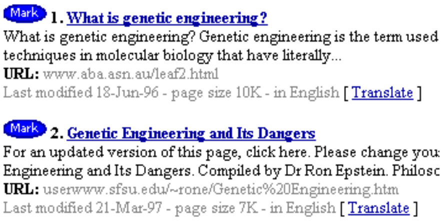
-like clustering or categorization interfaces earlier, but applied upfront

-Some require you to specify what the intended meaning is first, before any results are shown

- -Slows down search
- -Alternatively, show all results, and provide refinements
- -Still moves focus to list, rather than search results

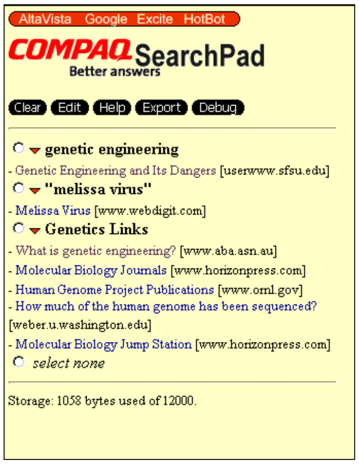
SearchPad [Bhar00b]

- Mark and save interesting search results



1. **What is genetic engineering?**
What is genetic engineering? Genetic engineering is the term used techniques in molecular biology that have literally...
URL: www.aba.asn.au/leaf2.html
Last modified 18-Jun-96 - page size 10K - in English [Translate]

2. **Genetic Engineering and Its Dangers**
For an updated version of this page, click here. Please change you: Engineering and Its Dangers. Compiled by Dr Ron Epstein. Philosc
URL: userwww.sfsu.edu/~rone/Genetic%20Engineering.htm
Last modified 21-Mar-97 - page size 7K - in English [Translate]



○ **genetic engineering**
- Genetic Engineering and Its Dangers [userwww.sfsu.edu]

○ **"melissa virus"**
- Melissa Virus [www.webdigit.com]

○ **Genetics Links**
- What is genetic engineering? [www.aba.asn.au]
- Molecular Biology Journals [www.horizonpress.com]
- Human Genome Project Publications [www.ornl.gov]
- How much of the human genome has been sequenced? [weber.u.washington.edu]
- Molecular Biology Jump Station [www.horizonpress.com]

○ *select none*

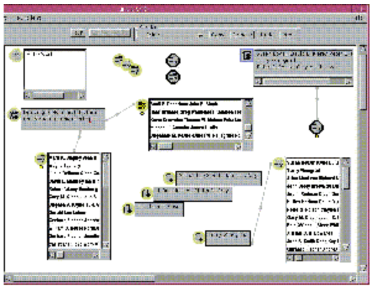
Storage: 1058 bytes used of 12000.

-MSN had a feature in which results could be saved (no longer)

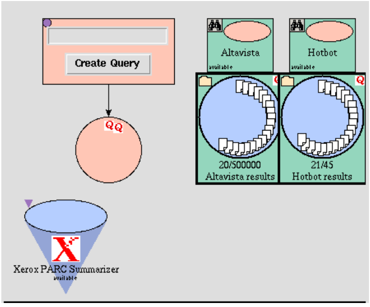
Search workspaces

- Support entire search process
- Manipulate existing queries and searches

Sketchtrieve [Hend97]



DLITE [Cous97]



[Suh02] Bongwon Suh, Allison Woodruff, Ruth Rosenholtz, and Alyssa Glass. Popout Prism: Adding perceptual principles to overview+detail document interfaces. *CHI 2002*, 2002.

[Vanr79] C. J. van Rijsbergen. *Information Retrieval (2nd Edition),* London, UK. Butterworth. 1979.
http://www.dcs.gla.ac.uk/Keith/Preface.html

[Voor98] E. M. Voorhees. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *Proc. of SIGIR 1998*: 315-323.1998.
http://www.itl.nist.gov/iaui/894.02/works/papers/sigir98.dvi.ps

[Voor98] E. Voorhees. Using WordNet for Text Retrieval. *In C. Fellbaum, (Ed.), WordNet: An Electronic Lexical Databas*e (pp.285-303). Cambridge, Massachusetts, USA: The MIT Press. 1998.

[Witt97] Kent Wittenburg and Eric Sigman. Integration of browsing, searching, and filtering in an applet for web information access. *CHI'97*, 1997.

[Wood91] Allison Woodruff, Andrew Faulring, Ruth Rosenholtz, Julie Morrison, and Peter Pirolli. Using thumbnails to search the web. *CHI 2001*, 2001.

[Zami88] Oren Zamir and Oren Etzioni: Web Document Clustering: A Feasibility Demonstration. *SIGIR 1998:* 46-54

[Zami99] Oren Zamir and Oren Etzioni: Grouper: A Dynamic Clustering Interface to Web Search Results. *WWW8 / Computer Networks*, 31(11-16): 1361-1374 (1999)

[Zell98] Polle T. Zellweger, Bay-Wei Chang, and Jock Mackinlay. Fluid links for informed and incremental link transitions. *Hypertext'98*, 1998.