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........................................................................................................iii
Instructors ..................................................................................................iv
Objectives ................................................................................................v
Introduction ..............................................................................................v
Tutorial Slides ........................................................................................1
PART 1: Web IR & Search Engines ..........................................................3
   Introduction ............................................................................................4
      IR vs Web IR.....................................................................................5
      Role of HCI Specialists ..................................................................14
   Implementing Web Search .................................................................19
      Classic IR.........................................................................................20
      Web-Specific Challenges ..............................................................28
      Web IR.............................................................................................33
   Evaluation of Search Engines .............................................................55
      Quality..............................................................................................56
      Coverage ..........................................................................................62
PART 2: Web Search User Interfaces .......................................................67
   Web Search vs Traditional IR.............................................................69
   Search Task: Information Need.........................................................75
   Search Task: Query Formulation.......................................................78
   Search Task: Evaluating Results.......................................................91
   Search Task: Refining Queries / Results ........................................117
   Search Task: Collecting Information...............................................125
   Other Topics .......................................................................................129
Bibliography ............................................................................................145
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 hypertext 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
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/fcu/)
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/)

### Engine 2

1. Fuller, 'HUMAN-COMPUTER... INTERACTION: HOW COMPUTERS AFFECT INTERPERSONAL'  
   (http://hegel.lib.ncsu.edu/stacks/serials/aejve/aejve-v2n02-fuller-humancomputer-human.txt)
2. HUMAN-COMPUTER-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://e90fs4.ira.uka.de/bibliography/Misc/HBP/ACMTOCHI.html)
5. Informatics and Communication - 85401 Computer Human Interaction A  

---

Relevant & Authoritative

Relevant
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
Link Analysis Algorithms

- Query independent page quality
  – Pagerank (Google)
    *(global analysis)*

- Query specific page quality
  – Kleinberg’s algorithm & variants
    *(local analysis)*
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
The search box now
Initiate queries by selection

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

- Search term extraction from a wide area selection [Lowd98]

- XLibris document reader [Pric98]
Integrating multiple result types

- Types of results:
  - Web pages
  - Directory categories (e.g., 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
Scanning results

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

<table>
<thead>
<tr>
<th>Title</th>
<th>Type</th>
<th>Size</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>_200727_gllnn.i_feel_fine.rom</td>
<td>REAL</td>
<td>0.1KB</td>
<td>13/08/99</td>
</tr>
</tbody>
</table>

URL: http://news.bbc.co.uk/news/hq/200727/.../ ... the world's oldest astronaut, John Glenn, is talking about his space 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, J. 

  Brewer, L. P. 

Proxying results

- Allows enhancement of result pages
  - Highlighting search terms
  - Navigating to terms (Inquirus [Lawr98a])
  - Annotating links, adding info [Barr97, Barr98]
Multiple meanings

- 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

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

- Support entire search process
- Manipulate existing queries and searches


