Evaluating IR *In Situ*

Susan Dumais

Microsoft Research

Perspective for this Talk

- Information retrieval systems are developed to help people find information to satisfy their information needs
- Success depends critically on two general components
    - Content and ranking
    - User interface and interaction
- Data as a critical resource for research
- Cranfield/TREC-style resources
    - Great for some components and some user models
- Can we develop similar resources for understanding and improving the user experience?
- Can we study individual components in isolation, or do we need to consider the system as a whole?

$$ You have won 100 Million $$

- Challenge: *You have been asked to lead a team to improve the **AYoBig** Web search engine. You have a budget of 100 million dollars. How would you spend it?*
- Content
  - Ranking – query analysis; doc representation; matching …
  - Crawl - coverage, new sources, freshness, …
  - Spam detection
- User experience
  - Presentation (speed, layout, snippets, more than results)
  - Features like spelling correction, related searches, …
  - Richer capabilities to support query articulation, results analysis, …

$$ You have won 100 Million $$

- Challenge: *You have been asked to lead a team to improve the* ***AYoBig*** *Web search engine. You have a budget of 10 million dollars. How would you spend it?*
- Depends on:
  - What are the problems now?
  - What are you trying to optimize?
  - What are the costs and effect sizes?
  - What are the tradeoffs?
  - How do various components combine?
  - Etc.

Evaluating Search Systems

- Traditional test collections
  - Fix: Docs, Queries, RelJ (Q-Doc), Metrics
  - Goal: Compare systems, w/ respect to metric
  - NOTE: Search engines do this, but not just this …
- What’s missing?
  - Metrics: User model (pr@k, nncg), average performance, all queries equal
  - Queries: Types of queries, history of queries (session and longer)
  - Docs: The “set” of documents – duplicates, site collapsing, diversity, etc.
  - Selection: Nature and dynamics of queries, documents, users
  - Users: Individual differences (location, personalization including re-finding), iteration and interaction
  - Presentation: Snippets, speed, features (spelling correction, query suggestion), the whole page

Kinds of User Data

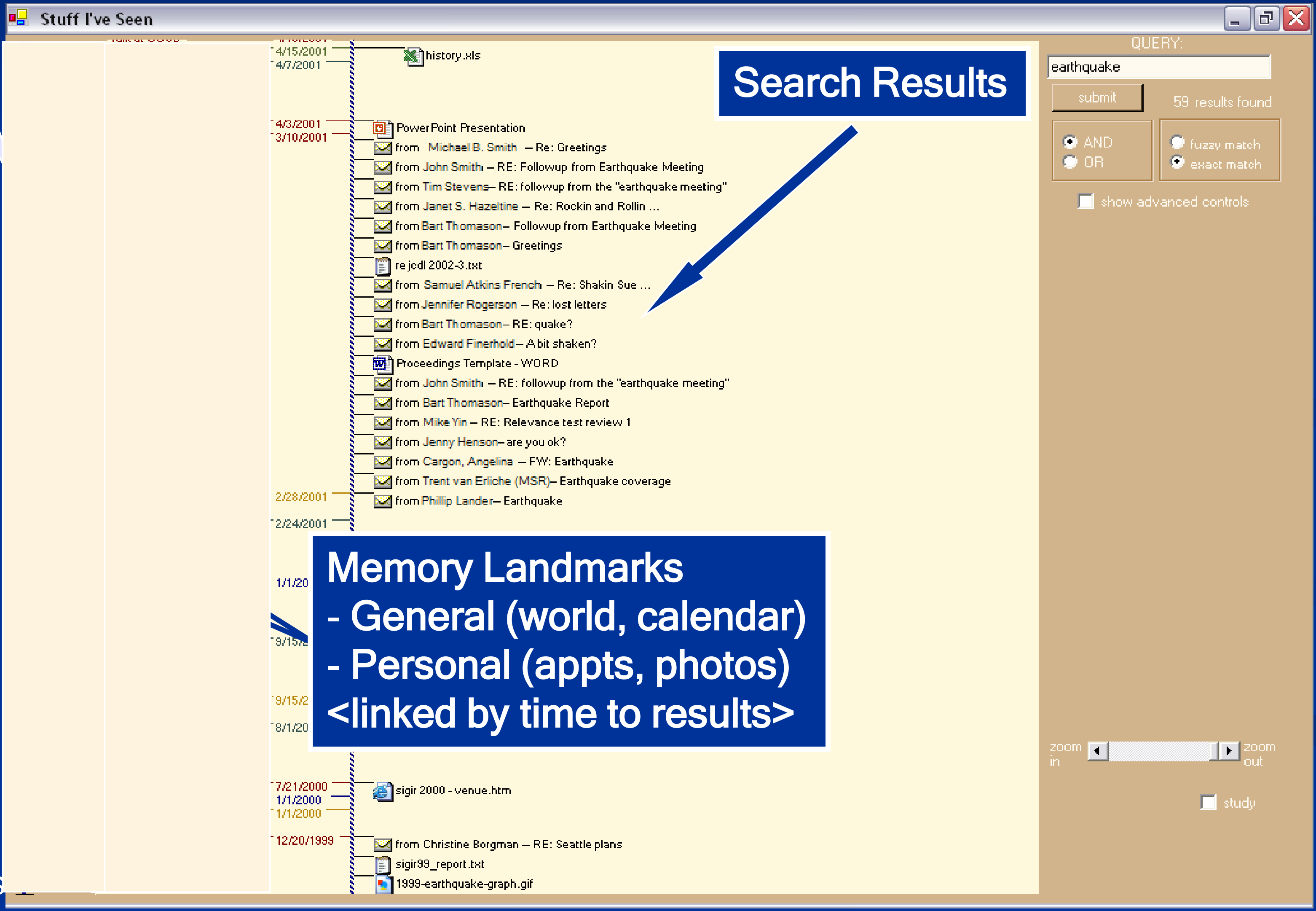
- User Studies
  - Lab setting, controlled tasks, detailed instrumentation (incl. gaze, video), nuanced interpretation of behavior
- User Panels
  - In-the-wild, user-tasks, reasonable instrumentation, can probe for more detail
- Log Analysis and Experimentation (in the large)
  - In-the-wild, user-tasks, no explicit feedback but lots of implicit indicators
  - The what vs. the why
- Others: field studies, surveys, focus groups, etc.

User Studies

- E.g., Search UX (timeline views, query suggestion)
- Memory Landmarks [Ringel et al., Interact 2003]

SIS, Timeline w/ Landmarks

Distribution of Results Over Time



QUERY: earthquake

59 results found

● AND ○ OR

○ fuzzy match ● exact match

☐ show advanced controls

☐ study

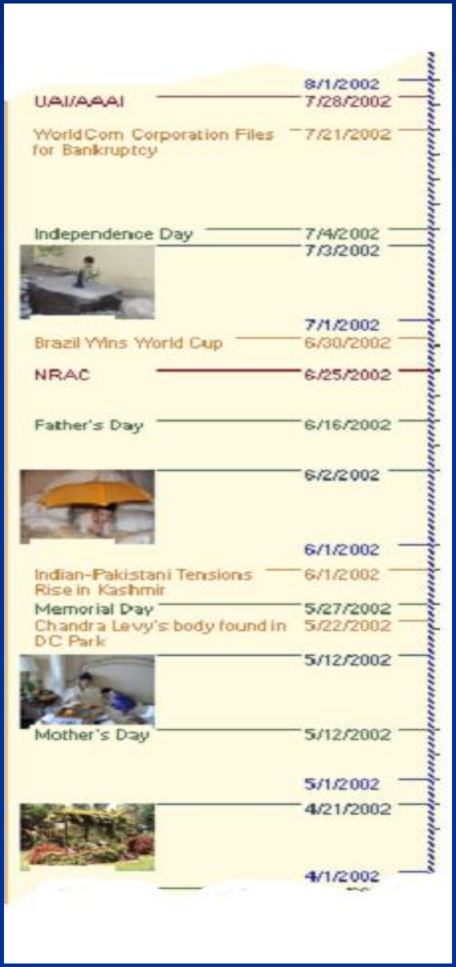
history.xls
PowerPoint Presentation
from Michael B. Smith – Re: Greetings
from John Smith – RE: Followup from Earthquake Meeting
from Tim Stevens– RE: followup from the "earthquake meeting"
from Janet S. Hazeltine – Re: Rockin and Rollin ...
from Bart Thomason– Followup from Earthquake Meeting
from Bart Thomason– Greetings
re jcdl 2002-3.txt
from Samuel Atkins French – Re: Shakin Sue ...
from Jennifer Rogerson – Re: lost letters
from Bart Thomason– RE: quake?
from Edward Finerhold– A bit shaken?
Proceedings Template - WORD
from John Smith – RE: followup from the "earthquake meeting"
from Bart Thomason– Earthquake Report
from Mike Yin – RE: Relevance test review 1
from Jenny Henson– are you ok?
from Cargon, Angelina – FW: Earthquake
from Trent van Erliche (MSR)– Earthquake coverage
from Phillip Lander– Earthquake
sigir 2000 - venue.htm
from Christine Borgman – RE: Seattle plans
sigir99_report.txt
1999-earthquake-graph.gif

Search Results

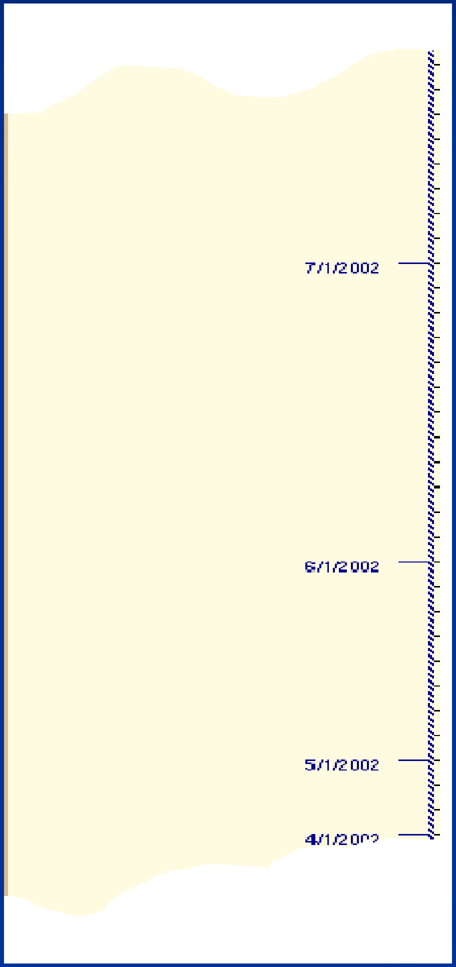
Memory Landmarks
- General (world, calendar)
- Personal (appts, photos)
<linked by time to results>

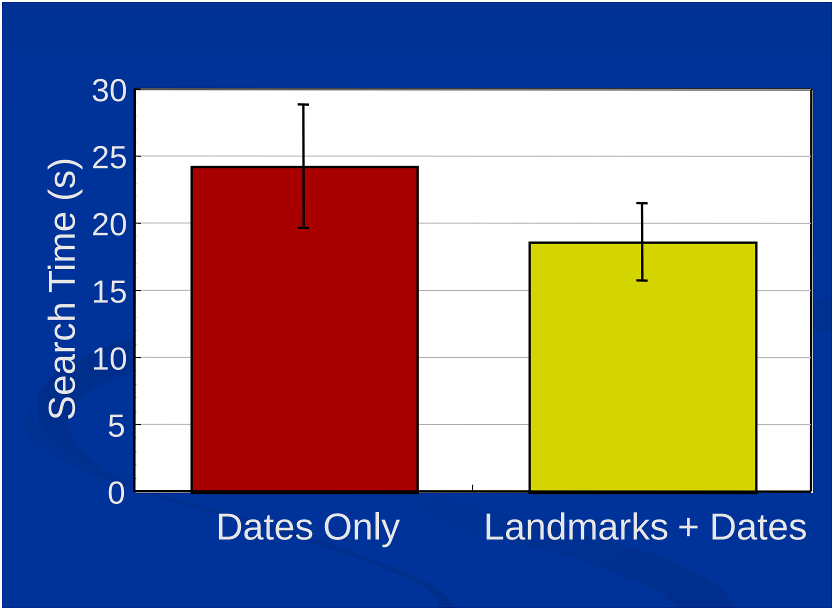
SIS, Timeline Experiment

With Landmarks



Without Landmarks





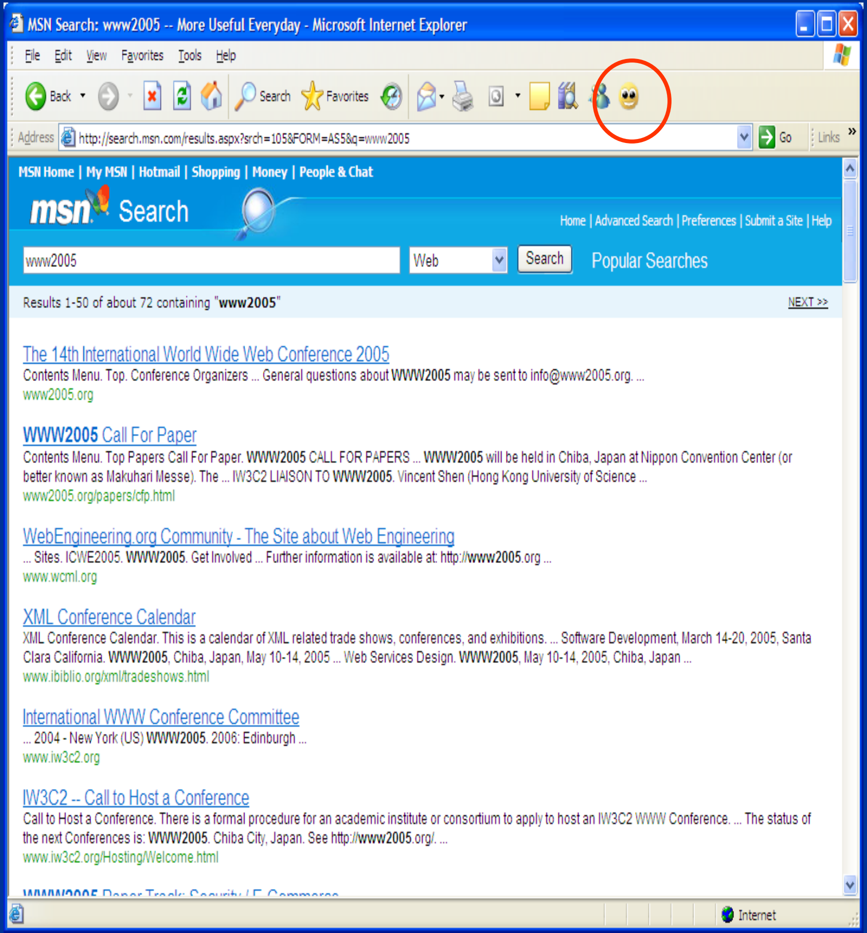
User Studies

- E.g., Search UX (timeline views, query suggestion)
- Laboratory (usually)
- Small-scale (10s-100s of users; 10s of queries)
- Months for data
- Known tasks and known outcome (labeled data)
- Detailed logging of queries, URLs visited, scrolling, gaze tracking, video
- Can evaluate experimental prototypes
- Challenges – user sample, behavior w/ experimenter present or w/ new features

User Panels

- E.g., Curious Browser, SIS, Phlat
- Curious Browser [Fox et al., TOIS 2005]

Curious Browser
(link explicit user judgments w/ implicit actions)


MSN Search: www2005 -- More Useful Everyday - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://search.msn.com/results.aspx?srch=105&FORM=AS5&q=www2005
Go Links
MSN Home | My MSN | Hotmail | Shopping | Money | People & Chat
msn Search
Home | Advanced Search | Preferences | Submit a Site | Help
www2005
Web
Search
Popular Searches
Internet

Results 1-50 of about 72 containing "**www2005**" NEXT >>

The 14th International World Wide Web Conference 2005
Contents Menu. Top. Conference Organizers ... General questions about WWW2005 may be sent to info@www2005.org. ...
www2005.org

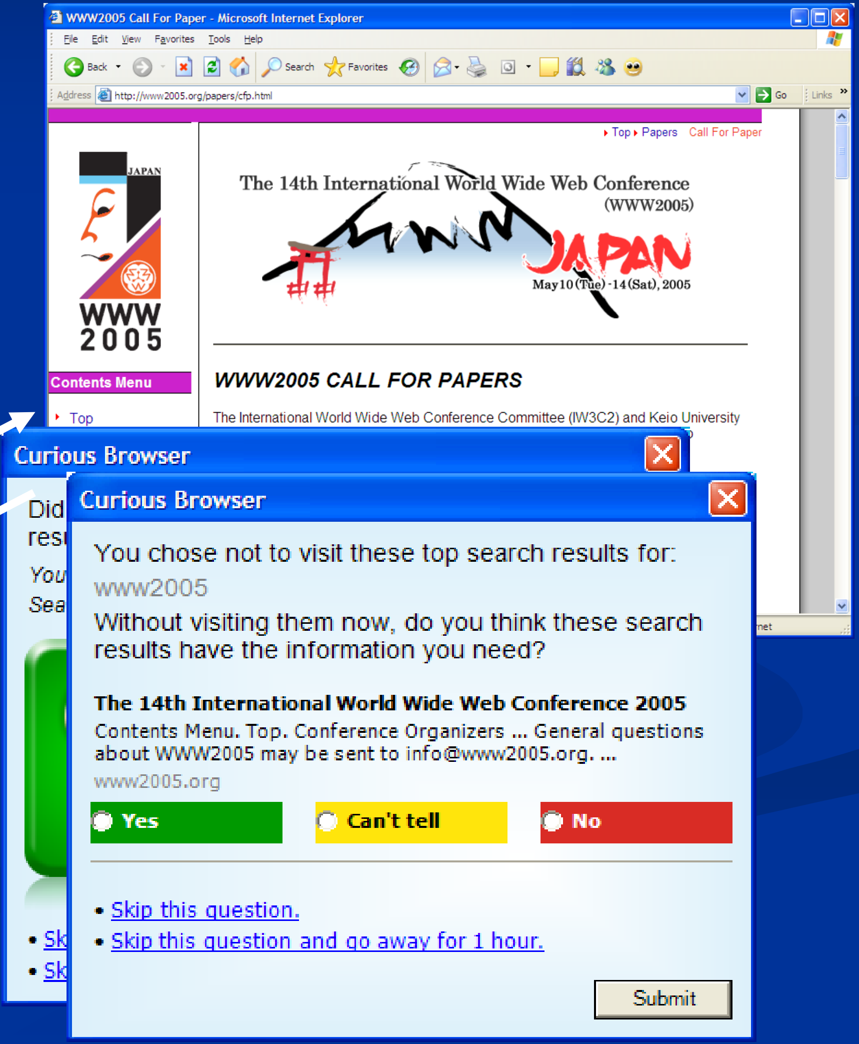
WWW2005 Call For Paper
Contents Menu. Top Papers Call For Paper. WWW2005 CALL FOR PAPERS ... WWW2005 will be held in Chiba, Japan at Nippon Convention Center (or better known as Makuhari Messe). The ... IW3C2 LIAISON TO WWW2005. Vincent Shen (Hong Kong University of Science ...
www2005.org/papers/cfp.html

WebEngineering.org Community - The Site about Web Engineering
... Sites. ICWE2005. WWW2005. Get Involved ... Further information is available at: http://www2005.org ...
www.wcml.org

XML Conference Calendar
XML Conference Calendar. This is a calendar of XML related trade shows, conferences, and exhibitions. ... Software Development, March 14-20, 2005, Santa Clara California. WWW2005, Chiba, Japan, May 10-14, 2005 ... Web Services Design. WWW2005, May 10-14, 2005, Chiba, Japan ...
www.ibiblio.org/xml/tradeshows.html

International WWW Conference Committee
... 2004 - New York (US) WWW2005. 2006: Edinburgh ...
www.iw3c2.org

IW3C2 -- Call to Host a Conference
Call to Host a Conference. There is a formal procedure for an academic institute or consortium to apply to host an IW3C2 WWW Conference. ... The status of the next Conferences is: WWW2005. Chiba City, Japan. See http://www2005.org/. ...
www.iw3c2.org/Hosting/Welcome.html


WWW2005 Call For Paper - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www2005.org/papers/cfp.html
Go Links
Top Papers Call For Paper
JAPAN
WWW 2005
The 14th International World Wide Web Conference (WWW2005)
JAPAN
May 10 (Tue) - 14 (Sat), 2005
Contents Menu
Top
Curious Browser
Did
resu
You
Sea
Sk
Sk
Curious Browser

WWW2005 CALL FOR PAPERS

The International World Wide Web Conference Committee (IW3C2) and Keio University

You chose not to visit these top search results for:
www2005
Without visiting them now, do you think these search results have the information you need?

The 14th International World Wide Web Conference 2005
Contents Menu. Top. Conference Organizers ... General questions about WWW2005 may be sent to info@www2005.org. ...
www2005.org

○ Yes ○ Can't tell ○ No

- Skip this question.
- Skip this question and go away for 1 hour.

Submit

SIGIR 2009

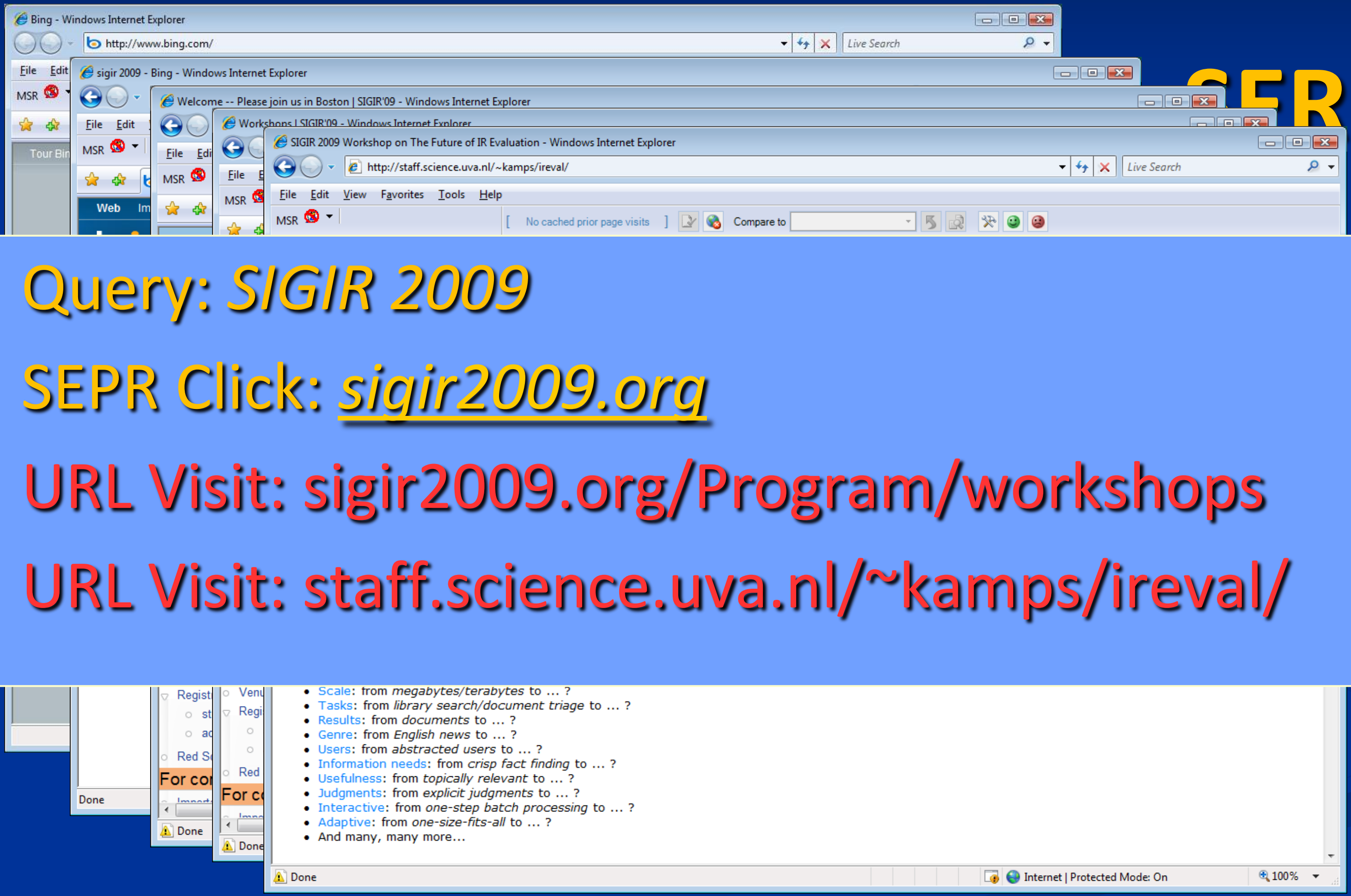
User Panels

- E.g., Curious Browser, SIS, Phlat
- Browser toolbar or other client code
- Smallish-scale (100s-1000s of users; queries)
- Weeks for data
- In-the-wild, search interleaved w/ other tasks
- Logging of queries, URLs visited, screen capture, etc.
- Can probe about specific tasks and success/failure (some labeled data)
- Challenges – user sample, drop out, some alteration of behavior

Log Analysis and Expts (in the large)

- E.g., Query-Click logs
  - Search engine vs. Toolbar
  - Search engine
    - Know lots of details about your application (e.g. results, features)
    - Only know activities on the SERP
  - Toolbar (or other client code)
    - Can see activity with many sites, including what happens after the SERP
    - Don’t know as many details of each page

SERP


Bing - Windows Internet Explorer
http://www.bing.com/
sigir 2009 - Bing - Windows Internet Explorer
Welcome -- Please join us in Boston | SIGIR'09 - Windows Internet Explorer
Workshops | SIGIR'09 - Windows Internet Explorer
SIGIR 2009 Workshop on The Future of IR Evaluation - Windows Internet Explorer
http://staff.science.uva.nl/~kamps/ireval/
Live Search
File Edit View Favorites Tools Help
MSR
No cached prior page visits
Compare to
Done
Internet | Protected Mode: On
100%

- Scale: from *megabytes/terabytes* to … ?
- Tasks: from *library search/document triage* to … ?
- Results: from *documents* to … ?
- Genre: from *English news* to … ?
- Users: from *abstracted users* to … ?
- Information needs: from *crisp fact finding* to … ?
- Usefulness: from *topically relevant* to … ?
- Judgments: from *explicit judgments* to … ?
- Interactive: from *one-step batch processing* to … ?
- Adaptive: from *one-size-fits-all* to … ?
- And many, many more…

- Query: *SIGIR 2009*
- SEPR Click: *sigir2009.org*
- URL Visit: sigir2009.org/Program/workshops
- URL Visit: staff.science.uva.nl/~kamps/ireval/

Log Analysis and Expts (in the large)

- E.g., Query-Click logs
  - Search engine - details of your service (results, features, etc.)
  - Toolbar – broader coverage of sites/services, less detail
- Millions of users and queries
- Real-time data
- In-the-wild
- Benefits – diversity and dynamics of users, queries, tasks, actions
- Challenges
  - Logs are very noisy (bots, collection errors)
  - Unlabeled activity – *the what, not the why*

Log Analysis and Expts (in the large)

- E.g., Experiential platforms
- Operational systems can (and do) serve as "experimental platforms"
  - A/B testing
  - Interleaving for ranking evaluation

Sharable Resources?

- User studies / Panel studies
  - Data collection infrastructure and instruments
  - Perhaps data
- Log analysis – Queries, URLs
  - Understanding how user interact with existing systems
    - What they are doing; Where they are failing; etc.
  - Implications for
    - Retrieval models
    - Lexical resources
    - Interactive systems
  - Lemur Query Log Toolbar – developing a community resource !

Sharable Resources?

- Operational systems as an experimental platform
  - Can generate logs, but more importantly …
  - Can also conduct controlled experiments *in situ*
    - A/B testing -- Data vs. the "hippo" [Kohavi, CIKM 2009]
    - Interleave results from different methods [Radlinski & Joachims, AAAI 2006]
  - Can we build a "Living Laboratory"?
    - Web search
      - Search APIs , but ranking experiments somewhat limited
      - UX perhaps more natural
    - Search for other interesting sources
      - Wikipedia, Twitter, Scholarly publications, …
  - Replicability in the face of changing content, users, queries

Closing Thoughts

- Information retrieval systems are developed to help people satisfy their information needs
- Success depends critically on
  - Content and ranking
  - User interface and interaction
- Test collections and data are critical resources
  - Today's TREC-style collections are limited with respect to user activities
  - Can we develop shared user resources to address this?
    - Infrastructure and instruments for capturing user activity
    - Shared toolbars and corresponding user interaction data
    - "Living laboratory" in which to conduct user studies at scale