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

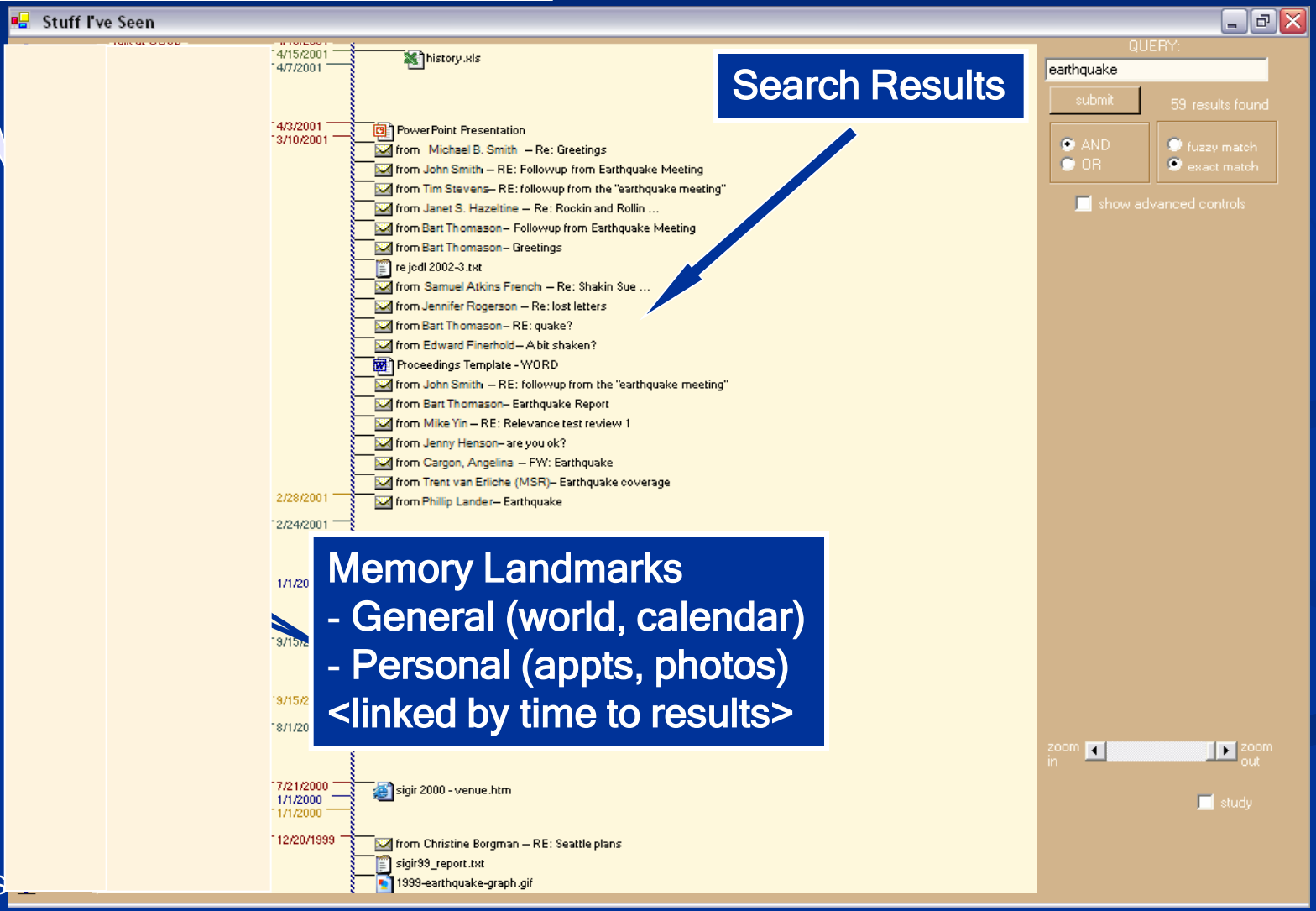
- User Studies
  - Lab setting, controlled tasks, detailed instrumentation (incl. gaze, video), nuanced interpretation of behavior
- User Panels
  - In-the-wild, user-tasks, reasonable instrumentation, can probe for more detail
- Log Analysis and Experimentation (in the large)
  - In-the-wild, user-tasks, 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

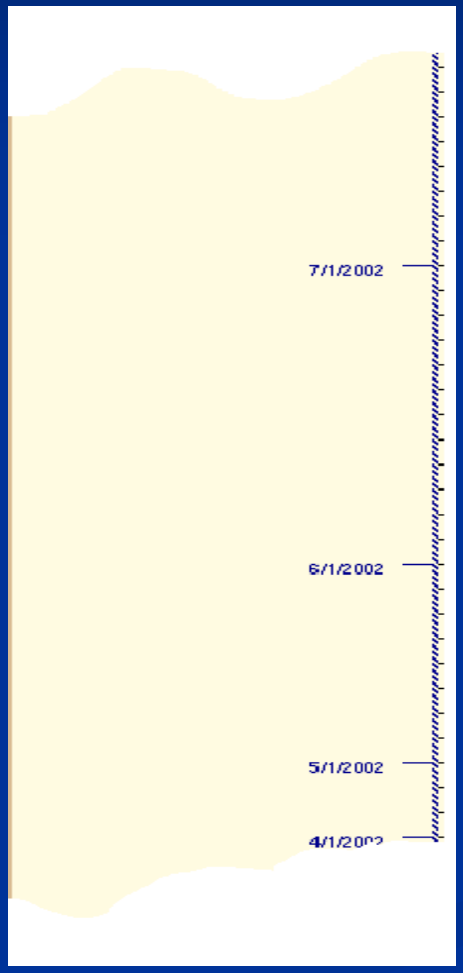
Distribution of Results Over Time

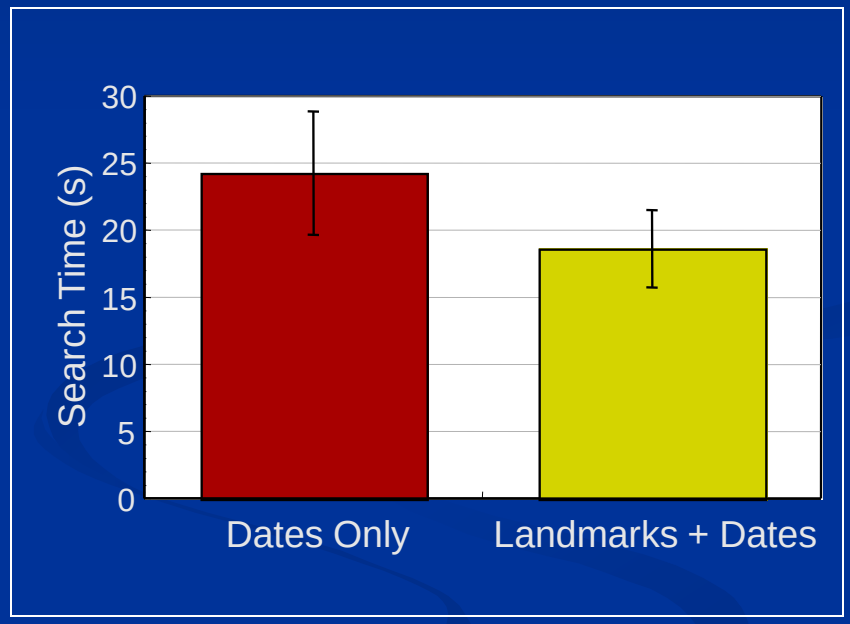


SIS, Timeline Experiment

With Landmarks

Without Landmarks





User Studies

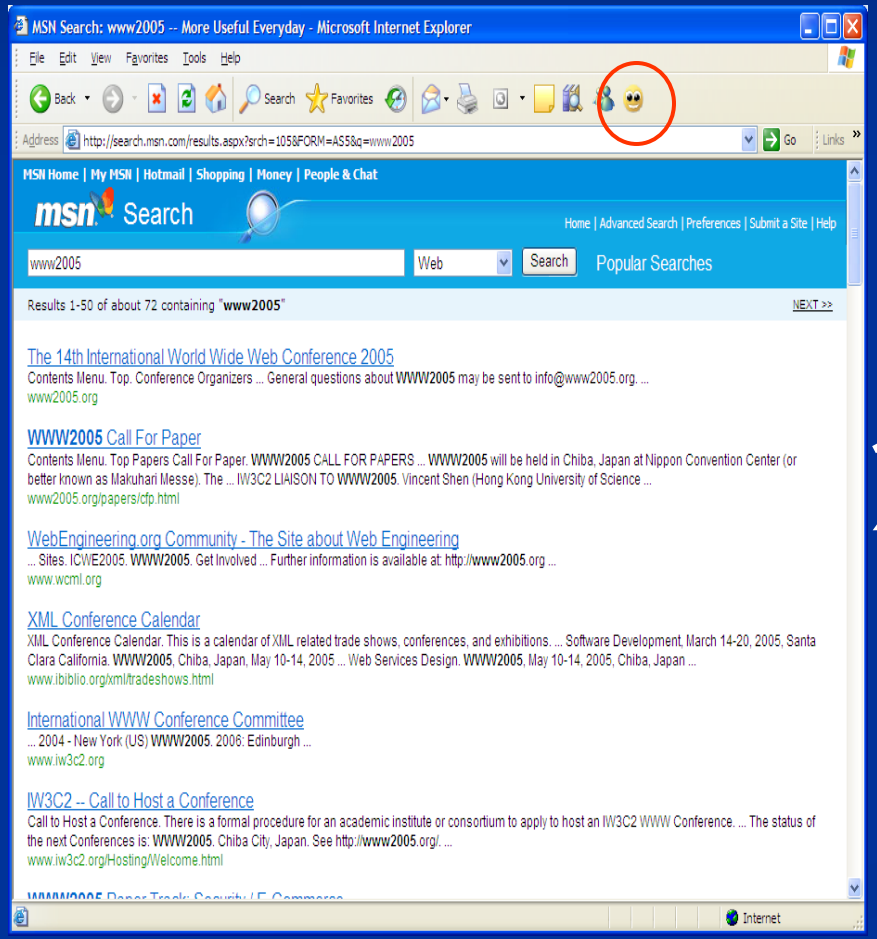
- E.g., Search UX (timeline views, query suggestion)
- Laboratory (usually)
- Small-scale (10s-100s of users; 10s of queries)
- Months for data
- Known tasks and known outcome (labeled data)
- Detailed logging of queries, URLs visited, scrolling, gaze tracking, video
- Can evaluate experimental prototypes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

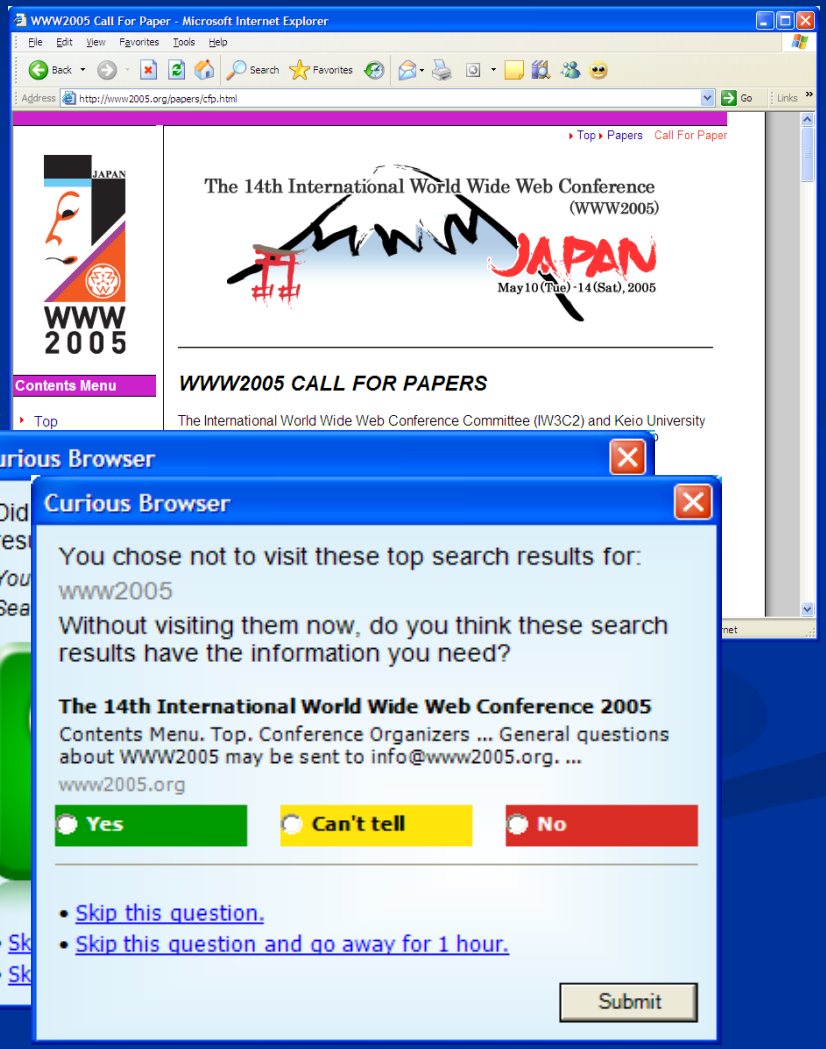
User Panels

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

Curious Browser

(link explicit user judgments w/ implicit actions)



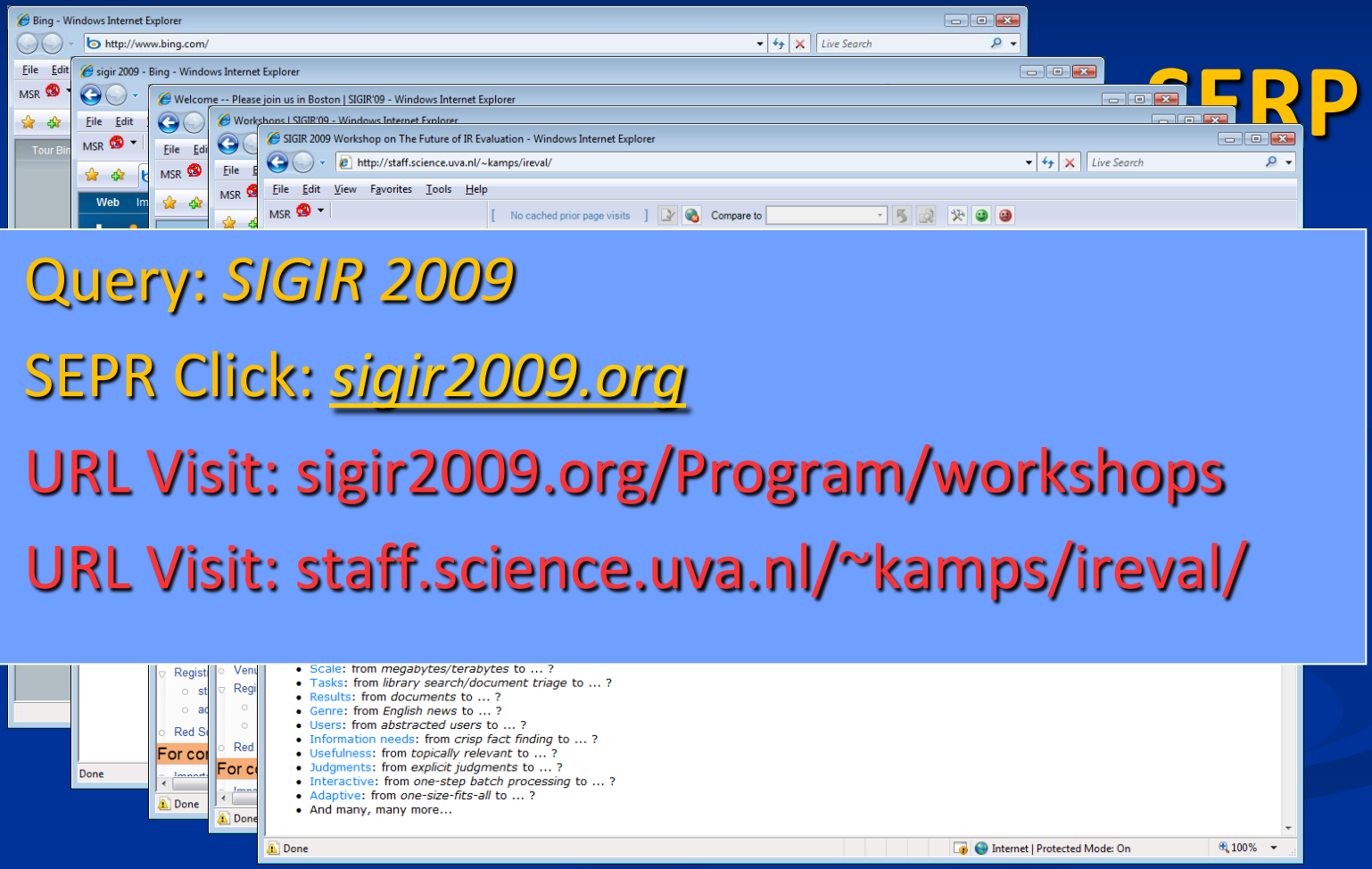


User Panels

- 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



- 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