Using Context to Support Searchers in Searching

Susan Dumais
Microsoft Research

http://research.microsoft.com/~sdumais
Using Context to Support Searchers

User Context

Query Words

RankedList

Ranked List

Document Context

Task/Use Context
Web Info through the Years

What’s available

- Number of pages indexed
  - 7/94 Lycos -
  - 95 - $10^6$ millions
  - 97 - $10^7$
  - 98 - $10^8$
  - 01 - $10^9$ billions
  - 05 - $10^{10}$ ...

- Types of content
  - Web pages, newsgroups
  - Images, videos, maps
  - News, blogs, spaces
  - Shopping, local, desktop
  - Books, papers
  - Health, finance, travel ...

How it’s accessed
Some Support for Searchers

- The search box
- Spelling suggestions
- Query suggestions
- Advanced search operators and options (e.g., “”, +/-, site:, language:, filetype:, intitle:)
- Richer snippets
- But, we can do better … using context
Key Contexts

- **Users:**
  - Individual, group (topic, time, location, etc.)
  - Short-term or long-term models
  - Explicit or implicit capture

- **Documents/Domains:**
  - Document-level metadata, usage/change patterns
  - Relations among documents

- **Tasks/Uses:**
  - Information goal - Navigational, fact-finding, informational, monitoring, research, learning, social, etc.
  - Physical setting - Device, location, time, etc.
Using Contexts

- **Identify:**
  - What context(s) are of interest?

- **Accommodate:**
  - What do we do differently for different contexts?
  - Outcome \((Q|\text{context}) \gg \text{Outcome (Q)}\)

- **Influence points within the search process**
  - Articulating the information need
    - Initial query, subsequent interaction/dialog
  - Selecting and/or ranking content
  - Presenting results
  - Using and sharing results
Context in Action

Research prototypes: provide insights about algorithmic, user experience, and policy challenges

- **User Contexts:**
  - Finding and Re-Finding (Stuff I’ve Seen)
  - Personalized Search (PSearch)
  - Novelty in News (NewsJunkie)

- **Document/Domain Contexts:**
  - Metadata and search (Phlat)
  - Visualizing patterns in results (GridViz)

- **Task/Use Contexts:**
  - Pages as context (Community Bar, IQ)
  - Richer collections as context (NewsJunkie, PSearch)
  - Working, understanding, sharing (SearchTogether, InkSeine)
**SIS: Stuff I’ve Seen**

- Unified index of *stuff you’ve seen*
  - Many info silos (e.g., files, email, calendar, contacts, web pages, rss, im)
  - Unified index, not storage
  - Index of content and metadata (e.g., time, author, title, size, access)
  - Re-finding vs. finding

*Vista Desktop Search (and Live Toolbar)*

Also, Spotlight, GDS, X1, ...
### SIS Demo

**Stuff I've Seen**

<table>
<thead>
<tr>
<th>Document</th>
<th>Date</th>
<th>Rank</th>
<th>Path</th>
<th>Author</th>
<th>Mail To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Briefing: AP/Sue Dumais 5/6@4:30 p.m.</td>
<td>5/6/2004 11:27</td>
<td>...</td>
<td>personal folders/media</td>
<td>Km Davis</td>
<td>Susan Dumais</td>
</tr>
<tr>
<td>Final Recap: The Economist Campus Visit April 15...</td>
<td>4/20/2004 6:21</td>
<td>...</td>
<td>personal folders/media</td>
<td>Kristen Birkeland</td>
<td>Karen Redetzki, Suzan DelBene, John</td>
</tr>
<tr>
<td>Microsoft: Longhorn to arrive in 2005 - News - ZD...</td>
<td>3/26/2004 10:2</td>
<td>...</td>
<td>temporary internet files/content.ie5/entry120001/765</td>
<td>Salim Roukos</td>
<td><a href="mailto:marcuc@ISI.EDU">marcuc@ISI.EDU</a>, Susan Dumais</td>
</tr>
<tr>
<td>News: All about Longhorn</td>
<td>3/26/2004 10:2</td>
<td>...</td>
<td>temporary internet files/content.ie5/entry120001/765</td>
<td>Salim Roukos</td>
<td><a href="mailto:marcuc@ISI.EDU">marcuc@ISI.EDU</a>, Susan Dumais</td>
</tr>
<tr>
<td>X1 instantly searches files &amp; email. For Outlook, Ou...</td>
<td>3/24/2004 10:1</td>
<td>...</td>
<td>temporary internet files/content.ie5/entry120001/765</td>
<td>Salim Roukos</td>
<td><a href="mailto:marcuc@ISI.EDU">marcuc@ISI.EDU</a>, Susan Dumais</td>
</tr>
<tr>
<td>Enhanced Microsoft: Exhibit Offers Sneak Peek at ...</td>
<td>3/23/2004 12:2</td>
<td>...</td>
<td>temporary internet files/content.ie5/entry120001/765</td>
<td>Salim Roukos</td>
<td><a href="mailto:marcuc@ISI.EDU">marcuc@ISI.EDU</a>, Susan Dumais</td>
</tr>
</tbody>
</table>

**Last 7 days**

<table>
<thead>
<tr>
<th>Document</th>
<th>Date</th>
<th>Rank</th>
<th>Path</th>
<th>Author</th>
<th>Mail To</th>
</tr>
</thead>
<tbody>
<tr>
<td>prefecforprogram</td>
<td>3/22/2004 5:01</td>
<td>...</td>
<td>personal folders/ia4cl04</td>
<td>personal folders/ia4cl04</td>
<td>personal folders/ia4cl04</td>
</tr>
<tr>
<td>prefecforprogram</td>
<td>3/22/2004 4:19</td>
<td>...</td>
<td>personal folders/ia4cl04</td>
<td>personal folders/ia4cl04</td>
<td>personal folders/ia4cl04</td>
</tr>
</tbody>
</table>

**Last 30 days**

<table>
<thead>
<tr>
<th>Document</th>
<th>Date</th>
<th>Rank</th>
<th>Path</th>
<th>Author</th>
<th>Mail To</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Future of Information Filtering</td>
<td>3/20/2004 12:0</td>
<td>...</td>
<td>temporary internet files/content.ie5/entry120001/765</td>
<td>Salim Roukos</td>
<td><a href="mailto:marcuc@ISI.EDU">marcuc@ISI.EDU</a>, Susan Dumais</td>
</tr>
<tr>
<td>Slashdot</td>
<td>Google buys Pyra Labs</td>
<td>3/17/2004 3:22</td>
<td>...</td>
<td>temporary internet files/content.ie5/entry120001/765</td>
<td>Salim Roukos</td>
</tr>
<tr>
<td>Recap: Steven Levy/Newswire on IQ</td>
<td>3/15/2004 6:27</td>
<td>...</td>
<td>temporary internet files/content.ie5/entry120001/765</td>
<td>Salim Roukos</td>
<td><a href="mailto:marcuc@ISI.EDU">marcuc@ISI.EDU</a>, Susan Dumais</td>
</tr>
<tr>
<td>Homepage for HLT-NAACL 2003</td>
<td>3/15/2004 5:34</td>
<td>...</td>
<td>temporary internet files/content.ie5/entry120001/765</td>
<td>Salim Roukos</td>
<td><a href="mailto:marcuc@ISI.EDU">marcuc@ISI.EDU</a>, Susan Dumais</td>
</tr>
<tr>
<td>Pedro Domingos</td>
<td>3/10/2004 4:00</td>
<td>...</td>
<td>temporary internet files/content.ie5/entry120001/765</td>
<td>Salim Roukos</td>
<td><a href="mailto:marcuc@ISI.EDU">marcuc@ISI.EDU</a>, Susan Dumais</td>
</tr>
<tr>
<td>NLP: Multimedia Features and Information Retrieval</td>
<td>3/8/2004 3:55</td>
<td>...</td>
<td>temporary internet files/content.ie5/entry120001/765</td>
<td>Salim Roukos</td>
<td><a href="mailto:marcuc@ISI.EDU">marcuc@ISI.EDU</a>, Susan Dumais</td>
</tr>
<tr>
<td>Bates' Bibliography: Information Seeking, Indexing...</td>
<td>3/8/2004 3:54</td>
<td>...</td>
<td>temporary internet files/content.ie5/entry120001/765</td>
<td>Salim Roukos</td>
<td><a href="mailto:marcuc@ISI.EDU">marcuc@ISI.EDU</a>, Susan Dumais</td>
</tr>
</tbody>
</table>

**Older than 30 days**

<table>
<thead>
<tr>
<th>Document</th>
<th>Date</th>
<th>Rank</th>
<th>Path</th>
<th>Author</th>
<th>Mail To</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SIS Usage Experiences

Internal deployment
- ~3000 internal Microsoft users
- Analyzed: Free-form feedback, Questionnaires, Structured interviews, Log analysis (characteristics of interaction), UI expts, Lab expts

Personal store characteristics
- 5k - 500k items

Query characteristics
- Short queries (1.6 words)
- Few advanced operators or fielded search in query box (~7%)
- Many advanced operators and query iteration in UI (48%)
  - Filters (type, date); modify query; re-sort results

<table>
<thead>
<tr>
<th>Susan's (Laptop) World</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td><strong>N</strong></td>
<td><strong>Size</strong></td>
</tr>
<tr>
<td>Web</td>
<td>3k</td>
<td>0.2 Gb</td>
</tr>
<tr>
<td>Files</td>
<td>28k</td>
<td>23.0 GB</td>
</tr>
<tr>
<td>Mail</td>
<td>60k</td>
<td>2.2 Gb</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>91k items</td>
<td>25.4 Gb</td>
</tr>
<tr>
<td><strong>Index</strong></td>
<td></td>
<td>190 Mb</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+1.5 Mb/week</td>
</tr>
</tbody>
</table>
Important of people, time, and memory

- **People**
  - 25% of queries contained names
  - People in roles (to:, from:) vs. people as entities in text

- **Time**
  - Age of items opened
    - 5% today; 21% last week
    - 50% of the cases in 36 days
      - Web (11); Mail (36); Files (55)
  - **Date** most common sort field, even when Rank was the default
    - Support for episodic memory

- Few searches for “best” topical match ... many other criteria
SIS Usage Data, cont’d

Observations about unified access

- Metadata quality is variable
  - Email: rich, pretty clean
  - Web: little, available to application
  - Files: some, but often wrong

- Memory depends on abstractions
  - “Useful date” is dependent on the object!
    - Appointment, when it happens
    - File, when it is changed
    - Email and Web, when it is seen
  - “People” attribute vs. contains
    - To, From, Cc, Attendee, Author, Artist
Ranked list vs. Metadata (for personal content)

Why Rich Metadata?

- People remember many attributes in re-finding
  - Often: time, people, file type, etc.
  - Seldom: only general overall topic
- Rich client-side interface
  - Support fast iteration/refinement
  - Fast filter-sort-scroll vs. next-next-next
Re-finding on the Web

- 50-80% URL visits are revisits
- 30-40% of queries are re-finding queries

Table 1. A classification of different query types.

<table>
<thead>
<tr>
<th>All queries: 13,060 queries (100%)</th>
<th>Overlapping Click Queries – 5072 queries (39%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Equal Click Queries – 3777 (29%)</td>
</tr>
<tr>
<td></td>
<td>Single Identical Click 3737 (29%)</td>
</tr>
<tr>
<td></td>
<td>Multiple Identical Clicks 40 (&lt; 1%)</td>
</tr>
<tr>
<td></td>
<td>Some Common Clicks 1295 (10%)</td>
</tr>
<tr>
<td></td>
<td>No Common Clicks 7988 (61%)</td>
</tr>
<tr>
<td>Equal Query Queries 4256 (33%)</td>
<td>Navigational Queries 3100 (24%)</td>
</tr>
<tr>
<td></td>
<td>36 (&lt; 1%)</td>
</tr>
<tr>
<td></td>
<td>635 (5%)</td>
</tr>
<tr>
<td></td>
<td>485 (4%)</td>
</tr>
<tr>
<td>Different Query 8804 (67%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>637 (5%)</td>
</tr>
<tr>
<td></td>
<td>4 (&lt; 1%)</td>
</tr>
<tr>
<td></td>
<td>660 (5%)</td>
</tr>
<tr>
<td></td>
<td>7503 (57%)</td>
</tr>
</tbody>
</table>
Shell for WDS; publically available

**Features:**
- Search / Browse (faceted metadata)
- Unified Tagging
- In-Context Search
Phlat: Faceted metadata

- Tight coupling of search and browse
- Q $\rightarrow$ Results &
  - Associated metadata w/ query previews
  - 5 default properties to filter on (extensible)
  - Includes tags
- Property filters integrated with query
  - Query = words and/or properties
  - No stuck filters
- Search == Browse
Phlat: Tagging

- **Apply a single set** of user-generated tags to all content (e.g., files, email, web, rss, etc.)

- **Tagging interaction**
  - Tag widget or drag-to-tag

- **Tag structure**
  - *Allow* but do not *require* hierarchy

- **Tag implementation**
  - Tags directly associated with files as NTFS or MAPI properties
Phat: In-Context Search

- Selecting a result …
- Linked view to show associated tags
- Rich actions
  - Open, drag-drop, etc.
- Pivot on metadata
- “Sideways search”
- Refine or replace query
Phlat shell for Windows Desktop Search

- Tight coupling of searching/browsing
- Rich faceted metadata support
  - Including unified tagging across data types
- In-context search and actions

Download: http://research.microsoft.com/adapt/phlat
Web Search using Metadata

Many queries include implicit metadata:
- portrait of barak obama
- recent news about midwest floods
- good painters near redmond
- starbucks near me
- overview of high blood pressure

Limited support for users to articulate this.
Search in Context

- Search is not the end goal ...
- Support information access in the context of ongoing activities (e.g., writing talk, finding out about, planning trip, buying, monitoring, etc.)
  - Search always available
  - Search from within apps (keywords, regions, full doc)
  - Show results within app
  - Maintains “flow” (Csikszentmihalyi)
  - Can improve relevance
Documents as (a simple) Context

Proactive “query” specification depending on current document content and activities

- Recommendations
  - People who bought this also bought ...
- Contextual Ads
  - Ads relevant to page
- Community Bar
  - Notes, Chat, Tags, Inlinks, Queries
- Implicit Queries (IQ)
  - Also Y!Q, Watson, Rememberance Agent
Document Contexts (Implicit Query, IQ)

- Proactively find info related to item being read/created
  - Quick links
  - Related content

- Challenges
  - Relevance, fine
  - When to show? (useful)
  - How to show? (peripheral awareness)

Score = tf_doc / log(tf_corpus+1)

Quick links for People and Subject.

Top matches for this Implicit Query (IQ).
**PSearch:** Personalized Search (Even Richer Context)

- Today: People get the same results, independent of current session, previous search history, etc.
- PSearch: Uses rich client-side info to personalize results

- Building a user profile
- Personalized ranking
- When to personalize?
- How to personalize display?
Building a User Profile

- **Type of information:**
  - Explicit: Judgments, categories
  - Content: Past queries, web pages, desktop
  - Behavior: Visited pages, dwell time

- **Time frame:** Short term, long term

- **Who:** Individual, group

- **Where the profile resides:**
  - Local: Richer profile, improved privacy
  - Server: Richer communities, portability
**Personalized Ranking**

- Personal Rank = $f(\text{Cont}, \text{Beh}, \text{Web})$
  - Pers_Content Match: $\text{sim}(\text{result, user_content_profile})$
  - Pers_Behavior Match: visited URLs
  - Web Match: web rank
When to Personalize?

- Personalization works well for some queries, ... but not for others
- Framework for understanding when to personalize
  - Personal ranking
    - Personal relevance (explicit or implicit)
  - Group ranking
    - Decreases as you add more people
  - Gap is “potential for personalization (p4p)”
More Personalized Search

- **PSearch** - rich long-term context; single individual
- **Short-term session/task context**
  - Session analysis
  - Query: *ACL*, ambiguous in isolation
    - Natural language ... summarization ... *ACL*
    - Knee surgery ... orthopedic surgeon ... *ACL*
- **Groups of similar people**
  - Groups: Location, demographics, interests, behavior, etc.
  - Mei & Church (2008)
    - $H(\text{URL}) = 22.4$
    - Search: $H(\text{URL}|Q) = 2.8$
    - Personalization: $H(\text{URL}|Q, \text{IP}) = 1.2$
- Many models ... smooth individual, group, global models
Beyond Search - Gathering Info

- Support for more than retrieving documents
  - Retrieve -> Analyze -> Use
- Lightweight scratchpad or workspace support
  - Iterative and evolving nature of search
  - Resuming at a later time or on other device
  - Sharing with others
Beyond Search - Sharing & Collaborating

- **SearchTogether**
  - Collaborative web search prototype
  - Sync. or async. sharing w/ others or self

- Collaborative search tasks
  - E.g., Planning travel, purchases, understanding medical info; researching joint project or report

- Today little support
  - Email links, instant messaging, phone

- SearchTogether adds support for
  - Awareness (history, metadata)
  - Coordination (IM, recommend, split)
  - Persistence (history, summaries)

**SearchTogether**
Looking Ahead …

- Continued advances in scale of systems, diversity of resources, ranking, etc.
- Tremendous new opportunities to support searchers by
  - Understanding user intent
    - Modeling user interests and activities over time
    - Representing non-content attributes and relations
  - Supporting the search process
    - Developing interaction and presentation techniques that allow people to better express their information needs
    - Supporting understanding, using, sharing results
  - Considering search as part of richer landscape
Using Context to Support Searchers
Think Outside the IR Box(es)

User Context

Query Words

Ranked List

Document Context

Task/Use Context
Thank You!

- Questions/Comments ...


- Phlat, http://research.microsoft.com/adapt/phlat

References

- **Stuff I’ve Seen**
  - *Download: http://toolbar.live.com* and Vista Search

- **Phlat**
  - *Download: http://research.microsoft.com/adapt/phlat*

- **Memory Landmarks**

- **Personalized Search**

- **Implicit Queries**

- **Revisitation on Web**

- **InkSeine**
  - *Download: http://research.microsoft.com/inkseine/"

- **Search Together**
  - *Download: http://research.microsoft.com/searchtogether/"