Thinking Outside the (Search) Box

Susan Dumais
Microsoft Research

http://research.microsoft.com/~sdumais
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, many formats
  ■ Health, finance, travel ...

How it’s accessed
Supporting Searchers

- The search box
- Spelling suggestions
- Query suggestions
- Advanced search operators and options (e.g., "", +/-, site:, filetype:)
- Inline answers
- Richer snippets

But, we can do better …
understanding context
Search and Context

Research prototypes: extend search algorithmic, capabilities, and user experiences

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

- **Document/Domain Contexts:**
  - Metadata and search (SIS, Phlat)
  - Visualizing patterns in results (MemoryLandmarks, GridViz)
  - Dynamic information environments (DiffIE)

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

- Unified index of *stuff you’ve seen*
  - Many types of info (e.g., files, email, calendar, contacts, web pages, rss, im)
  - Index of content and metadata (e.g., time, author, title, size, usage)
  - Rich UI possibilities
  - Supports re-finding vs. finding

Vista Desktop Search (and XP, Live Toolbar)

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

The image shows a software interface with a table titled "Stuff I've Seen." The table contains columns for Document, Date, Rank, Path, Author, and Mail To. The interface includes options for filtering by date (e.g., Today, Yesterday, Last 7 days, Last 30 days, Older than 30 days) and includes check boxes for different categories of documents (All, Web Pages, Outlook, Files).

The table entries include various dates and document titles such as:
- Briefing: AP/Sue Dumais 5/6/2004 11:27 AM 28 personal folders/media Km Davis Susan Dumais
- Microsoft: Longhorn to arrive in 2005 - News - ZDNet 3/26/2004 10:22 AM 6 temporary internet files\content ie5\vty70oub
- News: All about Longhorn 3/26/2004 10:22 AM 8 temporary internet files\content ie5\0xubghir

The interface has a search bar with the text "information overload" and buttons for "Go," "Exact Match," and "Clear All."
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, people); modify query; re-sort results

<table>
<thead>
<tr>
<th>Susan's (Laptop) World</th>
<th>Type</th>
<th>N</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web</td>
<td>3k</td>
<td>0.2 Gb</td>
<td></td>
</tr>
<tr>
<td>Files</td>
<td>28k</td>
<td>23.0 GB</td>
<td></td>
</tr>
<tr>
<td>Mail</td>
<td>60k</td>
<td>2.2 Gb</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>91k items</td>
<td>25.4 Gb</td>
<td></td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td>190 Mb</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>+1.5 Mb/week</td>
<td></td>
</tr>
</tbody>
</table>
SIS Usage Data, cont’d

Characteristics of items opened

- File types opened
  - 76% Email
  - 14% Web pages
  - 10% Files

- Age of items opened
  - 5% today
  - 21% within the last week
  - 47% within the last month
  - 50% of the cases -> 36 days
    - Web: 11 days
    - Mail: 36 days
    - Files: 55 days

Log(Freq) = -0.68 * log(DaysSinceSeen) + 2.02
UI Usage

- Small effects of: Top/Side, Previews/NoPreviews
- Large effect of Sort Order:
  - **Date** by far the most common sort field, even for people who had best-match Rank as default
  - Importance of time
  - Few searches for “best” match; many other criteria ...

---

**SIS Usage Data, cont’d**
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, 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% page visits are re-visits
- 30-40% of queries are re-finding queries

Table 1. A classification of different query types.

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

Teevan et al., SIGIR 2007
Phlat: Search and Metadata

- **Phlat** (Prototype for Helpful Lookup And Tagging)
  - Shell for WDS; Publically available
  - Tightly couples search and metadata
- **Features:**
  - Search / Browse (metadata)
  - Unified Tagging
  - In-Context Search
Phlat: Faceted metadata
(for filtering, sorting, querying, tagging)

- Tight coupling of search and browsing
- Q → 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

HCIR: Oct 23, 2008
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.
  - "Sideways search"
  - Pivot on metadata
  - Refine or replace query
Phlat

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
Metadata and the Web

Many queries contain implicit metadata:

- thomas edison image portrait
- latest lasik techniques, canada
- good nursing programs in baltimore
- cheap digital camera
- overview of active directory domains

Limited support for users to articulate this
Dynamic Info Environments

MSR Homepage

1996

2007

HCIR: Oct 23, 2008
Dynamic Info Environments

Content Changes

User Visitation/ReVisitation

Today’s Browse and Search Experiences

But, ignores ...

HCIR: Oct 23, 2008
What We Did

**Content:**
- Crawled 55k pages every hour for 1 year
- Varying #users, #visits/user, inter-visit interval

**Behavior:**
- Analyzed revisitation patterns for >600k users for these 55k pages
- Surveyed 20 people for richer understanding of intent

**Examined:**
- **User revisitation** patterns
- **Page change** patterns
- Relations between change and revisitation
What We Found
Revisitation patterns

- Revisitations to pages are very common
  - 50-80% of pages
- What makes one page’s revisits different from another?
- Examined four characteristics

HCIR: Oct 23, 2008
What We Found
Change patterns

- 66% of the pages change
  - Change every 123 hours (avg.)
  - Change by 0.21 (avg. dice coeff.)

- Which pages change?
  - Popular pages, .com pages change most

- Which terms change?
  - Term longevity analyses
What We Found

Change patterns

1998

Susan Dumais

Senior Researcher, Decision Theory & Adaptive Systems Group, Microsoft Research

Mail: One Microsoft Way, Redmond WA 98052-6399, USA

Research Activities:

I am interested in algorithms in information retrieval, as well as general issues in human-computer interaction. Research in July 1997 focused on a variety of information access and management issues, including collaborative filtering, user modeling, and text categorization.

What's New:


2007

Susan Dumais

Principal Researcher, Adaptive Systems & Interaction Group, Microsoft Research

Mail: One Microsoft Way, Redmond WA 98052-6399, USA

Research Activities:

We're hiring at MSR and LiveLabs...

We're looking for great folks to advance the state-of-the-art and influence new products in the search arena. We have internships and permanent positions in several areas including: internet search, desktop search, personalization, and novel interfaces for search.

Workshops, Collaborations and Papers:

What We Found
Change patterns - rate of change

Figure 7. Renderings of the lifespan of elements on a number of pages (darker red blocks are shorter life spans) including a) boston.com, b) televisionwithoutpity.com (note the groups of similarly colored content), c) the DVD bestseller list on Amazon, d) gas prices in various cities on GasBuddy.com, and e) a list of earthquakes at the USGS. Not all blocks marked.
What We Found
Change patterns - for your visits
Search in Task Contexts

- 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
InkSeine: Active Note Taking

- Tablet application for active note taking
- Unifies ink, search and gather functions into a fluid workflow
- Note taking, enriched with:
  - Search from ink
  - Show results in app
  - Integrate results, links and clippings into notes
  - Maintain work flow
- “Inking for thinking”

Download: http://research.microsoft.com/InkSeine/
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
  - Context search, Notes, Chat, Tags, Inlinks, Queries
  - http://www.communitybar.net

- Implicit Queries (IQ)
  - Also Y!Q, Rememberance Agent, Watson, Query-free search

- Even more possibilities for context-driven retrieval w/ rich sensors and ubiquitous networks
Documents as Context (Implicit Query, IQ)

- Proactively find info relevant to item being read/created
  - Quick links
  - Matching content (several sources)

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

Background search on top k terms, based on user’s index —
Score = \( \frac{tf_{doc}}{\log(tf_{corpus} + 1)} \)

Quick links for People and Subject.

Top matches for this Implicit Query (IQ).

Dumais et al., SIGIR 2004
**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

---

**Step 1**: retrieve >> 10 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 = \( f(\text{Cont}, \text{Beh}, \text{Web}) \)
  - P_Content Match: \( \text{sim}(\text{result}, \text{user}_\text{content}_\text{profile}) \)
  - P_Behavior Match: visited URLs and sites
  - 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)”
How to Personalize Display

- Presenting results
  - Inline display (for demo)
    - Also: tabs, slider, fisheye, metadata
  - Interleave results (for evaluation)
  - Behind the scenes (for the curious)
  - Balance consistency, novelty

- Summarizing results
  - Highlight results that were seen before
  - Highlight new result content
  - Personalized snippets
More “Personalized” Search

- **PSearch** - rich long-term context; single individual
- **Short-term session/task content**
  - Query: *ACL*, ambiguous in isolation
    - austin music ... tickets alison krauss ... ACL
    - natural language processing ... summarization ... ACL
    - knee surgery ... orthopedic surgeon ... ACL
- **Groups of similar people**
  - Groups: Location, demographics, interests, behavior, etc.
    - Freyne & Smyth (2006); Smyth (2007); Teevan & Morris (2008)
    - Mei & Church (2008)
      - H(URL) = 22.4
      - Search: H(URL|Q) = 2.8
      - “Personalization”: H(URL|Q, IP) = 1.2
  - Many models ... smooth individual, group, global models
Beyond Search - Gathering Info

- Support for more than “retrieving” documents
  - Analyze -> Use -> Share
  - Exploratory search

- Lightweight scratchpad or workspace support
  - Iterative and evolving nature of search
  - Resuming at a later time or on other device
  - Sharing with others

ScratchPad
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, events; 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)

**Download:** http://research.microsoft.com/searchtogether

---

**Figure 1.** The SearchTogether client: (a) integrating messaging, (b) query awareness, (c) current results, (d) recommendation queue, (e)(f)(g) search buttons, (h) page-specific metadata, (i) toolbar, (j) browser
Looking Ahead …

- Continued advances in scale of systems, diversity of resources and quality of ranking, etc.
- Tremendous new opportunities to support information retrieval and analysis by …
  - Understanding user intent
    - Representing non-content attributes and relations
    - Modeling user interests and activities over time
  - Supporting the search process
    - Developing interaction and presentation techniques that allow people to better express their information needs
    - Supporting analysis, use and sharing of results
  - Considering search as part of richer landscape
Thinking Outside the (Search) Box

User Context

Query Words

Ranked List

Document Context

Task/Use Context
Thank You!

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