Desktop Search

Susan Dumais
Microsoft Research

http://research.microsoft.com/~sdumais
Outline

- Search ≠ Web Search
  - E.g., Desktop Search, and many verticals
- Desktop Search ≈ My Stuff
- Stuff I’ve Seen (SIS)
  - Case study: Research prototype system, deployment experiences, usage data
  - MSN Desktop Search; MS Vista Search (http://toolbar.msn.com)
- Future directions
  - Contextualized search
  - Personalized search
My Stuff

- Information acquisition vs. access
  - Easy to create or encounter lots of information
    - Types: email, docs, web pages, calendars, pictures, music, etc.
    - Amount: 100+ gig drives
  - Hard to organize
  - And, even harder to re-find

- Information discovery vs. recovery
  - Many tools for finding information (discovery)
  - Fewer tools for keeping information (recovery)
  - Yet, many tasks involve re-using information
Desktop Search Today

- Information silos
  - Many locations, interfaces for finding things (e.g., web, mail, contacts, docs, photos, notes)

- Often slow

SIMS 141: October 17, 2005
Desktop Search With SIS

- Unified index of **stuff you’ve seen**
- All types of info (e.g., files, email, calendar, contacts, web pages, rss, im)
- Index **content plus metadata** (e.g., time, author, title, size, usage)
- Automatic and immediate update of index
- Rich UI possibilities, since it’s **your content** (e.g., consider usage)

- Get back to information you’ve seen
- Recovery vs. discovery
Related Work

- Research projects
  - Haystack [Adar et al., 1999; Huynh et al. 2002]
  - Keeping Found Things Found [Jones et al., 2001]
  - MyLife Bits [Gemmell, Bell et al., 2002]
  - Lifestreams/Scopeware [Fertig, Freeman, Gelemter, 1996]

- Commercial systems/software
  - OS: Mac OS X Spotlight, MS Vista Search
  - DS Apps: Enfish, dtSearch, Copemic, X1/Yahoo!, G-DS, MSN-DS, etc.

- What’s new with SIS …
  - Full content and metadata for many different sources
  - Extensible architecture (gather, filter, word break)
  - Focus on user interface and user experience
  - Iterative design guided by usage and experimental data
SIS Design Principles

- Indexing experience ...
  - No additional work is required
  - User sees something, and it gets indexed

- Retrieval experience ...
  - Fast, flexible
  - Interactive refinement
    - Sort and filter on metadata
    - Note: Sort/filter automatically triggers query
  - UI innovations
    - Previews, Top/Side, Sort order
    - Richer visualizations
SIS Demo

Stuff I've Seen

File View Options Help

2767 rows returned

Document Date Path Author Mail To

Updated: Stuff I've Seen... Fast mail... 11/4/2002 1:00 PM  mailbox - susan dumais/sent items  Susan Dumais  Marc Olson, Will Kennedy, Jensen Harris, ... When Monday, November 04, 2002 1:00 PM (PST-0700) Pacific Time (US & Canada, Tijuana) Where: 18/2488 Update Time since Marc is OFF Friday. We are working on a prototype called Stuff I've Seen (SIS). SIS provides an integrated index of all the things you look at.

Today

stuff I've seen - outlook

11/1/2002 5:18 PM  d'personaKopiers/mics mail  Susan Dumais

Stuff I've Seen Susan Dumais, Ed Cutrell, W Cadz, Savin Janke, Raman Sain, Microsoft Research Search Today... and tomorrow... SIS Details Unified index of stuff you've seen Web pages, office docs, email... and more. A full-text index of content plus metadata attributes (e.g., creation time, author, Hot new hits elsewhere)

RE: Local store vs. Server hits in SIS

11/1/2002 5:15 PM  mailbox - susan dumais/sent items  Susan Dumais  Edward Cutrell

Ed - Can you send the xls file? I can't seem to cut and paste the figure below into ppt. Thanks, Sue. It —— Original Message —— From: Edward Cutrell  Sent: Friday, November 01, 2002 3:34 PM (PST-0700) Pacific Time (US & Canada, Tijuana) Where: 18/2488 We are working on a prototype called Stuff I've Seen (SIS). SIS provides an integrated index of all the things you look at, including files, web pages, email, and more. A full-text index of content plus metadata attributes (e.g., creation time, author, Hot new hits elsewhere).

Local store vs. Server hits in SIS

11/1/2002 3:33 PM  mailbox - susan dumais/inbox  Edward Cutrell  Susan Dumais, Adrian Klein

Some data from our logs: 6% of all mail items opened (aggregated across all users), 43% were local hits (i.e., hits from the local store vs. Server hits in SIS). I determined the % of all mail opened that was local and the percent from server. Then I did a histogram of these values for Stuff I've Seen... Fast mail indexing and... 11/1/2002 3:00 PM  mailbox - susan dumais/sent items  Susan Dumais  Marc Olson, Will Kennedy, Jensen Harris, ...

When Friday, November 01, 2002 3:00 PM (PST-0700) Pacific Time (US & Canada, Tijuana) Where: 18/2488 We are working on a prototype called Stuff I've Seen (SIS). SIS provides an integrated index of all the things you look at, including files, web pages, email, and more. A full-text index of content plus metadata attributes (e.g., creation time, author, Hot new hits elsewhere).

RE: Things have quieted down again

11/1/2002 12:47 AM  mailbox - susan dumais/sent items  Susan Dumais  Eugene Samsonov, Raman Sain, (Eoset D...

If I simply net stop/delay "is search" I get 12 gig I/O read, and 17 meg I/O write, the exact amount varies a bit from time to time. Incremental crawl is all that is happening. However, it almost seems as if someone is reading all the dir content to check for file changes. I'm not sure if this is happening or not, but it's possible. The incremental crawl is all that is happening.

RE: Things have quieted down a...


If I simply net stop/delay "is search" I get 12 gig I/O read, and 17 meg I/O write, the exact amount varies a bit from time to time. Incremental crawl is all that is happening.
Indexing infrastructure uses MS Search components (note: IR platform)
- **Gatherer** - interface to content sources, e.g., files, http, MAPI
- **Filters** - decode different file types, e.g., word, powerpoint, html, pdf, journal notes
- **Tokenizer** - break into words, including date normalization, stemming, etc.
- **Indexer** - standard inverted index
- **Retriever** - Boolean, best match (Okapi), fielded

**User interface**
**Client side indexing and storage**
Evaluating SIS

- Internal deployment
  - ~3000 users
  - Users include: program management, test, sales, development, administrative, executives, etc.

- Research techniques
  - Free-form feedback
  - Questionnaires; Structured interviews
  - Usage patterns from log data
  - UI experiments (randomly deploy different versions)
  - Lab studies for richer UI (e.g., timeline, trends)
SIS Usage Data

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>Type</th>
<th>N</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<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><strong>91k items</strong></td>
<td><strong>25.4 Gb</strong></td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td>190 Mb</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+1.5 Mb/week</td>
</tr>
</tbody>
</table>
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
### Top vs. Side Views

#### User Interface (UI) Alternatives

<table>
<thead>
<tr>
<th>Sort By Date vs. Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
</tr>
<tr>
<td>Rank</td>
</tr>
</tbody>
</table>

#### Previews vs. Not

<table>
<thead>
<tr>
<th>Document</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 Okapi Rank as default
  - Importance of time
  - Few searches for “best” match; many other criteria …
Metadata vs. Best-match list

Google Desktop Search results

Desktop:
- 11 emails
- 8 files
- 22 web history
- 3 chats

San Diego Vacation

San Diego Vacation planning


www.sandiego-usa.com/ - 2 cached - 4:23 PM

*mark prince: When I'm in San Diego I'll visit...

I'm excited about my vacation in San Diego. I'll make sure to hit all the cool spots. Do you have any hotel recommendations?

petersonwilliams 10:18am
Observations about unified access

- **Metadata quality is variable**
  - Email: rich, pretty clean
  - Web: little, not very useful for retrieval
  - Files: some, but often wrong

- **Need abstractions**, e.g., "Useful date", "People", "Picture"
  - Initially, used ‘date seen’
  - But …
    - Appointment, when it happens
    - File, when it is changed
    - Email and Web, when it is seen
  - “Useful date” abstraction
Ease of finding information

- Easier after SIS for web, email, files
- Non-SIS search decreases for web, email, files

Additional benefits

- “The ability to find misfiled documents and email has been extremely helpful.” -- A sales executive in Washington D.C.
- “Thanks again for the MARVELOUS tool! I find myself unable to live without it! It saves me at least 10-15 minutes a day looking for information; saves even more time not having to file things. It makes me more effective, as more time goes to thinking and deciding, and less to overhead.” -- An executive in Redmond.
SIS, Timeline w/ Landmarks

- SIS: time as important access cue
- Importance of “landmarks” in human memory
- Identify and use landmarks to facilitate information management and search
- Timeline interface, augmented with landmarks
  - General landmarks: holidays, world events
  - Personal landmarks: important photos, appointments
  - Heuristics or Bayesian models to identify memorable events
SIS, Timeline w/ Landmarks

Distribution of Results Over Time

Search Results

Memory Landmarks
- General (world, calendar)
- Personal (appts, photos)
<linked by time to results>
Landmarks, key dependencies
(from learned graphical model)
SIS, Visualizing Patterns

- Summarize the results of a search
  - Abstraction beyond individual results

- Grid-based design
  - Axes represent topic, time, people, etc.
  - Cells encode frequency, recency

- Supports activities like:
  - What newsgroups are active (on topic x)?
  - What people are active, authoritative (on topic x)?
  - When did I last interact w/ people?
SIS, Visualizing Patterns
SIS, Grid vs. List Experiment

Grid View

List View
Contextualized Search

- Search is not the end goal …
- Need to support information management in the context of ongoing work activities
- Search always available
- Search from within apps (select keywords/regions, full docs)
- Show results in context of apps
Background search on top $k$ interesting terms from message, based on user’s index —
Score = $\frac{tf_{doc}}{\log(tf_{corpus}+1)}$

Quick searches for people associated with the message and Subject.

Top N hits for this Implicit Query (IQ). Open items directly.

Query autofills with IQ terms.

Go to SIS for detailed search. Query autofills with IQ terms.
Personalized Search

Today:
- All users get the same results, independent of previous search history, current context, etc.

Personalized Search Prototype:
- Use rich client-side info to personalize search results
- includes personal content (e.g., SIS index), Activities
- No profile setup or maintenance required
- All profile storage and processing client-side
Desktop Search Summary

- Desktop search with SIS
  - Unified index to *stuff you’ve seen*
  - Heterogeneous content: files, email, web, etc.
  - Fast and flexible access

- Supports new capabilities for personal information management
  - Rich metadata vs. single hierarchy or ranked list
  - Landmarks, patterns, implicit queries, etc.

- New directions
  - Contextualized search
  - Personalized search
Vannevar Bush’s Vision


- Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, "memex" will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory.

Memex, 1945

NLS/Augment, 1968

DesktopSearch, 2003
Thank You

- Questions/Comments …


- MSN Toolbar and Desktop Search, http://toolbar.msn.com