

Desktop Search

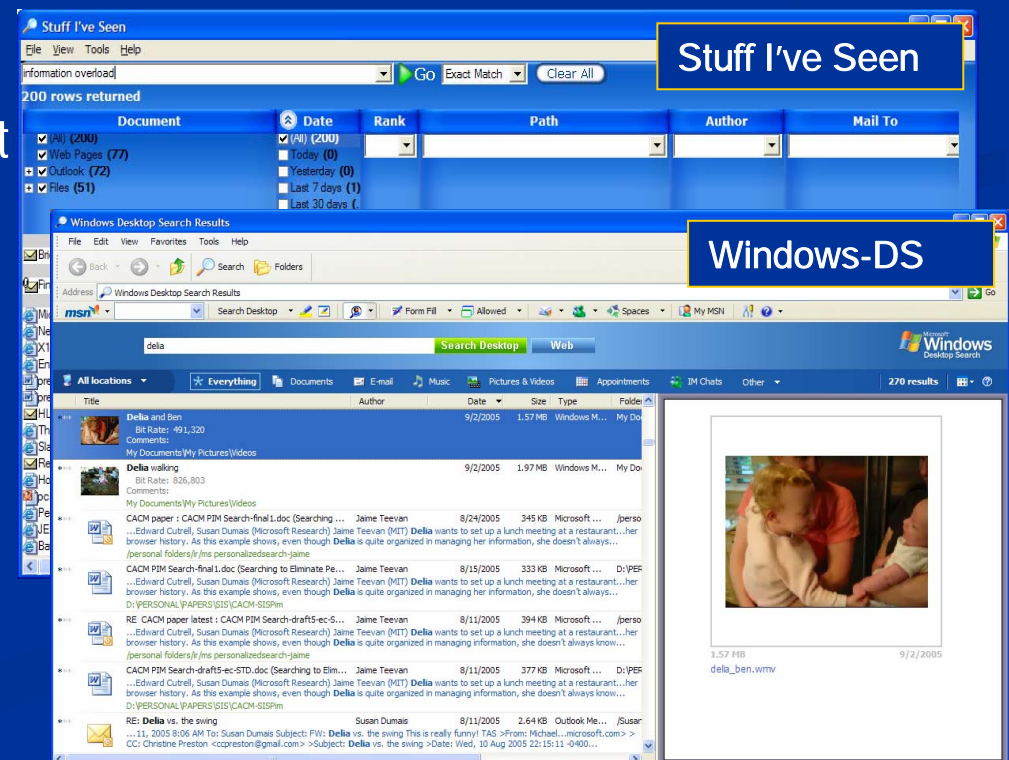
Susan Dumais

Microsoft Research

<http://research.microsoft.com/~sdumais>

Outline

- Search \neq Web Search
 - E.g., Desktop Search, and many verticals
- Desktop Search \approx *My Stuff*
- Stuff I've Seen (SIS)
 - *Case study*: Research prototype system, deployment experiences, usage data
 - \rightarrow MSN Desktop Search; MS Vista Search (<http://toolbar.msn.com>)
- Future directions
 - Contextualized search
 - Personalized search

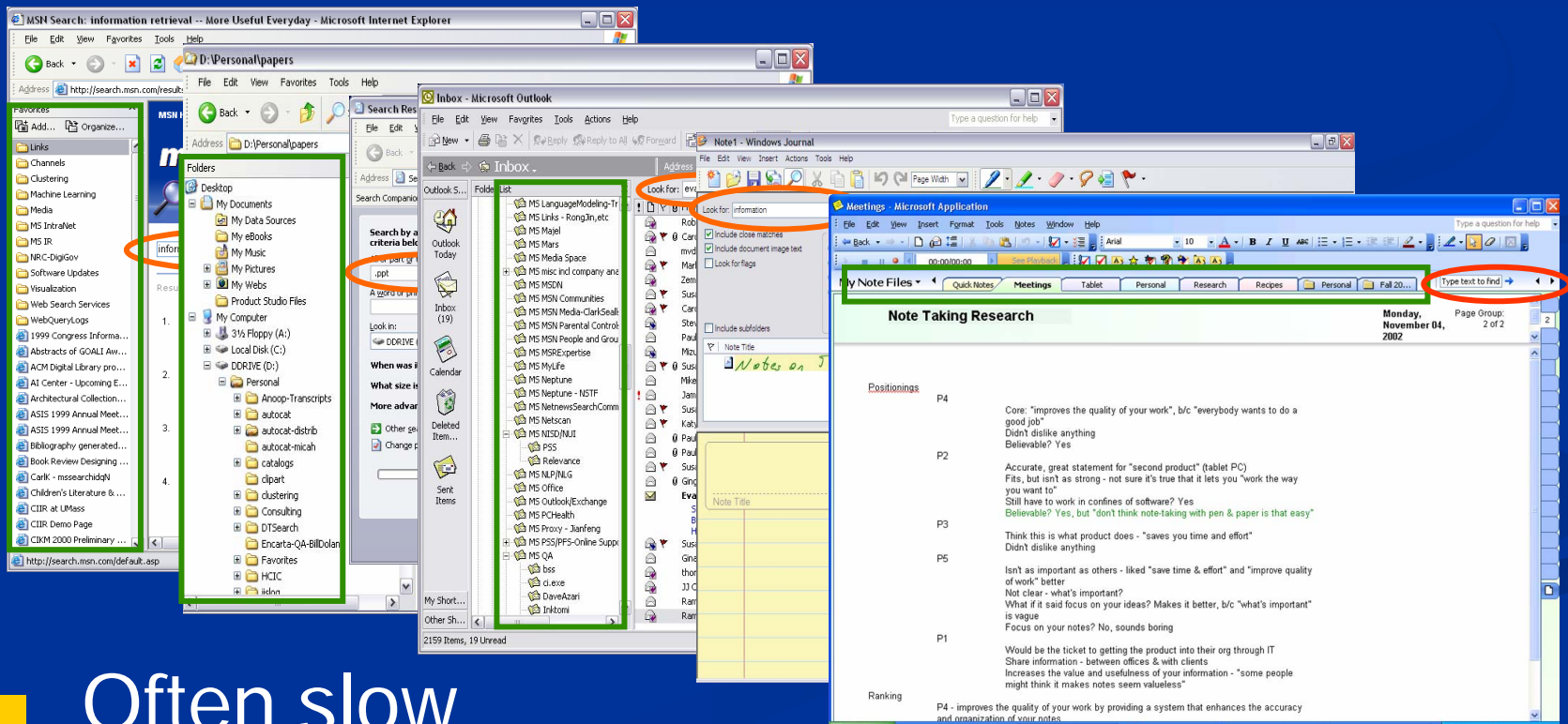


My Stuff

- 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)

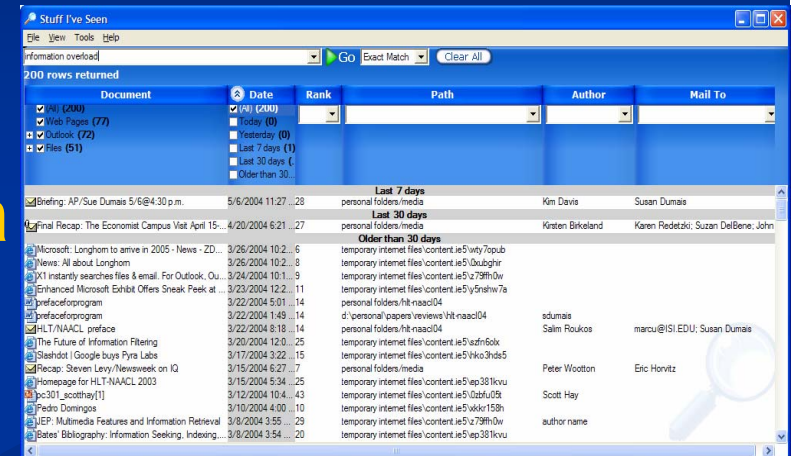


■ Often slow

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, Gelernter, 1996]
- Commercial systems/software
 - OS: Mac OS X Spotlight, MS Vista Search
 - DS Apps: Enfish, dtSearch, Copernic, 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

sis Exact Match

2767 rows returned

Document	Date	Path	Author	Mail To
<input checked="" type="checkbox"/> (All) [2767]	<input checked="" type="checkbox"/> (All) [2767]			
<input checked="" type="checkbox"/> Web Pages [7]	<input type="checkbox"/> Today [25]			
+ <input checked="" type="checkbox"/> Outlook [2625]	<input type="checkbox"/> Yesterday [17]			
+ <input checked="" type="checkbox"/> Files [135]	<input type="checkbox"/> Last 7 days [96]			
	<input type="checkbox"/> Last 30 days [486]			
	<input type="checkbox"/> Older than 30 days [...]			
Future				
Updated: Stuff I've Seen ... Fast mail ind... 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-2:00 PM (GMT-08:00) Pacific Time (US & Canada); Tijuana Where: 18/2498 UPDATING since Marc is OOF Friday. We are working on a prototype called Stuff I've Seen (SIS). SIS provides an integrated index of all the things you look				
Today				
stuff i've seen - outlook	11/1/2002 5:18 PM	d:\personal\papers\misc ms-ir	Susan Dumais	
Stuff I've Seen Susan Dumais, Ed Cutrell JJ Cadiz, Gavin Janke, Raman Sarin Microsoft Research Search Today ... and Tomorrow SIS Details Unified index of stuff you've seen Web pages, office docs, email ... and more Full-text index of content plus metadata attributes (e.g., creation time, author,				
RE: Local store vs. Server hits in SIS	11/1/2002 5:16 PM	mailbox - susan dumais/sent items	Susan Dumais	Edward Cutrell
Ed - Can you send the xls file? I can't seem to cut and past the figure below into ppt. Thanks, Sue -----Original Message----- From: Edward Cutrell Sent: Friday, November 01, 2002 3:34 PM To: Susan Dumais; Adrian Klein Cc: JJ Cadiz Subject: Local store vs. Server hits in SIS				
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: Of all mail items opened (aggregated across all users), 43% were local files (pst, etc), & 57% were located on the server. On a per-user basis, 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-4:00 PM (GMT-08:00) Pacific Time (US & Canada); Tijuana Where: 18/2498 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				
RE: Things have quieted down again	11/1/2002 10:47 AM	mailbox - susan dumais/sent items	Susan Dumais	Eugene Samsonov; Raman Sarin (Excell D...
If I simply net stop/start "rs search", I get 12 gig I/O read, and 174 meg I/O write <the exact amount varies a bit from time to time>. Incremental crawl is all that is happening. However, if almost seems as if msearch is reading all the disk content to check for file changes. I'm				
RE: Things have quieted down a...	11/1/2002 10:47 AM	mailbox - susan dumais/inbox	Susan Dumais	Eugene Samsonov; Raman Sarin (E...
If I simply net stop/start "rs search", I get 12 gig I/O read, and 174 meg I/O write <the exact amount varies a bit from time to time>. Incremental crawl is all that is happening.				

SIS Architecture

- 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

Sue's (Laptop) World		
Type	N	Size
Web	3k	0.2 Gb
Files	28k	23.0 GB
Mail	60k	2.2 Gb
Total	91k items	25.4 Gb
Index		190 Mb
		+1.5 Mb/week

Query characteristics

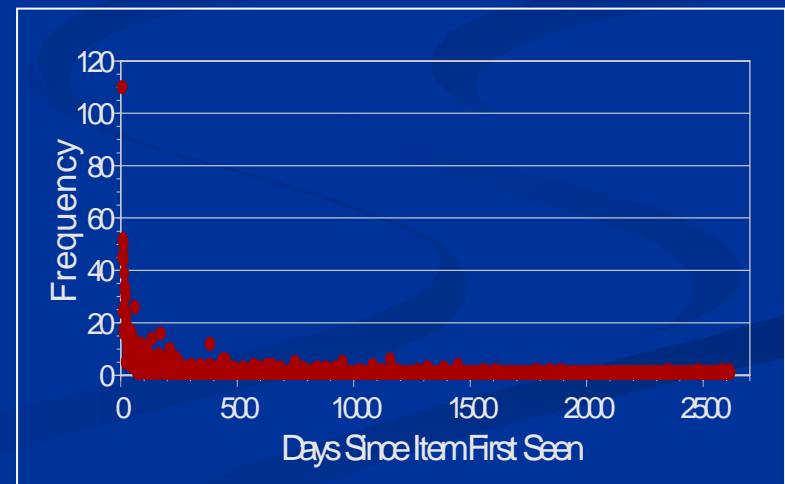
- Short queries (1.6 w)
- Few advanced operators
- Many advanced operators and query iteration in UI (48%)
 - Filters (type, date); modify query; re-sort results

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

$$\text{Log}(\text{Freq}) = -0.68 * \text{log}(\text{DaysSinceSeen}) + 2.02$$



Top vs. Side Views

3038 rows returned

3038 rows returned

287 rows returned

190 rows returned

Indexing... (8941)

Sort By Date vs. Rank

Document Date

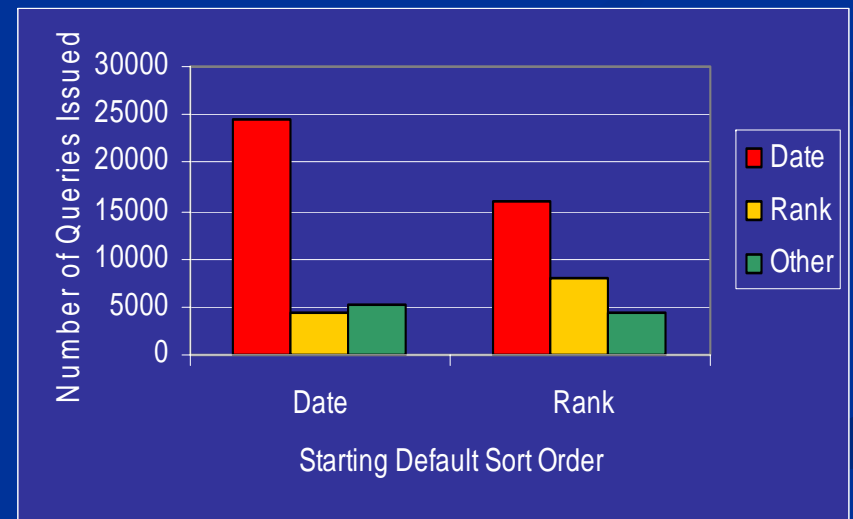
Document Date Rank Path Author Mail To

Document	Date	Rank	Path	Author	Mail To
[All] (287)	[All] (287)				
Web Pages (9)	Today (0)				
Outlook (182)	Yesterday (0)				
Files (96)	Last 7 days (0)				
	Last 30 days (7)				
	Older than 30 day				
RE: FW: HLT-NAACL Program C...	7/3/2003 12:27 PM				
RE: FW: HLT-NAACL Program Co...	7/3/2003 10:56 AM				
RE: HLT Program Question	6/25/2003 10:33 AM				
Re: HLT Program Question	6/25/2003 10:32 AM				
Javed Aslam: Publications	6/17/2003 2:51 PM				
Summary of Cochlea	6/10/2003 4:03 PM				
SearchDay #546 - To Google, an...	6/9/2003 10:43 AM				
gnidviz_ozchi2003_v1	6/6/2003 1:48 PM				
gnidviz_ozchi2003_v1	6/6/2003 1:48 PM				
knowledge_interchange_final	6/2/2003 2:33 PM				
knowledge interchange - executiv...	6/2/2003 2:33 PM				
knowledge_interchange_final	6/2/2003 2:33 PM				
knowledge interchange - executiv...	6/2/2003 2:33 PM				
Coverage: Seattle Times (Memor...	5/26/2003 11:53 AM				
Recap: Seattle Times Interview/E...	5/23/2003 4:35 PM				
tasist03	5/19/2003 8:57 AM				
[All] (190)	[All] (190)				
Web Pages (9)	Today (0)				
Outlook (85)	Yesterday (0)				
Files (96)	Last 7 days (0)				
	Last 30 days (0)				
	Older than 30 day (190)				
t0.txt	4/18/2003 10:34 AM	395	my documents\sis\datafomui-032103		
logs-sis queryissued_all.txt	4/16/2003 7:14 PM	238	my documents\sis\datafomui-032103		
queries_for_nui.txt	3/28/2003 10:19 AM	193	my documents\sis\datafomui-032103		
queries_for_nui.txt	3/28/2003 10:25 AM	193	personal folders\ir\ms nisd\hui\relevance		
queries.txt	3/28/2003 10:02 AM	185	my documents\sis\datafomui-032103		
tr grant proposal, intro, for review	10/23/2002 6:12 AM	106	personal folders\ir\keepingfoundthingsfound		
grant, with mike's comments	9/5/2000 4:37 PM	75	personal folders\ir\keepingfoundthingsfound		
grant, minus graph	7/10/2000 11:32 PM	64	personal folders\ir\keepingfoundthingsfound		
grant, minus graph3	7/11/2000 5:34 PM	64	personal folders\ir\keepingfoundthingsfound		
bad response from server	2/7/2003 3:54 PM	60	my documents\differenceengine\fileswebqueries		
logs-sis execute-2	3/21/2003 5:16 PM	55	my documents\sis\datafomui-032103	cutrell	
2nd brain	2/8/2002 2:57 PM	53	personal folders\ir\ms stuffiveseen		
2nd brain	8/26/2002 1:44 PM	53	personal folders\ir\ms stuffiveseen/feedback-alp...		
enclosure b - revised in response ...	7/7/1999 1:09 PM	52	personal folders\nrc-diggovt		
grant 16	9/9/2000 3:38 PM	52	personal folders\ir\keepingfoundthingsfound		
grant 17finalsubmitted	12/18/2000 9:57 AM	45	my documents\papers\keepingfoundthingsfound	William Jones	
grant 17	12/18/2000 10:58 AM	45	personal folders\ir\keepingfoundthingsfound		

SIS Usage Data, cont'd

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

Stuff I've Seen

File View Options... Help

sis Exact Match

2767 rows returned

Document	Date	Path	Author	Mail To
<input checked="" type="checkbox"/> [All] (2767)	<input checked="" type="checkbox"/> [All] (2767)			
<input checked="" type="checkbox"/> Web Pages (7)	<input type="checkbox"/> Today (25)			
+ <input checked="" type="checkbox"/> Outlook (2625)	<input type="checkbox"/> Yesterday (17)			
+ <input checked="" type="checkbox"/> Files (135)	<input type="checkbox"/> Last 7 days (96)			
	<input type="checkbox"/> Last 30 days (486)			
	<input type="checkbox"/> Older than 30 days [...]			

Future

Updated: Stuff I've Seen ... Fast mail ind... 11/14/2002 1:00 PM mailbox - susan dumais/sent items S
When: Monday, November 04, 2002 1:00 PM-2:00 PM (GMT-08:00) Pacific Time (US & Canada); Tijuana. Where: Friday. We are working on a prototype called Stuff I've Seen (SIS). SIS provides an integrated index of all the things...

Today

stuff i've seen - outlook 11/11/2002 5:18 PM d:\personal\papers\misc ms-ir S
Stuff I've Seen Susan Dumais, Ed Cutrell, JJ Cadiz, Gavin Janke, Raman Sarin, Microsoft Research Search Today pages, office docs, email ... and more Full-text index of content plus metadata attributes (e.g., creation time, author, ...)

RE: Local store vs. Server hits in SIS 11/11/2002 5:16 PM mailbox - susan dumais/sent items S
Ed - Can you send the xls file? I can't seem to cut and past the figure below into ppt. Thanks, Sue #1-----Original Message-----
From: Susan Dumais
To: Ed Cutrell; Susan Dumais; Adrian Klein
Cc: JJ Cadiz
Subject: Local store vs. Server hits in SIS

Local store vs. Server hits in SIS 11/11/2002 3:33 PM mailbox - susan dumais/inbox E
Some data from our logs: Of all mail items opened (aggregated across all users), 43% were local files (pst, etc), & 5% 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 ... 11/11/2002 3:00 PM mailbox - susan dumais/sent items S
When: Friday, November 01, 2002 3:00 PM-4:00 PM (GMT-08:00) Pacific Time (US & Canada); Tijuana. Where: ... called Stuff I've Seen (SIS). SIS provides an integrated index of all the things you look at, including files, web page ...

RE: Things have quieted down again 11/11/2002 10:47 AM mailbox - susan dumais/sent items S
If I simply net stop/start "rs search", I get 12 gig I/O read, and 174 meg I/O write <the exact amount varies a bit fr ... However, it almost seems as if msearch is reading all the disk content to check for file changes. I'm ...

RE: Things have quieted down a... 11/11/2002 10:47 AM mailbox - susan dumais/inbox S
If I simply net stop/start "rs search", I get 12 aia I/O read, and 174 mea I/O write <the exact amount varies a bit fr ...

Google Desktop Search results

Web Images Groups News Froogle Local Desktop more...

Google Desktop [Desktop Preferences](#) [Remove Items](#)

Desktop: All - 7 emails - 8 files - 22 web history - 3 chats 1-10 of about 3,356 (0.02s)

[Sort by relevance](#) [Sorted by date](#)

[Going on vacation!](#)
I'm going on vacation to San Diego! I'll be back on August 5th, so let me know if you want me to bring you back anything. I'm really excited, I can't wait. See you soon, Charles
James Martin - 2 messages - 10:35am

[San Diego Vacation planning](#)
San Diego Vacation planning Page 1 San Diego Vacation planning -Go to LegoLand -Hit the beach -Swim at the hotel pool -Go to great restaurants -Go to SeaWorld -Sleep Sight see My Documents\San Diego Vacation planning pdf - Open folder - 1 cached - 4:22pm

[san diego.jpg](#)
320 x 198 pixels, 15k
My Documents\san_diego.jpg - Open folder - 4:23pm

[San Diego Hotel, San Diego Hotel Guide](#)
... San Diego restaurants, san Diego attractions, San Diego zoo, San Diego real estate, San Diego golf courses, San Diego vacation rentals, San Diego nightlife.
...
www.san-diego.us/ - 2 cached - 10:26am

["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?
peterwilliams - 10:18am

SIS Usage Data, cont'd

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

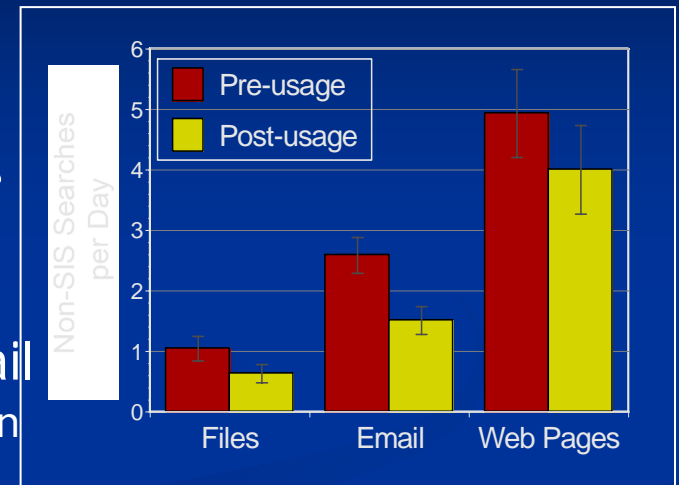
SIS Usage Data, cont'd

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.

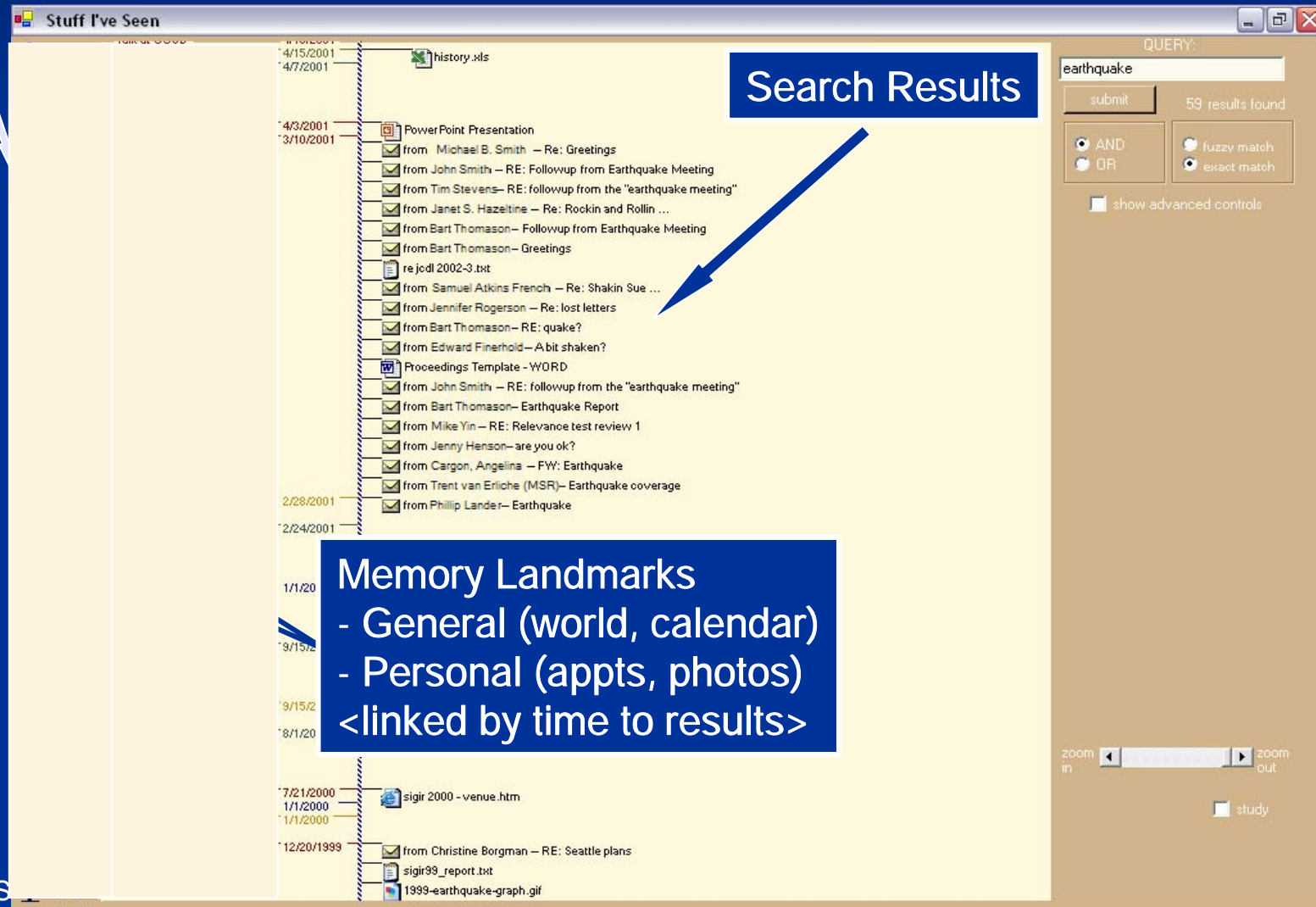


SIS, Timeline w/ Landmarks

- 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

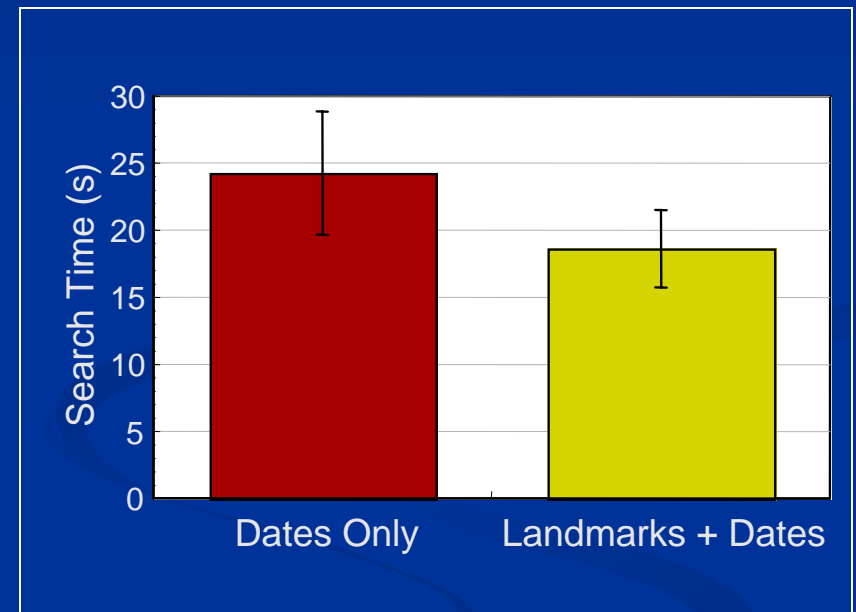
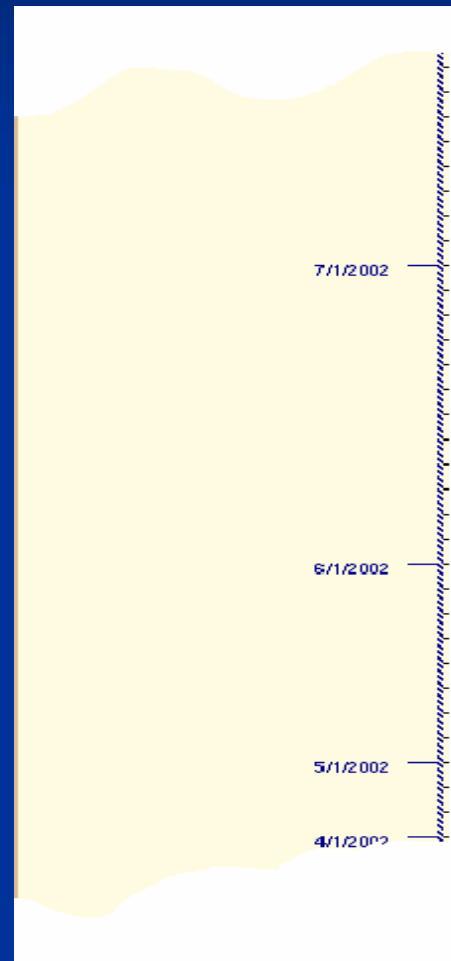


SIS, Timeline Experiment

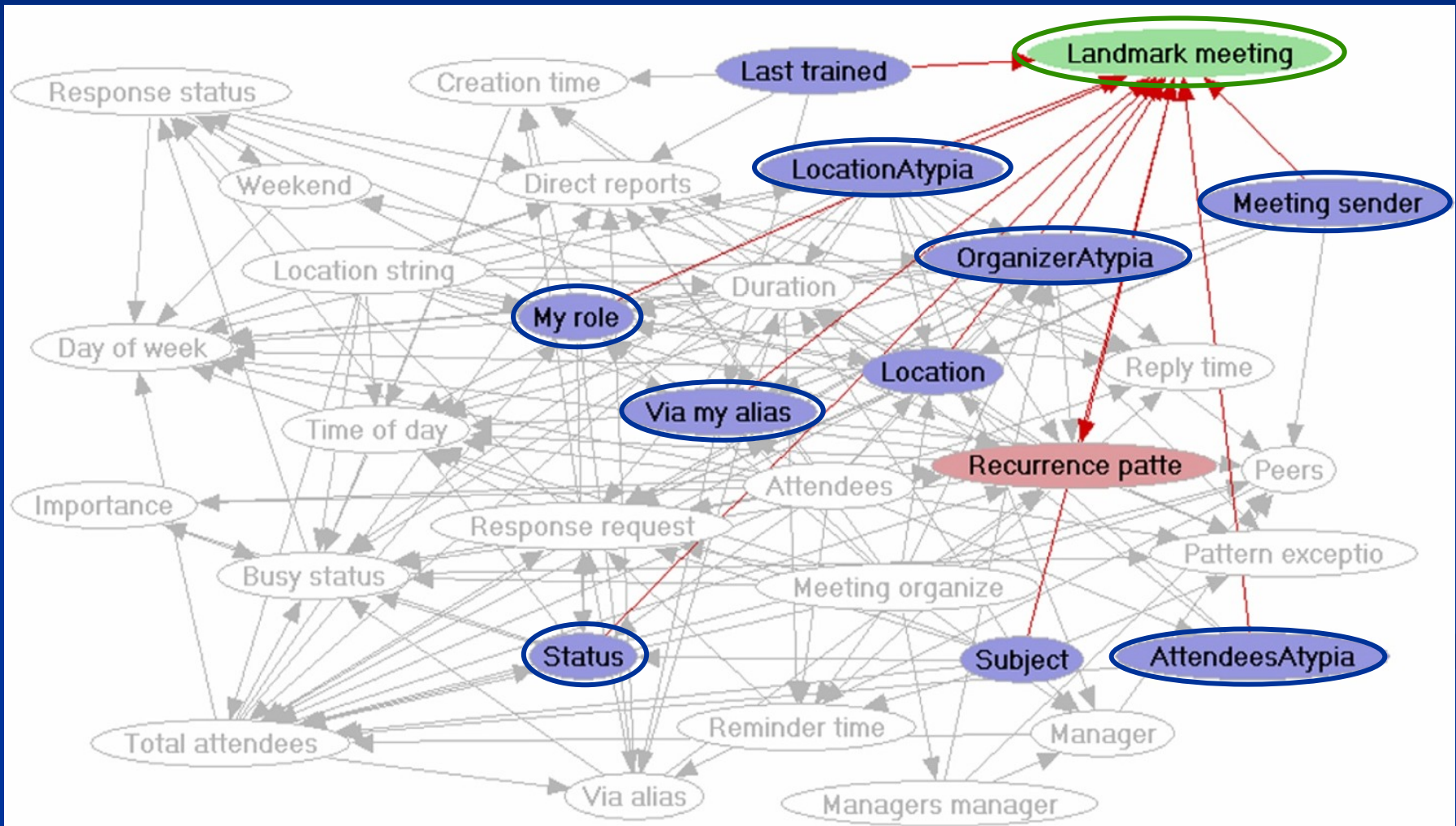
With Landmarks



Without Landmarks



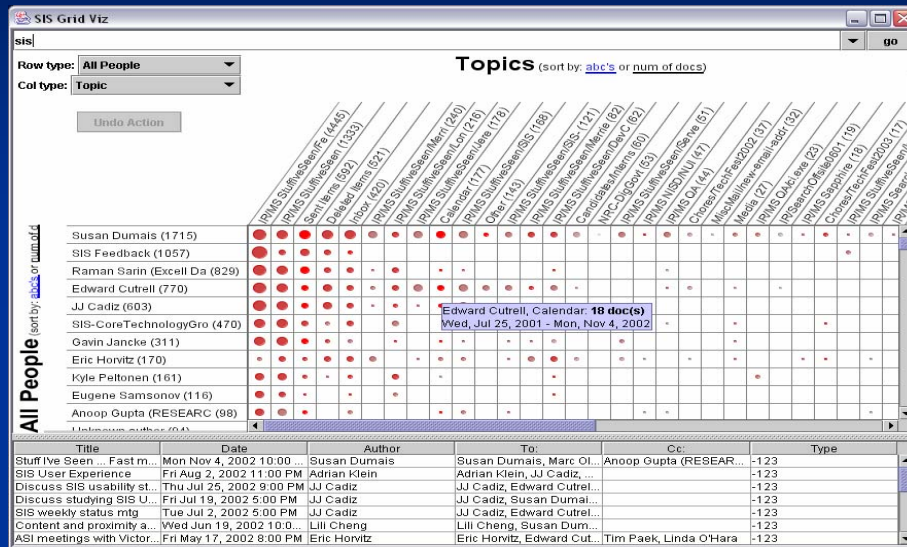
Landmarks, key dependencies (from learned graphical model)



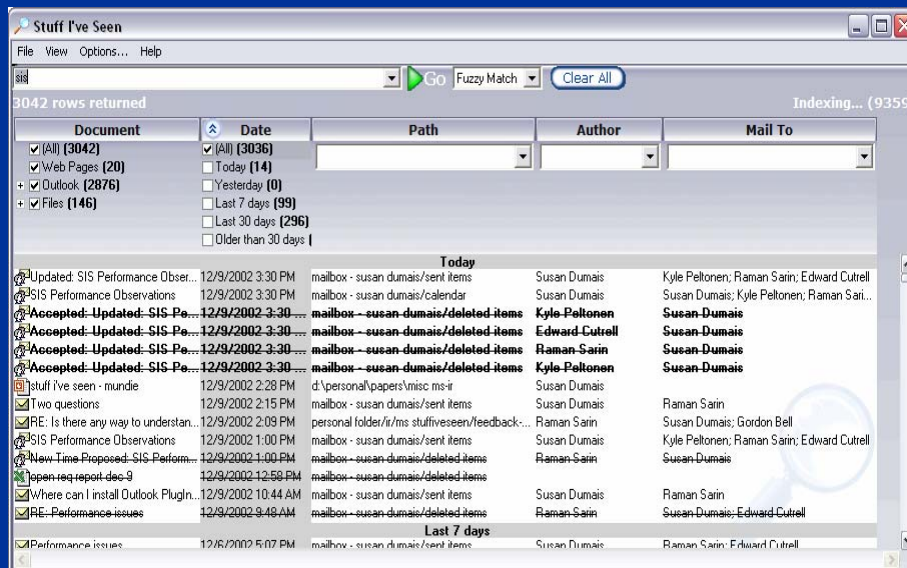
SIS, Visualizing Patterns

- 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?

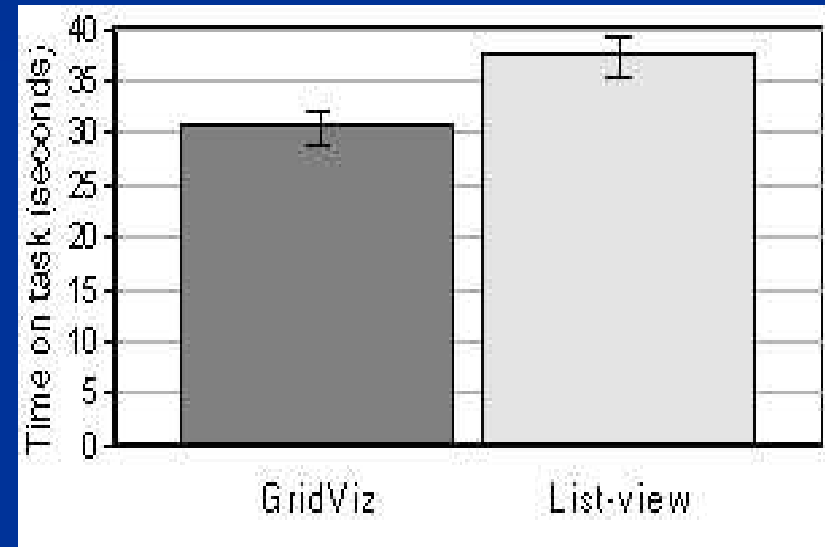
SIS, Grid vs. List Experiment



Grid View



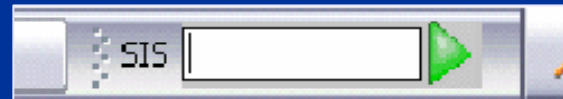
List View



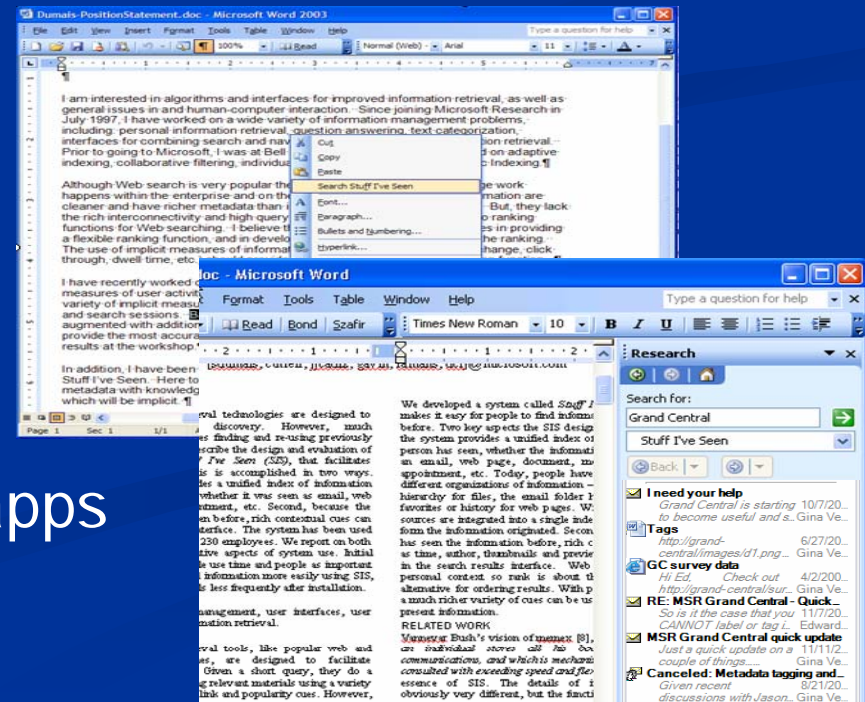
Contextualized Search

- 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

Context, Implicit Queries

The screenshot shows an email client window titled "I need your help - Message (HTML)". The message header includes: From: Gina Venolia, To: ASI and Affiliates, Cc: (empty), Subject: I need your help. The message body contains text about "Grand Central" and a link to "http://grand-". To the right, an "IQPane" window is open, showing "Search SIS for:" with results for "Gina Venolia" and "ASI and Affiliates". Below the search results, the subject "I need your help" is listed. The IQPane window is partially obscured by a yellow callout box.

Quick searches for people associated with the message and Subject.

Background search on top k interesting terms from message, based on user's index —

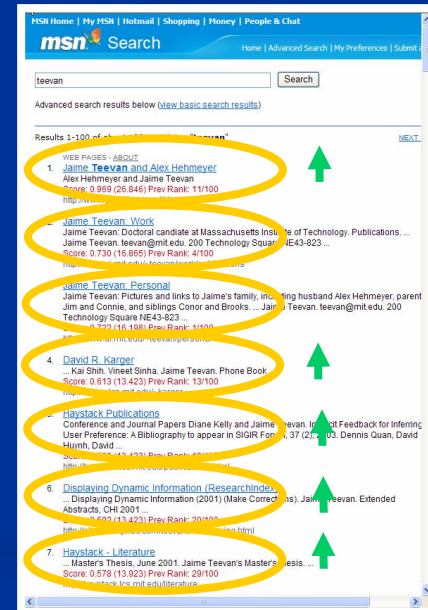
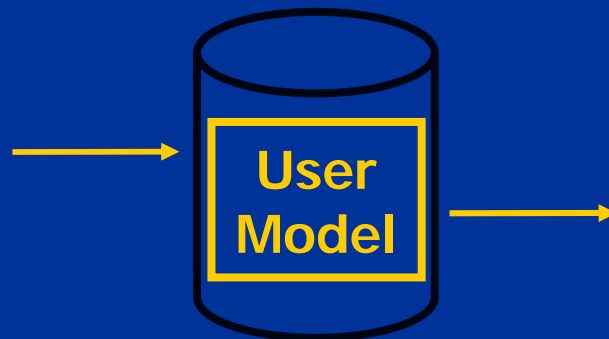
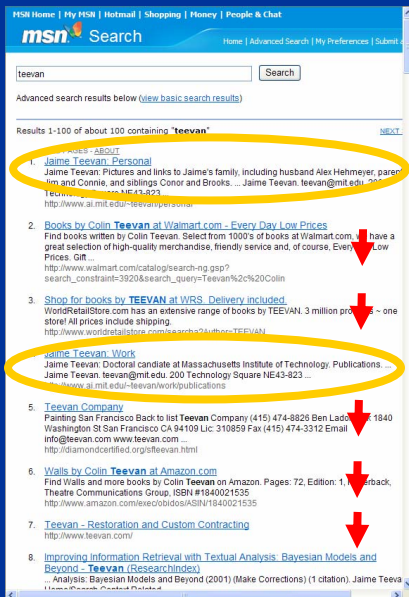
$$\text{Score} = \text{tf}_{\text{doc}} / \log(\text{tf}_{\text{corpus}} + 1)$$

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

S for 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:



Desktop Search Summary

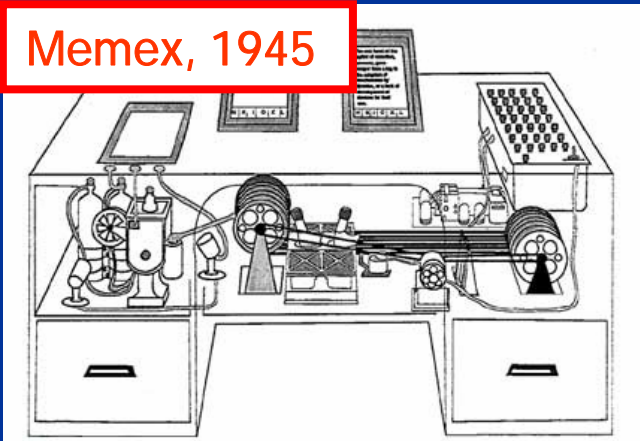
- Desktop search with SIS
 - Unified index to *stuff you've seen*
 - Heterogeneous content: files, email, web, etc.
 - Fast and flexible access
- Supports new capabilities for personal information management
 - Rich metadata vs. single hierarchy or ranked list
 - Landmarks, patterns, implicit queries, etc.
- New directions
 - Contextualized search
 - Personalized search

Vannevar Bush's Vision

V. Bush (1945). As we may think. *Atlantic Monthly*, 176, July 1945, 101-108.

- 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

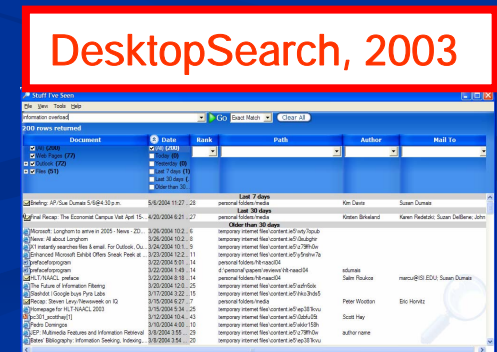


NLS/Augment, 1968



The oN-Line System display, keyboard and mouse

DesktopSearch, 2003



Thank You

- Questions/Comments ...
- More info,
<http://research.microsoft.com/~sdumais>
- MSN Toolbar and Desktop Search,
<http://toolbar.msn.com>