

Trento, Italy
22-26 June 2009

First and Seventeenth International Conference on
User Modeling, Adaptation, and Personalization

Thinking Outside the (Search) Box

Susan Dumais

Microsoft Research

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

Outline

- Search Yesterday and Today
- Search and Context
 - User, Domain and Task Contexts
- Examples of Prototypes w/ Richer Context
 - Stuff I've Seen (desktop)
 - Personalized Search (web)
 - Etc.
- Challenges and Opportunities

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
 - Auto complete
 - Inline answers
 - Richer snippets
- But, we can do better

... by using content



A stack of screenshots showing search results for 'suzan dumais' and 'weather redmond'. The top screenshot shows the search box with 'suzan dumais' entered. The middle screenshot shows the search box with 's' entered. The bottom screenshot shows the search results for 'weather redmond', featuring a 'Best match' for Microsoft Corporation with a list of links and a search box for 'Search within microsoft.com'. A red circle highlights the 'Search within microsoft.com' section.

Search and Context

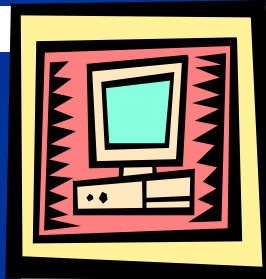
User
Context



Query Words



Query Words



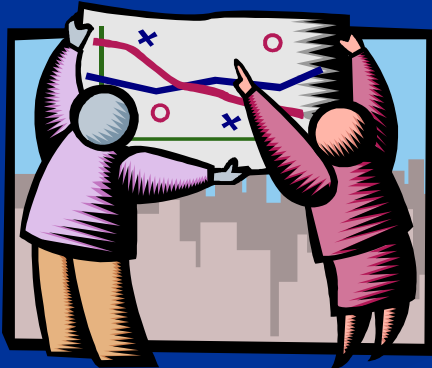
Ranked List



Document
Context



Task/Use
Context



Search and Context

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

■ User Contexts:

- Finding and Re-Finding (Stuff I've Seen)
- Novelty in news (NewsJunkie)
- Personalized search (PSearch)

■ Document/Domain Contexts:

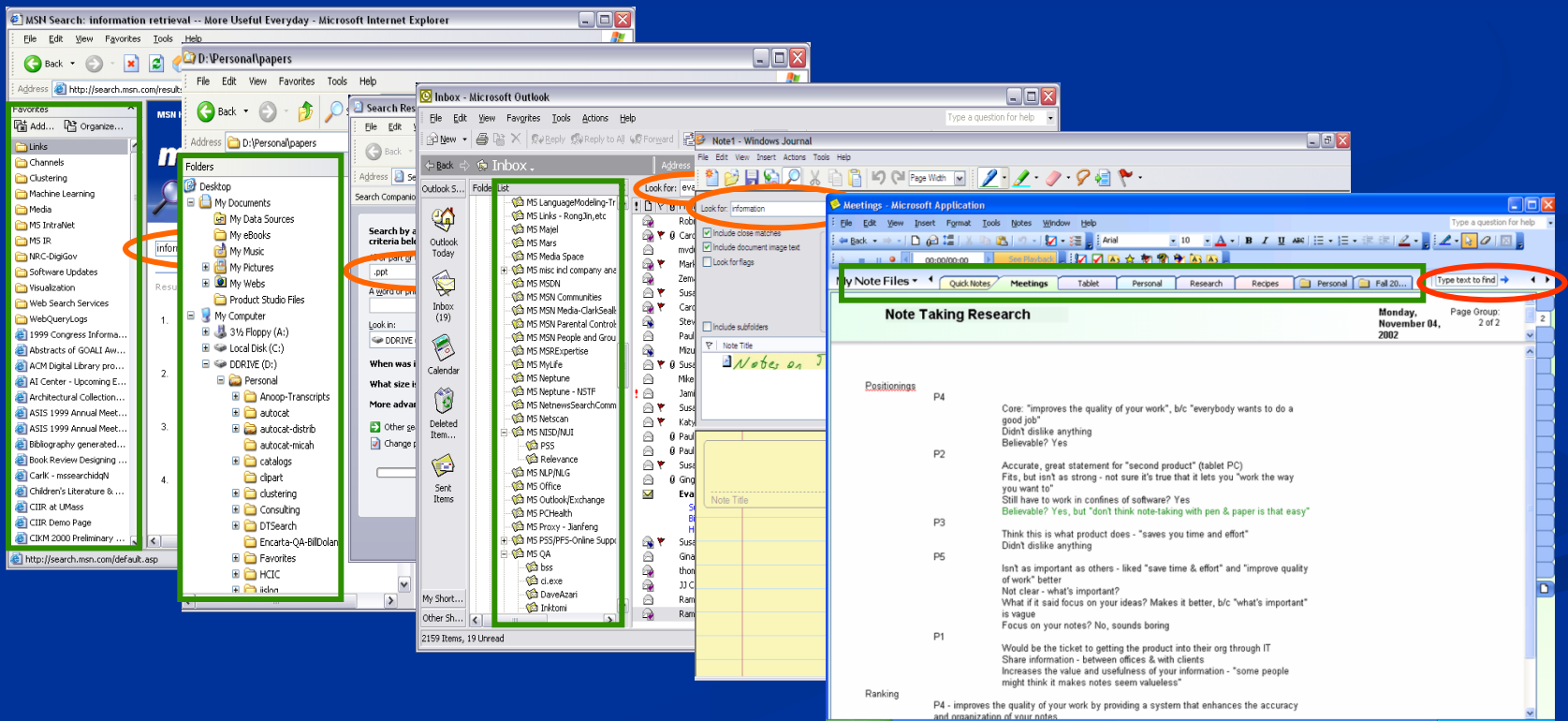
- Metadata and search (SIS, Phlat)
- Visualizing patterns in results (MemoryLandmarks, GridViz)
- Dynamic information environments (DiffIE)

■ Task/Use Contexts:

- Pages as context (Community Bar, IQ)
- Richer collections as context (NewsJunkie, PSearch)
- Understanding, sharing (uRank, SearchTogether)

Information Solos

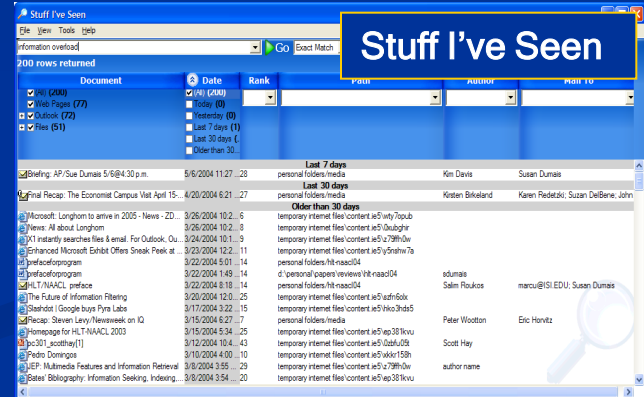
- Many sources of information, and methods for finding and organization (e.g., web, mail, contacts, docs, photos, notes)



Stuff I've Seen (SIS)

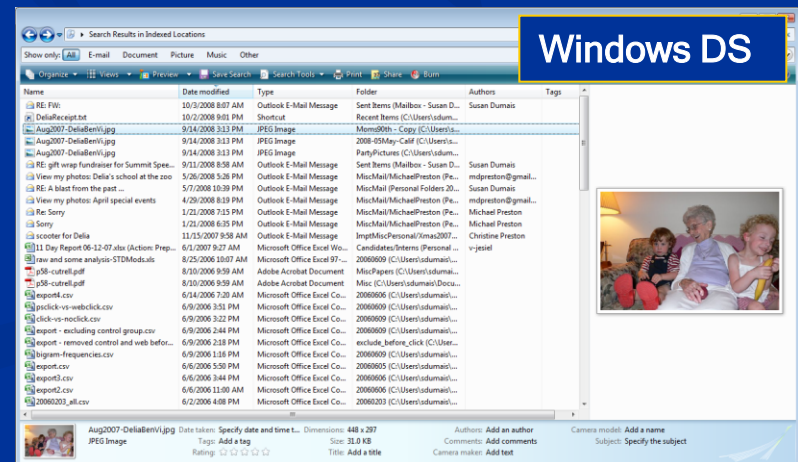
- Unified index of *stuff you've seen*

- 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
- Re-finding vs. finding



→ Vista Desktop Search (and XP, Live Toolbar)

Also, Spotlight, GDS, X1, ...



SIS Demo

Stuff I've Seen

File View Tools Help

information overload | Go Exact Match Clear All

200 rows returned

Document	Date	Rank	Path	Author	Mail To
<input checked="" type="checkbox"/> (All) (200) <input checked="" type="checkbox"/> Web Pages (77) <input checked="" type="checkbox"/> Outlook (72) <input checked="" type="checkbox"/> Files (51)					
<input checked="" type="checkbox"/> Today (0) <input type="checkbox"/> Yesterday (0) <input type="checkbox"/> Last 7 days (1) <input type="checkbox"/> Last 30 days (. <input type="checkbox"/> Older than 30...					
Last 7 days					
<input checked="" type="checkbox"/> Briefing: AP/Sue Dumais 5/6@4:30 p.m.	5/6/2004 11:27 ...28		personal folders/media	Kim Davis	Susan Dumais
Last 30 days					
<input checked="" type="checkbox"/> Final Recap: The Economist Campus Visit April 15-...	4/20/2004 6:21 ...27		personal folders/media	Kirsten Birkeland	Karen Redetzki; Suzan DelBene; John
Older than 30 days					
<input type="checkbox"/> Microsoft: Longhom to arrive in 2005 - News - ZD...	3/26/2004 10:2... 6		temporary internet files\content.ie5\wty7opub		
<input type="checkbox"/> News: All about Longhom	3/26/2004 10:2... 8		temporary internet files\content.ie5\0xubghir		
<input type="checkbox"/> X1 instantly searches files & email. For Outlook, Ou...	3/24/2004 10:1... 9		temporary internet files\content.ie5\z79ffh0w		
<input type="checkbox"/> Enhanced Microsoft Exhibit Offers Sneak Peek at ...	3/23/2004 12:2... 11		temporary internet files\content.ie5\y5nshw7a		
<input type="checkbox"/> prefaceforprogram	3/22/2004 5:01 ...14		personal folders/hlt-naacl04		
<input type="checkbox"/> prefaceforprogram	3/22/2004 1:49 ...14		d:\personal\papers\reviews\hlt-naacl04	sdumais	
<input checked="" type="checkbox"/> HLT/NAACL preface	3/22/2004 8:18 ...14		personal folders/hlt-naacl04	Salim Roukos	marcu@ISI.EDU; Susan Dumais
<input type="checkbox"/> The Future of Information Filtering	3/20/2004 12:0... 25		temporary internet files\content.ie5\szfn6ox		
<input type="checkbox"/> Slashdot Google buys Pyra Labs	3/17/2004 3:22 ...15		temporary internet files\content.ie5\hko3hds5		
<input checked="" type="checkbox"/> Recap: Steven Levy/Newsweek on IQ	3/15/2004 6:27 ...7		personal folders/media	Peter Wootton	Eric Horvitz
<input type="checkbox"/> Homepage for HLT-NAACL 2003	3/15/2004 5:34 ...25		temporary internet files\content.ie5\vp381kvu		
<input type="checkbox"/> pc301_scotthay[1]	3/12/2004 10:4... 43		temporary internet files\content.ie5\0zbfu05t	Scott Hay	
<input type="checkbox"/> Pedro Domingos	3/10/2004 4:00 ...10		temporary internet files\content.ie5\dxkr158h		
<input type="checkbox"/> JEP: Multimedia Features and Information Retrieval	3/8/2004 3:55 ... 29		temporary internet files\content.ie5\z79ffh0w	author name	
<input type="checkbox"/> Bates' Bibliography: Information Seeking, Indexing,...	3/8/2004 3:54 ... 20		temporary internet files\content.ie5\vp381kvu		

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%)
- But ... many advanced operators and query iteration in UI (48%)
 - Filters (type, date, people); modify query; re-sort results

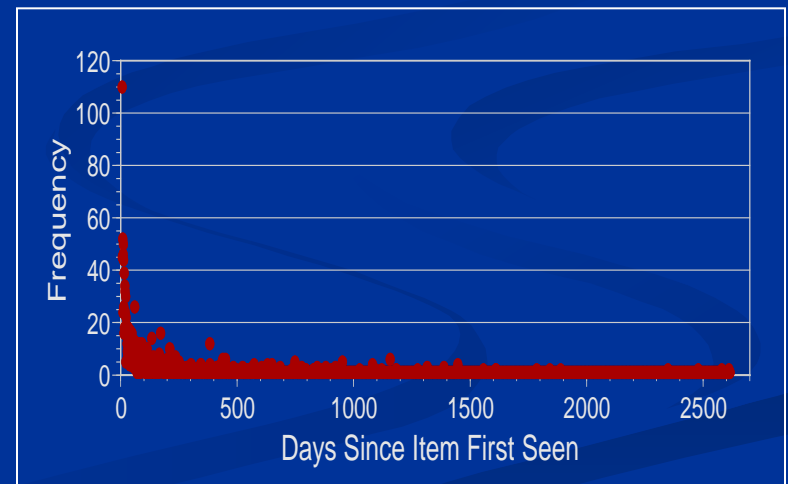
Susan's (Laptop) World		
Type	N	Size
Web	3k	0.2 Gb
Files	28k	23.0 GB
Mail	60k	2.2 Gb
Total	91k items	25.4 Gb
Index		190 Mb
		+1.5 Mb/week

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

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



Top vs. Side Views

User Interface (UI) Alternatives

Sort By Date vs. Rank

3038 rows returned

3038 rows returned

287 rows returned

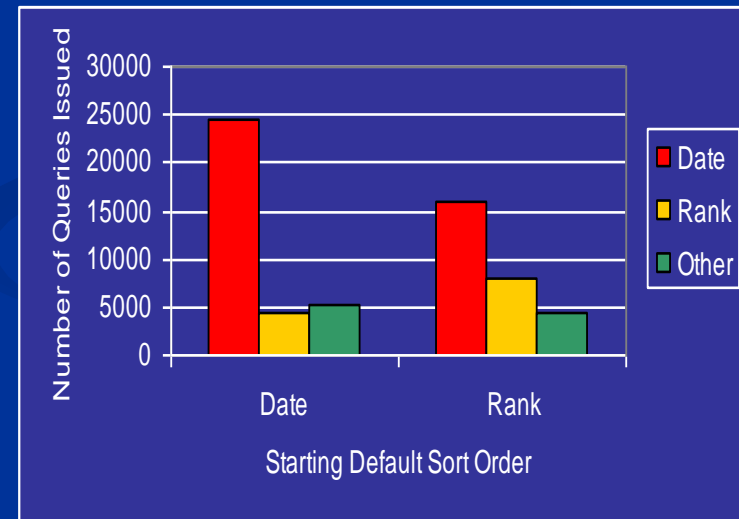
190 rows returned

Document	Date	Rank	Path	Author	Mail To
<input type="checkbox"/> (All) (190)	<input type="checkbox"/> (All) (190)				
<input checked="" type="checkbox"/> Web Pages (9)	<input type="checkbox"/> Today (0)				
<input type="checkbox"/> Outlook (85)	<input type="checkbox"/> Yesterday (0)				
<input checked="" type="checkbox"/> Files (96)	<input type="checkbox"/> Last 7 days (0)				
	<input type="checkbox"/> Last 30 days (0)				
	<input checked="" type="checkbox"/> 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//nui/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\ncr-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 best-match Rank as default
 - Importance of time
 - Few searches for “best” match; many other criteria

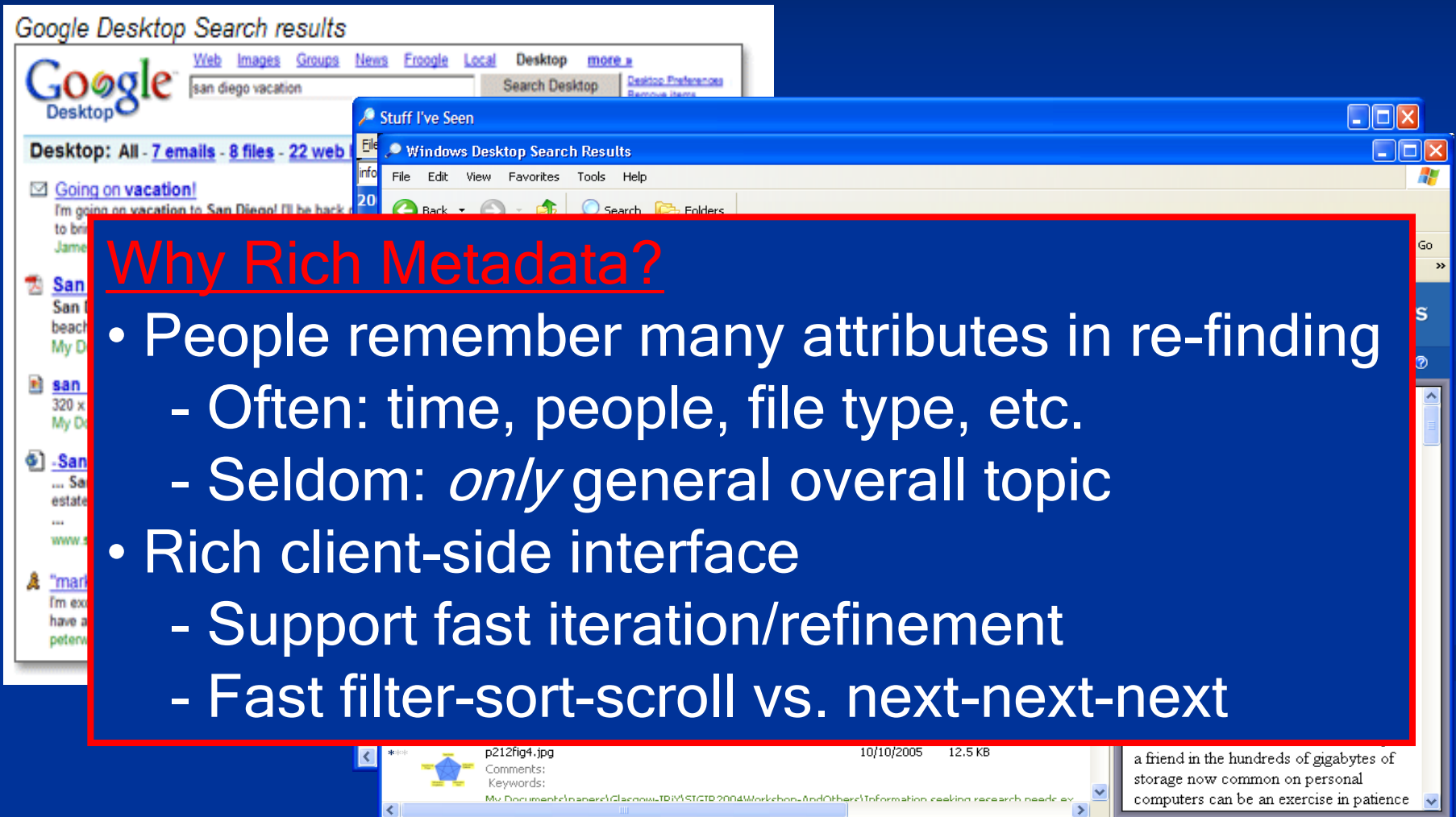


SIS Usage Data, cont'd

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, Author, Artist

Ranked list vs. Metadata (for personal content ... and beyond)



The screenshot shows a Windows desktop environment. In the background, a Google Desktop search results window is visible, displaying search results for 'san diego vacation'. In the foreground, a file explorer window titled 'Windows Desktop Search Results' is open, showing a file named 'p212fig4.jpg' with a date of '10/10/2005' and a size of '12.5 KB'. The file's metadata is displayed, including 'Comments:' and 'Keywords:'. A red box highlights the text 'Why Rich Metadata?' and a list of bullet points.

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

Metadata on the Web

The screenshot shows a Bing search results page for the query "trento, italy". The search bar is at the top center, and the results are displayed in a grid format. On the left side, there is a navigation menu with various filters. The "color" filter is highlighted with a red box, and the "photograph" filter is also highlighted with a red box. The search results show a variety of images related to Trento, Italy, including views of the city, mountains, and buildings. The page also features a navigation bar at the top with links to "Web", "Images", "Shopping", "News", "Maps", "More", "MSN", and "Windows Live".

See all results that mention **Digital Cameras** New Nikon D200 On Sale

Welcome | [Login](#) | [Register](#) My Recipe Box · Newsletters

Web Images Shopping News Maps More | MSN | Windows Live Sign in | Italy | Extras ▾

bing™
Beta

trento, italy

ALL RESULTS

IMAGES 1-15 of 35,600 results safe search moderate (change)

Images

- ⊕ SIZE
- ⊕ LAYOUT
- ⊖ COLOR
 - all
 - color**
 - black & white
- ⊖ STYLE
 - all
 - photograph**
 - illustration
- ⊕ PEOPLE

The search results grid contains 15 images showing various scenes from Trento, Italy, including city views, mountains, and buildings.

Re-finding on the Web

- 50-80% page visits are re-visits
- 30-50% of queries are re-finding queries

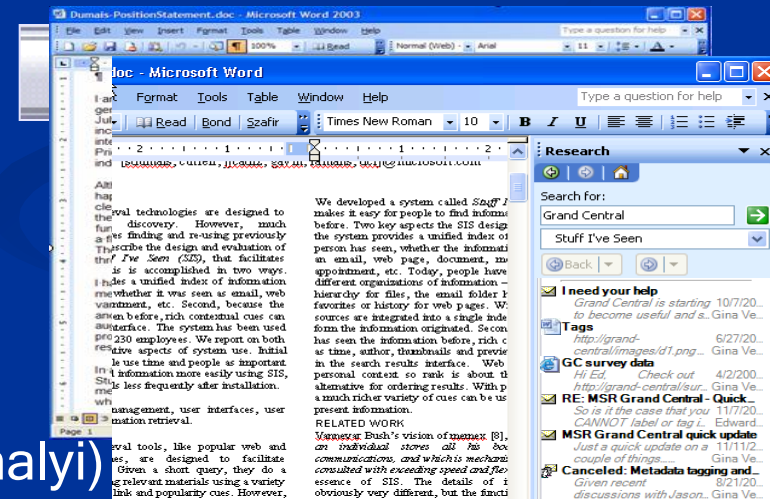
		Repeat Click	New Click
Repeat Query	33%	29%	4%
New Query	67%	10%	57%
		39%	61%

Total = 43%
Big opportunity to support re-finding on the Web

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.)

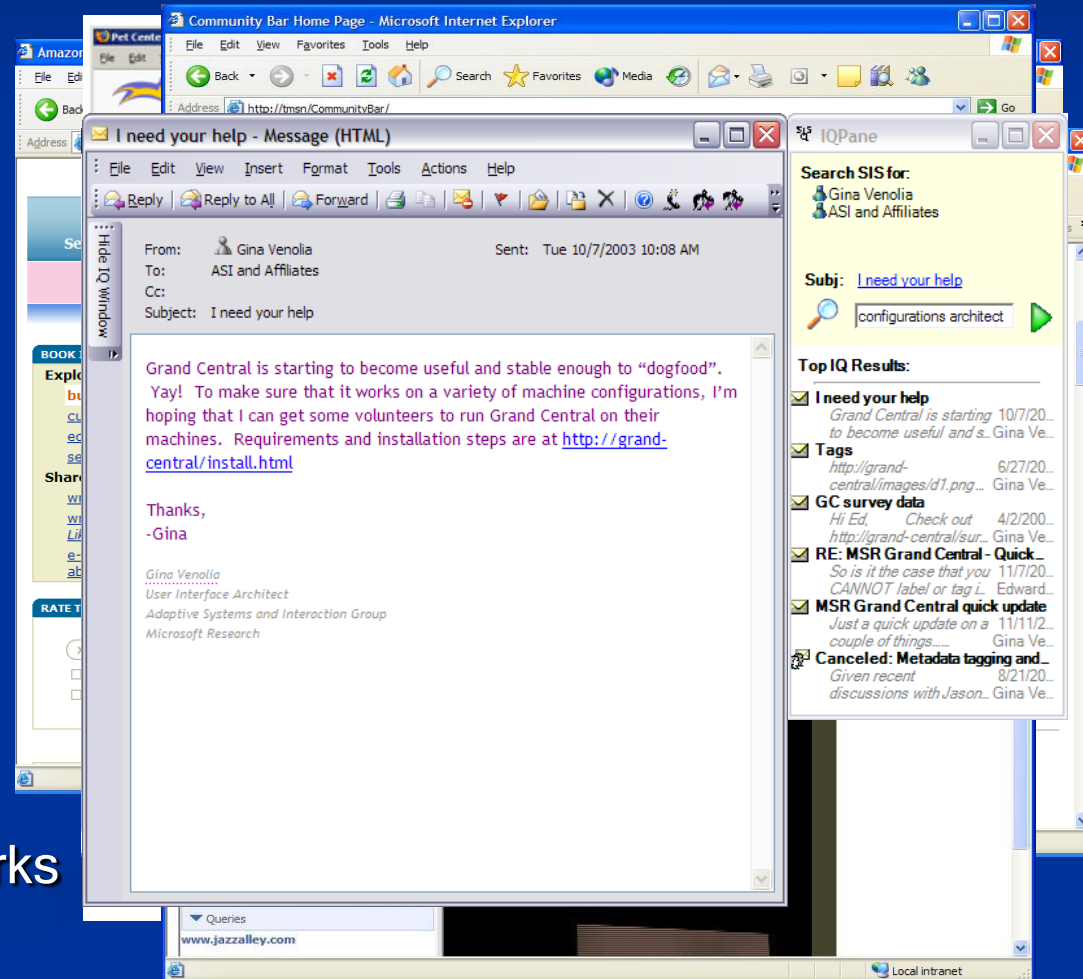
- Search always available
- Search from within apps (keywords, regions, full doc)
- Show results within app
- Maintains “flow” (Csikszentmihalyi)
- Can improve relevance



Documents as (a simple) Context

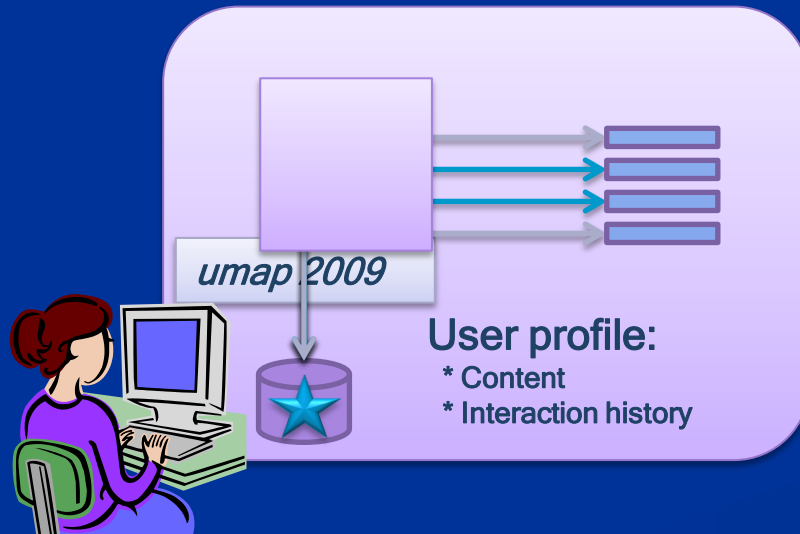
Proactive “query” specification depending on current document content and activities

- 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



*P*Search: Personalized Search (Much Richer Context)

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



PSearch Demo

■ Query: *SIGIR*

sigir - Bing - Windows Internet Explorer
http://www.bing.com/search?q=sigir&go=&form=QBRE

File Edit View Favorites Tools Help
MSR [Page loaded at 7:41 PM] Compare to Today at 7:41 PM

sigir - Bing

Web Images Videos Shopping News Maps More MSN Windows Live Sign in | United States | Extras

bing sigir

ALL RESULTS ALL RESULTS 1-10 of 161,000 results - advanced

» Personalized Search: Web Desktop

Results

Welcome | SIGIR'09
SIGIR is the major international forum for the presentation of new research results and the demonstration of new systems and techniques in the broad field of information retrieval.
sigir2009.org

Industry Track | SIGIR'09
Following up on the highly successful SIGIR 2007 Industry Event and a similar event at CIKM 2008, this year's SIGIR conference includes an Industry Track, to be held on ...
sigir2009.org/Program/industry

... show more personalized results

Special Inspector General for Iraq Reconstruction : SIGIR Homepage
Welcome to the Office of the Special Inspector General for Iraq Reconstruction (SIGIR), a temporary federal agency serving the American ...
www.sigir.mil · cached page · mark as spam

التقارير العربية	Reports
Hard Lessons	SIGIR Lessons Learned
Reports To Congress	Audits
About SIGIR	Testimony

Show more results from www.sigir.mil

ACM SIGIR Special Interest Group on Information Retrieval Home Page
"Addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization, storage, retrieval, and distribution of information."
sigir.org · cached page · mark as spam

Done Internet | Protected Mode: On 100%

PSearch ... How it Works

- Key components and challenges
 - Building a user profile
 - Personalizing the ranking
 - Personalizing the presentation
 - When to personalize?
- Beyond PSearch ...
 - Other types of profiles and applications
 - Privacy and security
 - Evaluation
 - End of serendipity?

Building a User Profile

- Type of information
 - Content: Past queries, web pages, desktop
 - Behavior: Visited pages, explicit feedback
- Time frame: Short term, long term
- Who: Individual, group
- Where the profile resides:
 - Local: Richer profile, improved privacy
 - Server: Richer communities, portability

PSearch

Personalized Ranking

■ Personal Rank = $f(\text{Cont}, \text{InterHist}, \text{Web})$

The screenshot shows a search engine interface with the search term 'sigir'. The results are ranked from top to bottom. The top result is 'Special Inspector General for Iraq Reconstruction :SIGIR Homepage', which is highlighted with a red box. Below it is 'ACM SIGIR Special Interest Group on Information Retrieval Home Page', followed by 'SIGIR'08 - Singapore', 'Welcome | SIGIR'09', and 'SIGIR 2004'. The bottom result is 'SIGIR 2004'. A red arrow points from the top result to the 'P_Content Match' factor (0.5). A yellow arrow points from the second result to the 'P_InteractHist Match' factor (1). A yellow arrow points from the third result to the 'Web Match' factor (8.5). A yellow arrow points from the fourth result to the 'P_Content Match' factor (0). A yellow arrow points from the fifth result to the 'Web Match' factor (15). A yellow arrow points from the sixth result to the 'Web Match' factor (4). A green arrow points from the bottom result to the 'Web Match' factor (15).

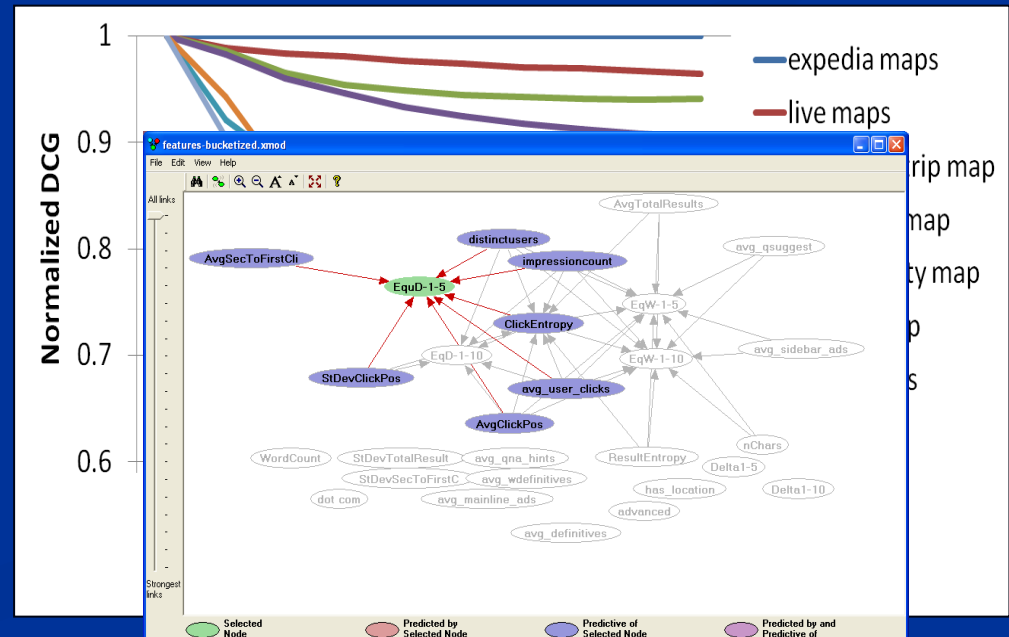
■ P_Content Match:
 $sim(result, user_content_profile)$

■ P_InteractHist 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
- Models for predicting when to personalize (using features of query, user-query)
 - Personal ranking
 - Personal relevance (explicit or implicit)
 - Group ranking
 - Decreases as you add more people
 - Gap is “*potential for personalization (p4p)*”



When to Personalize?

- High impact / low risk opportunities
 - “Re-finding” queries - depend on user-query

UMAP 2009

UMAP Trento



- * 43% coverage
- * high accuracy

- Personalization/presentation strategies that do not affect ranking
 - Annotate results that are personally relevant (allows for filtering/sorting)
 - Personalize snippets to aid relevance assessment

Beyond PSearch: Other User Profiles



- Type of information
 - Content: Past queries, web pages, desktop
 - Behavior: Visited pages, explicit feedback
- Time frame: Short term, long term
- Who: Individual, group
- Where the profile resides:
 - Local: Richer profile, improved privacy
 - Server: Richer communities, portability

PSearch

Groupization

*Session
Memory*

*Query
Suggest*

uRank

Personalization and Privacy



- PSearch
 - Local profile, local computation
 - Nothing sent to the server except the original query
- Need profile and web content in same place to rank
- When information is stored in the cloud
 - Send query and user profile, or store user profile in cloud
 - Transparency
 - Control
- Other approaches we are exploring
 - Matching an individual to a group
 - Light weight profiles (e.g., queries in a session)
 - Public or semi-public profiles (e.g., Tweets, Facebook status, blogs)

Evaluating Personalized Search Systems



- Explicit judgments (offline and *in situ*)

- Evaluate components before system
- NOTE: What's relevant

- Deploy system

- Verbatim feedback
- Measure behavioral reformulation, abandonment
- Click biases - order
- Interleaving for unbiased
- Link implicit and explicit

- Beyond a single query

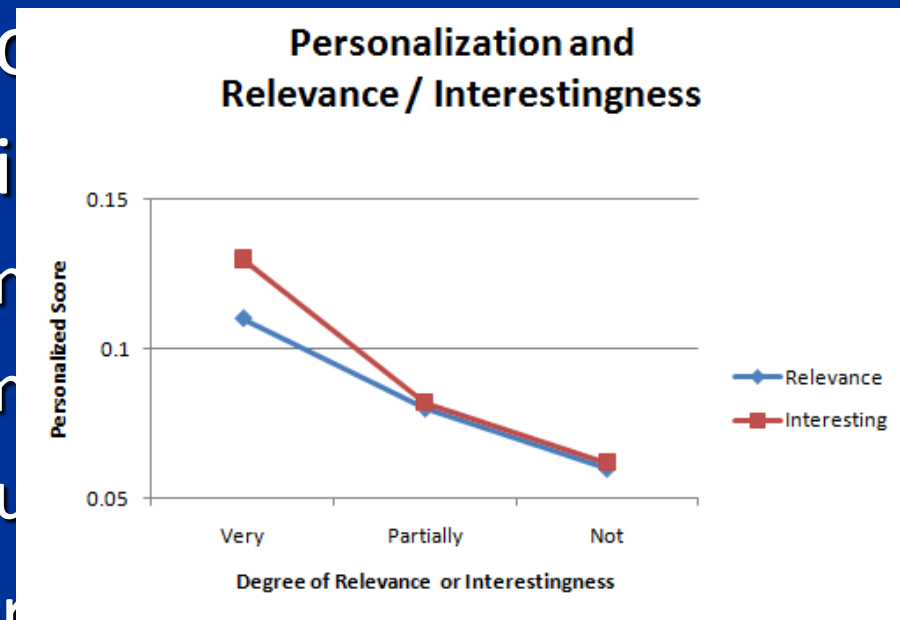
The screenshot shows a Microsoft Internet Explorer browser window displaying search results for 'www2005'. The search results list several links related to the 14th International World Wide Web Conference 2005 (WWW2005). A 'Curious Browser' pop-up window is overlaid on the search results, asking 'Did you find the information you needed at this search result?' and providing a search result: 'WWW2005 Call For Paper'. The pop-up window includes the same three emotion icons: 'Yes' (green), 'Sort of' (yellow), and 'No' (red).

Curious Browser Study (~4k)
* 45% w/ just click
* 75% w/ click + dwell + session

End of Serendipity?



- Does great search and personalization mean the end of serendipity?
- No, actually better po
- Relevance vs. interestingness
 - Personalization finds m
 - Personalization finds m
 - Many not relevant resu
- Need to be “ready” for serendipity



Beyond Search - Sharing & Collaborating

- **uRank** - Edit, organize and share search results w/ your friends

uRank

- Edit: move results up/down, annotate w/ notes
 - Results as first class objects
- Lists: create lists while you're researching
- Sharing: share results, lists w/ your friends

The screenshot displays the uRank web application interface. At the top, there's a search bar with the text 'favorite restaurants' and a search icon. Below the search bar, the page shows search results for 'Pair, Small Plates, Food and Wine in Seattle'. Each result includes a title, a brief description, a URL, and a list of user comments with dates. For example, one result is 'Pair Food and Wine Seattle ... Welcome to Pair, a neighborhood restaurant featuring european-inspired dishes and a beautifully'. Comments include 'Johanna Jacobsen Kiciman said "European-feel restaurant with delicious food." Thu, Oct 09' and 'You said "A nice neighborhood restaurant with great food" Mon, Sep 15'. To the right of each result, there are icons for 'move me', 'copy me', 'add note', and 'delete me'. A red box highlights the 'uRank' logo in the top right corner of the page. The page also features a sidebar on the right with a search bar and a list of related items like 'Emerald City Presse Re', 'Le Pichet a Capitol Hill', 'Junglecty', and 'Central Dis'.

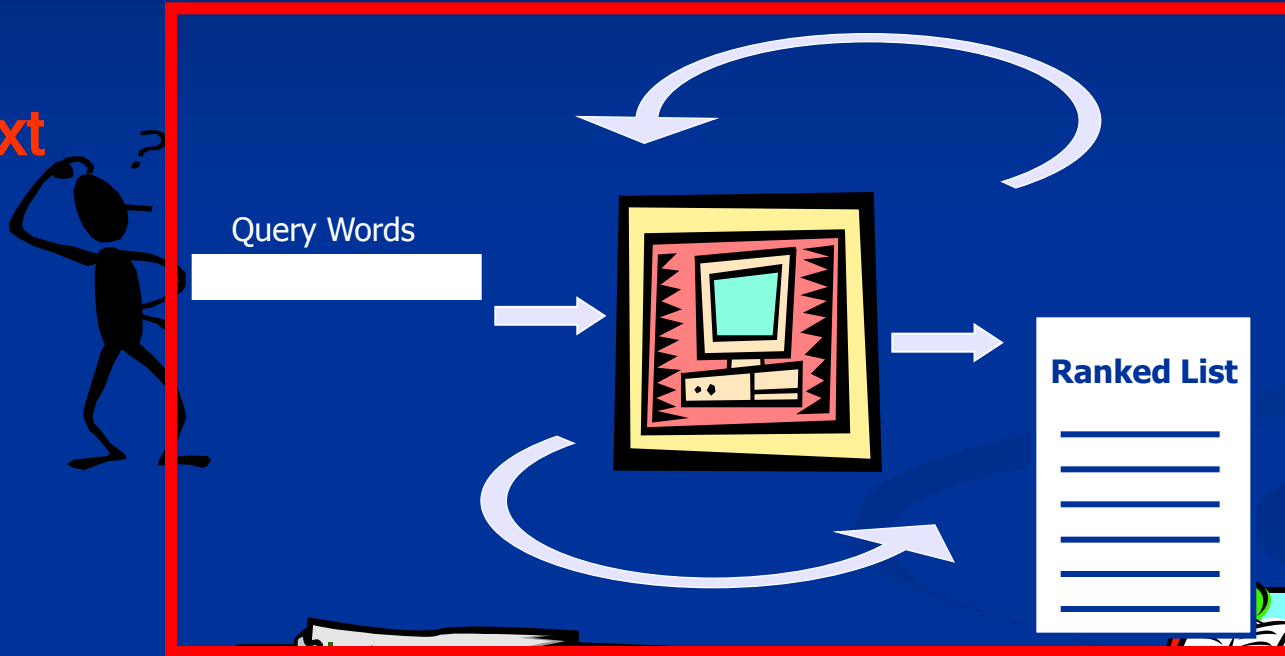
Try It: <http://research.microsoft.com/projects/urank>

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
 - 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 analysis, use and sharing of results
 - Considering search as part of richer landscape

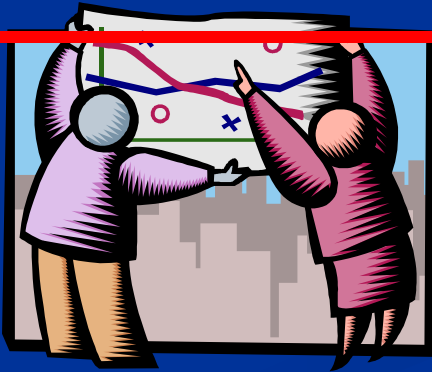
Thinking Outside the (Search) Box

User
Context



Document
Context

Task/Use
Context



Thank You !

- Questions/Comments ...
- More info,
<http://research.microsoft.com/~sdumais>

■ Stuff I've Seen & Memory Landmarks

- S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin & D. C. Robbins (2003). Stuff I've Seen: A system for personal information retrieval and re-use. *SIGIR 2003*.
- M. Ringel, E. Cutrell, S. Dumais & E. Horvitz (2003). Milestones in time: The value of landmarks in retrieving information from personal stores. *Interact 2003*.
- E. Horvitz, S. Dumais & P. Koch (2004). Learning predictive models of human memory landmarks. *Cognitive Science 2004*.

■ Personalized Search, Groupization & Serendipity

- J. Teevan, S. Dumais & E. Horvitz (2005). Personalizing search via automated analysis of interests and activities. *SIGIR 2005*.
- F. Radlinksy & S. Dumais (2006). Improving personalized web search using results diversification. *SIGIR 2006*.
- J. Teevan, S. Dumais & E. Horvitz (2006). To personalize or not: Modeling queries with variation in user intent. *SIGIR 2008*.
- J. Teevan, S. Dumais & E. Horvitz (in press). Potential for personalization. *TOCHI, in press*.
- P. Andre, J. Teevan & S. Dumais (2009). From x-rays to silly putty via Uranus: Serendipity and its role in Web search. *CHI 2009*.

- J. Teevan, M. Ringel & S. Bush(2009). Discovering and using groups to improve personalized search. *WSDM 2009.*
- R. White, S. Dumais & J. Teevan (2009). Characterizing the influence of domain expertise on Web search behavior. *WSDM 2009.*
- Re-Finding & Metadata on the Web
 - J. Teevan, E. Adar, R. Jones & M. Potts (2007). Information retrieval re-retrieval: Repeat queries in Yahoo's logs. *SIGIR 2007.*
 - J. Teevan, S. Dumais & Z. Gutt (2008). Challenges in supporting faceted search in large, heterogeneous corpora like the Web. *HCIR.*
- Implicit Queries & NewsJunkie
 - S. Dumais., E. Cutrell, R. Sarin & E. Horvitz (2004). Implicit queries (IQ) for contextualized search. *SIGIR 2004.*
 - . E. Gabrilovich, S. Dumais & E. Horvitz (2004). NewsJunkie: Providing personalized newsfeed analysis via information novelty. *WWW 2004.*
- Curious Browser
 - S. Fox, K. Jamawat, M. Mydland, S. Dumais & T. White (2005). Evaluating implicit measures to improve the research experience. *ACM:TOIS, 23(2), 2005.*
- Search Together and uRank
 - M. Morris & E. Horvitz (2007). SearchTogether: An interface for collaborative Web search. *UIST 2007.*