NSF WORKSHOP ON TASK-BASED SEARCH

TASK-BASED SEARCH: A SEARCH ENGINE PERSPECTIVE

**#nsftasks**

Susan Dumais, Microsoft Research

Outline

- Why tasks are important

- A taxonomy of Web browsing/searching tasks

- Detecting tasks
  - Within and across sessions
  - Implicit detection using behavioral interaction

- Supporting tasks
  - Ranking/recommending
  - Interaction

- Evaluating task support

Search and Context

User Context

Query Words

Query Words

Ranked List

Document Context

Task Context
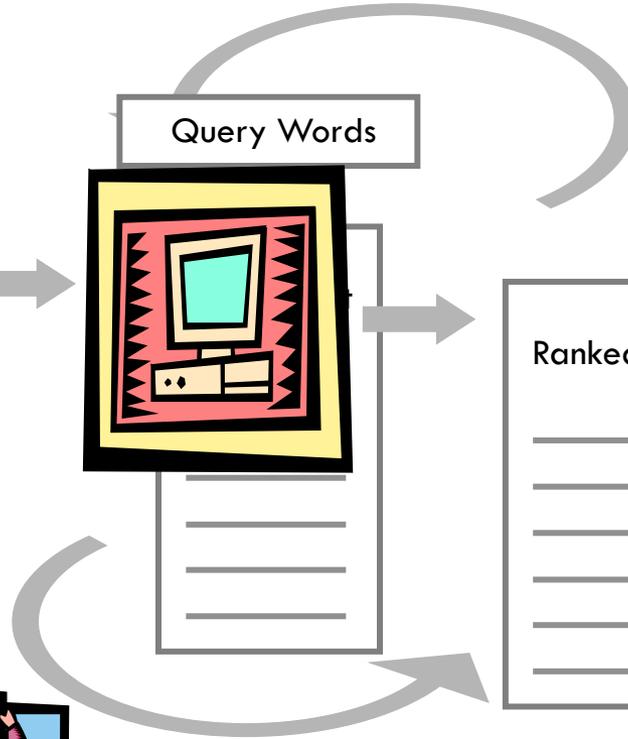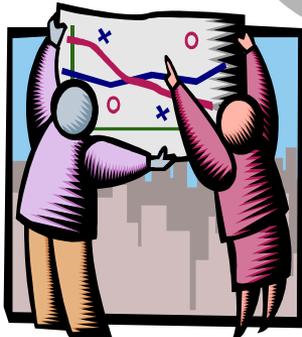
Why "Tasks" Are Important

- Long sessions are common
  - 40% of sessions contain multiple queries
  - 40% of sessions take 3+ minutes
    - 5% of these take 30+ minutes
- Long sessions are more likely to continue
- Long sessions account for most of the search time
  - ~50% of time spent in sessions of 30+ mins
- What are people doing in these long sessions?
- How can we support them?
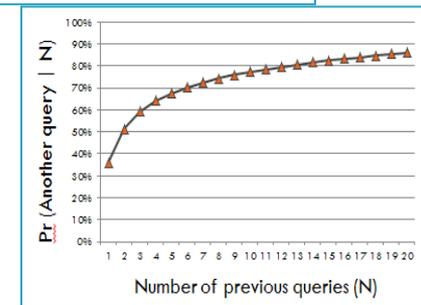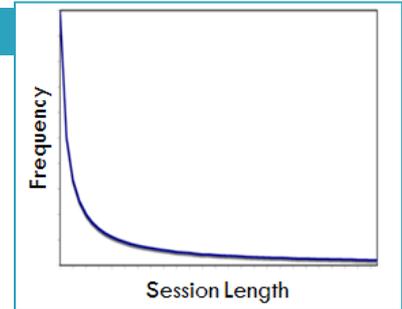
Tasks != Sessions

- Are interleaved with other tasks
- Extend across sessions and across devices
- Can extend over long periods of time
  - E.g., Queries related to "mortgage" over time

Time Period				
0–30 min	1–7 days	7–30d	30–90d	90–365d
mortgage	realtors	llc	kohls	patio
mortage	owner	associates	bath	harbor
mortgage	homes	insurance	overstock	outdoor
calculator	mls	lowes	barn	replacement
mortgages	remax	notary	sears	pools
lenders	property	depot	linens	hampton
calculators	financial	savings	beyond	lawn
countrywide	appraisers	construction	kmart	enterprise
gmac	builders	condo	pottery	ymca
refinance	prudential	business	walmart	vehicle
rates	zillow	secretary	outlet	supply
interest	bankruptcy	furniture	costco	resorts
broker	real	allstate	target	lake
lending	keller	companies	pier	rv
lender	properties	contractors	bed	walgreens
payment	agreement	cost	grill	newport
loan	appraisals	reverse	kitchen	lumber

Which Tasks To Focus On

- Several factors considered when deciding which tasks to support
  - Coverage
    - What proportion of tasks are impacted?
  - Accuracy
    - How accurately can we detect a task?
  - Impact
    - What can we do about it?
   - How does it changes behavior or outcome?

A Taxonomy of Browsing Behavior

- Developed a new task taxonomy for Web browsing and query-focused browsing behavior
- Log data from: 187 users, 453 sessions, 1913 tasks
- Used iterative taxonomy development, 5 judges
- **Verb**-based, not domain- or search-activity based
  - Action-topic pairs
    - E.g., find-address, plan-trip, download-song, compare-product

- NOTE: many other methods for identifying tasks of interest including ethnography, surveys, critical incidents

Compare to Broder & Russell et al.

| [Broder 2002]	[Google 2009]	Action	Topic
Informational	*Find-Simple*	Find	a specific miscellaneous fact
		Find	an address or contact information
	Find-Complex	Find	a date online
		Find	a job
		Find	facts about a person
		Find	real estate
	Explore/Learn	Browse	a social network
		Compare	products or services for use
		Discover	more information about a specific topic
		Discover	leisure activities
		Learn	how to perform a task
		Plan	event
		Plan	travel
Transactional	*Locate/Acquire*	Buy	a product
		Download	files
		Plan	event
		Plan	travel
		Sell	a product
	Play	Look	at online pictures for entertainment
		Watch	online videos for entertainment
		Listen	to online music
		Play	online games
	Meta		
Navigational	*Navigate*		
Other	*Other*	Communicate	with people online
		Do	education related tasks (online homework, etc.)
		Manage	personal accounts
		Read or write	on blogs or forums

Top Web Tasks by Session

- Web Tasks: All web activities including browsing behavior and search behavior
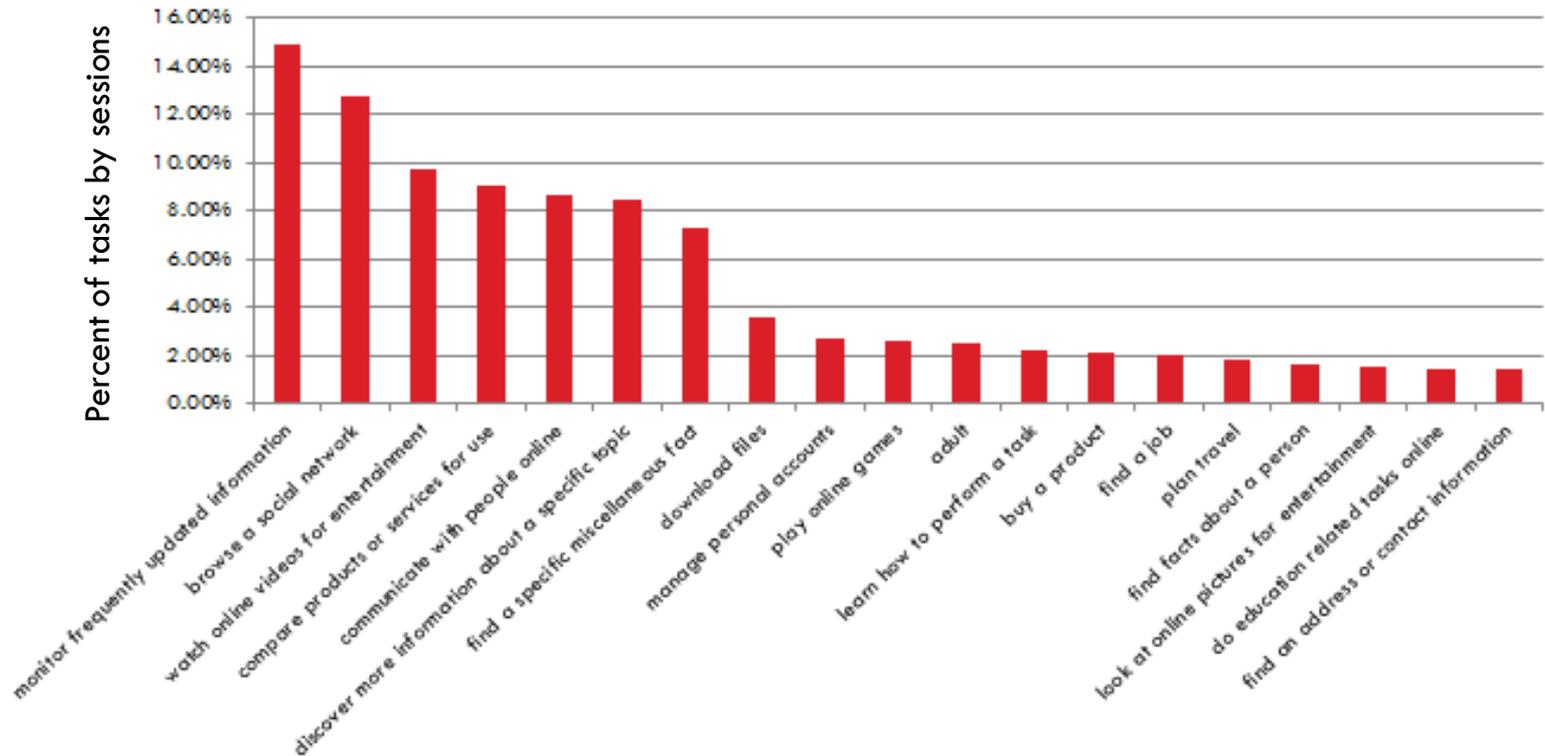
Top Query Tasks by Session

☐ Query Tasks: Contiguously labeled tasks within a session which contain a query issued to Google, Yahoo! or Bing
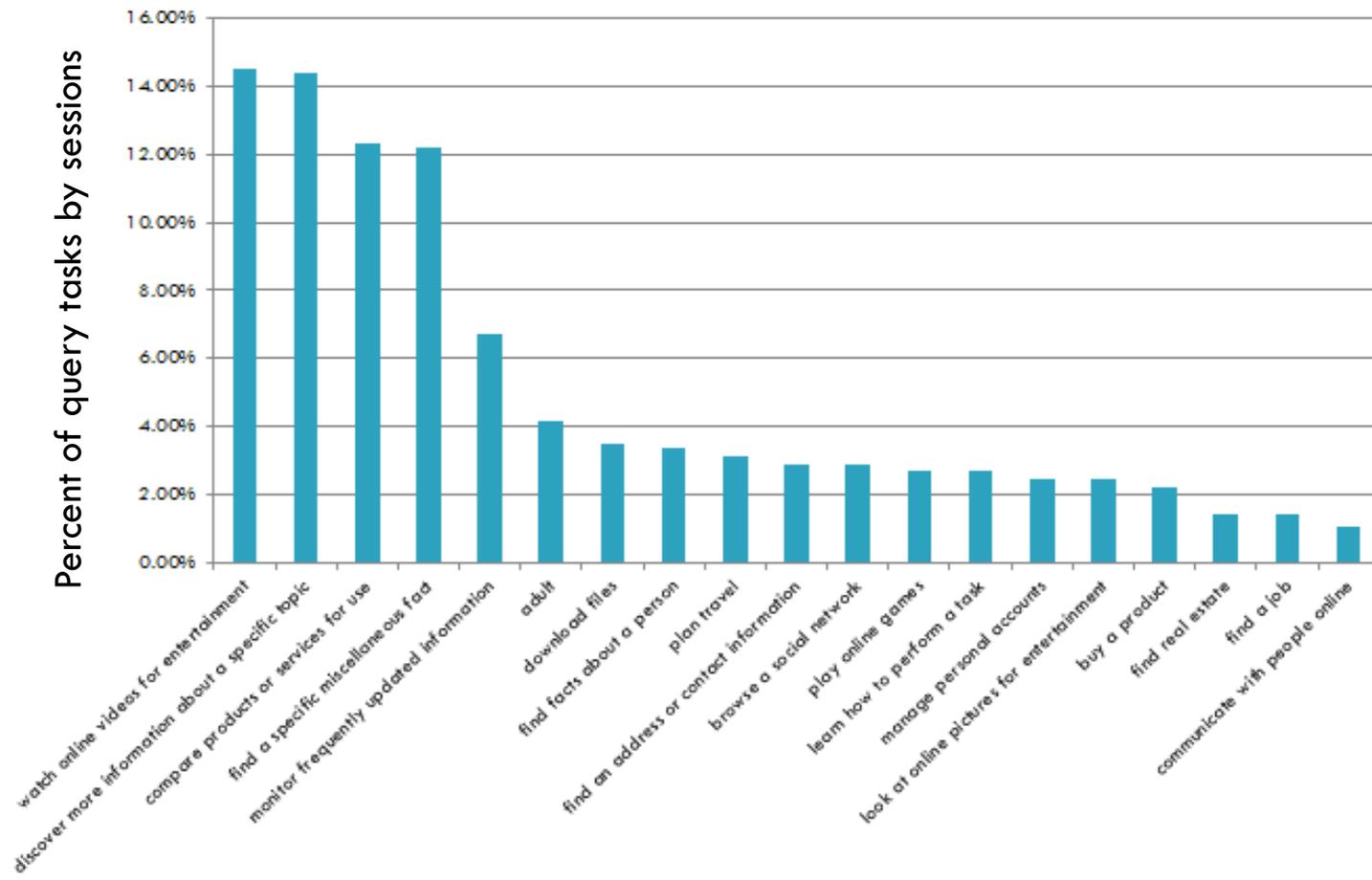
Other Task Properties

Task	Queries per task	Avg. events per task	Avg. length (mins)
adult	18.7	18.3	48.6
look at online pictures for entertainment	16.2	50.7	15.7
learn how to perform a task	**13.0**	**11.9**	**8.5**
download files	11.7	31.2	15.0
watch online videos for entertainment	7.5	19.5	19.0
find facts about a person	**6.9**	**18.9**	**4.8**
discover more information about a specific topic	**6.8**	**24.8**	**13.5**
compare products or services for use	**6.8**	**22.3**	**24.8**
find real estate	5.1	11.7	14.9
plan travel	4.7	5.1	12.0
find an address or contact information	4.2	48.5	7.5
not a task	3.9	10.3	21.6
monitor frequently updated information	3.6	24.1	20.6
find a specific miscellaneous fact	3.2	40.0	7.9
buy a product	3.1	15.4	8.9
play online games	2.0	21.1	16.2
manage personal accounts	1.8	42.0	7.5
find a job	1.8	29.4	18.0
communicate with people online	1.8	9.0	5.6
browse a social network	1.5	7.1	24.7

E.g., Compare Task

- Average compare task has lots of engagement
  - Contains 7 queries, 22 actions
  - Lasts for 25 minutes
- Compare sub-tasks
  - Explore dimensions for comparison (e.g., size, color, capacity)
  - **Compile and refine list of choices**
  - **Find details**
  - Read reviews
  - Read side-by-side
  - Act on a comparison decision

Effort Distribution of Comparison Sub-task

% of event

70%
60%
50%
40%
30%
20%
10%
0%

Compile and…   Find details about…   Explore…   Read reviews…   Read side by side…   Act on a…
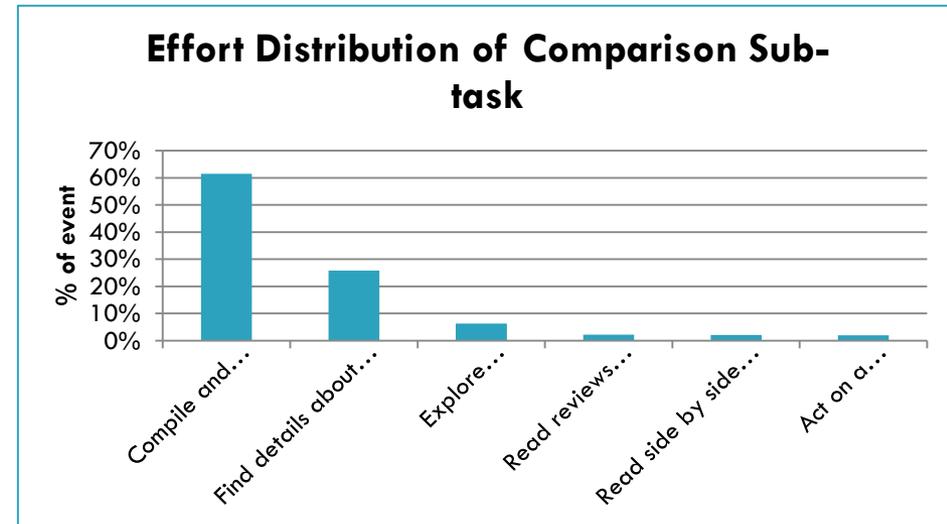
Automatically Detecting Tasks

- Approach
  - Encode search interaction features
  - Label some tasks
  - Learn a model that links implicit behaviors to explicit task labels
  - Run learned model in open-loop
- Used to model
  - Tasks within sessions
  - Tasks that extend across session and/or device

Cross-Session Tasks

☐ Many tasks continue across sessions

 ▪ E.g., Medical diagnosis and treatment, event planning, how-to advice, shopping research, academic research, etc.

Time	Query	NClicks
1/22/2011 1:11pm	ken more	
1/22/2011 1:13pm	kenmore elite parts	
1/22/2011 1:14pm	kenmore elite washer	
1/22/2011 1:15pm	kenmore elite washer troubleshooting	3
1/22/2011 3:11pm	hotmail	
1/22/2011 4:15pm	weather nyc	
1/22/2011 4:19pm	hotmail	
1/22/2011 4:46pm	ebay	10
1/22/2011 6:40pm	how to fix kenmore elite washer	2
1/22/2011 6:44pm	kenmore elite error codes	
1/23/2011 6:20pm	kenmore elite error codes	
1/23/2011 6:22pm	kenmore elite error codes ce	1
1/24/2011 6:30pm	rangers scores	
1/24/2011 6:40pm	kenmore elite washer manual	
1/24/2011 6:55pm	kenmore elite calypso washer manual	1
1/24/2011 7:01pm	elite calypso wash motion	3
1/24/2011 8:31pm	hotmail	
1/25/2011 6:29pm	sears washer/dryer reviews	3
1/25/2011 7:31pm	kenmore elite parts	
1/25/2011 7:35pm	kenmore elite parts sears	1
1/25/2011 7:37pm	sears nyc hours	1
1/25/2011 8:00pm	nytimes.com	5
1/25/2011 8:35pm	regal cinema schedule	1

fixing a kenmore washing machine
- 16 queries
- 25 clicks
- several hours
- over 3 days

Cross-Session Tasks

☐ Many tasks continue across sessions

　　◻ E.g., Medical diagnosis and treatment, event planning, how-to advice, shopping research, academic research, etc.

☐ Can we *identify same task*, and *predict task resumption?*

☐ Data and labeling

Time	Query	NClicks
1/22/2011 1:11pm	ken more	
1/22/2011 1:13pm	kenmore elite parts	
1/22/2011 1:14pm	kenmore elite washer	
1/22/2011 1:15pm	kenmore elite washer troubleshooting	3
1/22/2011 3:11pm	hotmail	
1/22/2011 4:15pm	weather nyc	
1/22/2011 4:19pm	hotmail	
1/22/2011 4:46pm	ebay	10
1/22/2011 6:40pm	how to fix kenmore elite washer	2
1/22/2011 6:44pm	kenmore elite error codes	
1/23/2011 6:20pm	kenmore elite error codes	
1/23/2011 6:22pm	kenmore elite error codes ce	1
1/24/2011 6:30pm	rangers scores	
1/24/2011 6:40pm	kenmore elite washer manual	
1/24/2011 6:55pm	kenmore elite calypso washer manual	1
1/24/2011 7:01pm	elite calypso wash motion	3
1/24/2011 8:31pm	hotmail	
1/25/2011 6:29pm	sears washer/dryer reviews	3
1/25/2011 7:31pm	kenmore elite parts	
1/25/2011 7:35pm	kenmore elite parts sears	1
1/25/2011 7:37pm	sears nyc hours	1
1/25/2011 8:00pm	nytimes.com	5
1/25/2011 8:35pm	regal cinema schedule	1

　　◻ 270k people w/ 5+ search sessions and 10+ queries

　　◻ Identify an early-dominant task

　　　　■ Task occurs during first 2 days

　　　　■ 2 or more <u>unique</u> queries on same task

　　◻ Automatically cluster 10k queries into tasks

　　◻ Manually label 1.2k of these tasks

Cross-Session Tasks (cont'd)

- 15% of early-dominant tasks continue across sessions
    - 25% of queries are from multi-session tasks
- Two prediction tasks
    - Identify Same Task: *Given current query, find (previous) queries on this task*
    - Predict Task Resumption: *Will searcher resume the task (within the week)*

- Behavioral interaction features
    - Query, Session, History, Pair-wise
- Labels (automatic, human)
- Learned model (logistic regression, MART)

- Prediction accuracy
    - Same Task: F1= 0.67
    - Task Continuation: F1 = 0.75

Feature	Weight	Feature Type
SameQueryHist	1.11	History-based
NumSessHist	0.60	History-based
NumDomQueriesHist	0.39	History-based (Table 6)
AvgInterQTimeHist	0.24	History-based (Table 6)
FreqDomQueriesHist	0.24	History-based (Table 6)
NumDwell30Hist	0.22	History-based (Table 6)
NumQueryHist	0.21	History-based
NumTop10Clicks	-0.16	Query-based
AvgInterQTimeSess	-0.17	Session-based (Table 6)
NumClicksHist	-0.18	History-based
NumQueryChars	-0.21	Query-based
SubQueryHist	-0.23	History-based
SupQuerySess	-0.40	Session-based
SupQueryHist	-0.40	History-based
SubQuerySess	-0.49	Session-based

Search Missions and Goals

- Hierarchical segmentation of search logs into missions and goals
- Data and labeling
  - 312 user sessions, 3 days
  - 1820 missions, 2922 goals, 8266 queries
  - Label all goals and missions within session
- Supervised learning
  - Task boundary detection
 - Same task identification
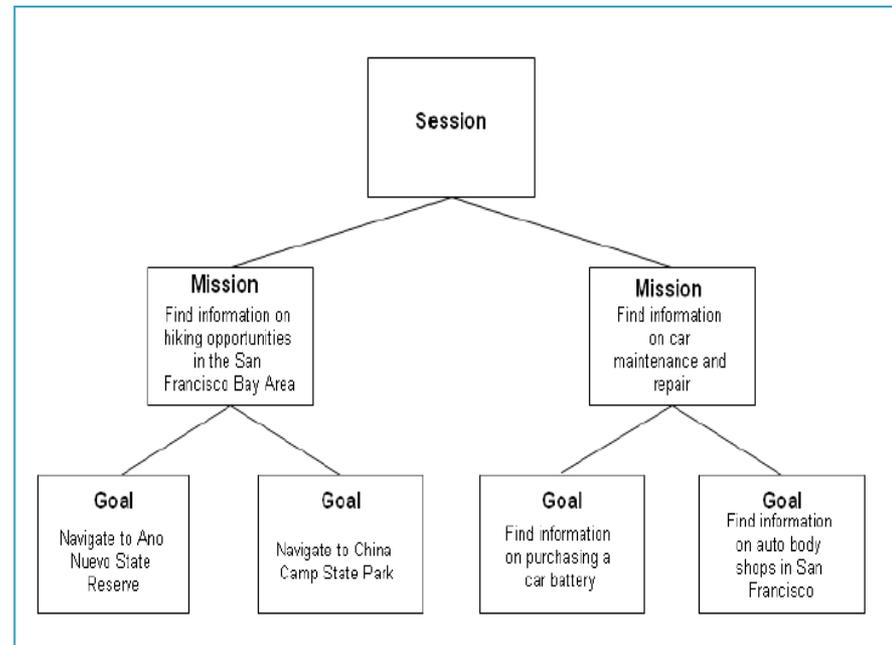
Research Missions and Search Pad

- Search missions and goals (from Jones & Klinkner)
  - 10% of search sessions contain "research" missions
- Can research missions be identified on-the-fly during a search session?
  - Input features representing tasks and engagement
    - Textual – similarity q1, q1
    - Session-based – queries, clicks, queries since last click, etc.
    - Time-based – time between q1, q2, total session time, etc.
  - Three general signals compute probability of research mission
    - Research_mission (q1, q2) boosted dt classifier
    - Same_mission (q1, q2) boosted dt classifier
    - Similarity (topics (q1), topics (q2))
- If research mission detected, show Yahoo! Search Pad

Research Missions and Search Pad

- If research mission detected, show Yahoo! Search Pad
- Deployed broadly, for a while

Scenario:

Searcher is planning a trip to Barcelona

Issues a few related queries:

"*Barcelona cheap flights*" (several clicks)

"*Barcelona airport*" (several clicks)

Then, when issuing the query "*Barcelona airport transfer*", she might see …

Research Mission detected
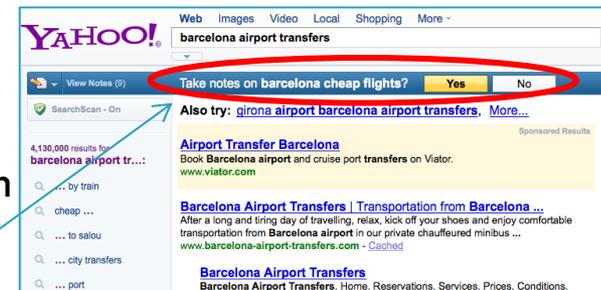
Search Pad invoked
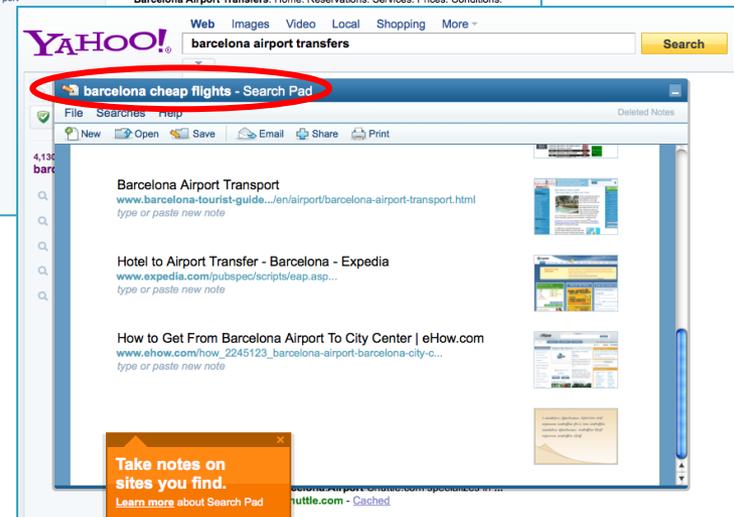
How to Support Tasks

□ (Better Ranking, Query Suggestion, etc.)

□ History of queries and/or URLs

□ Richer snippets

□ Verticals

How to Support Tasks

- (Better Query Suggestion, Ranking, etc.)
- History of queries and/or URLs
- Richer snippets
- Verticals
- Inline "answers"

How to Support Tasks

- (Better Query
- History of que
- Richer snippet
- Verticals
- Inline "answers
- **Custom experiences e.g., entities, exploration**

How to Support Tasks

□ (Better Query Sugg

□ History of queries

□ Richer snippets

□ Verticals

□ Inline "answers"

□ Custom experience

□ Richer Sensemaking
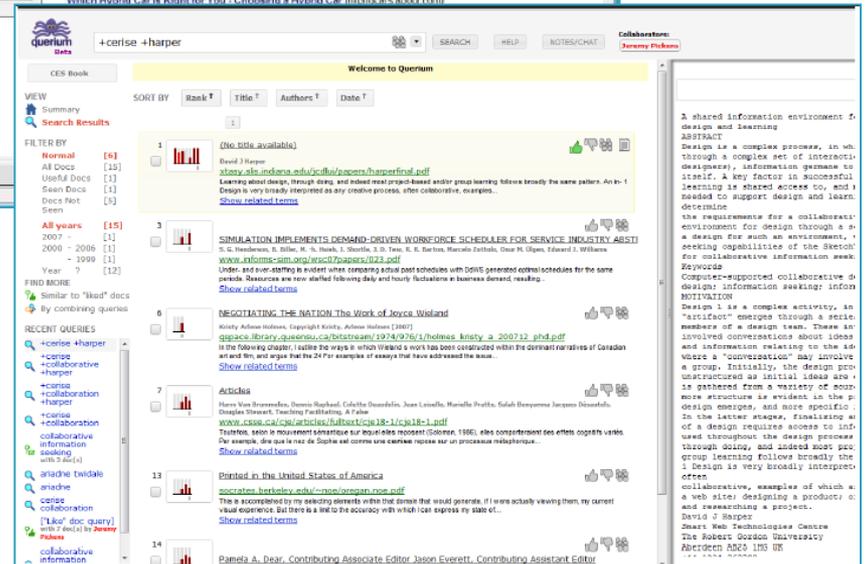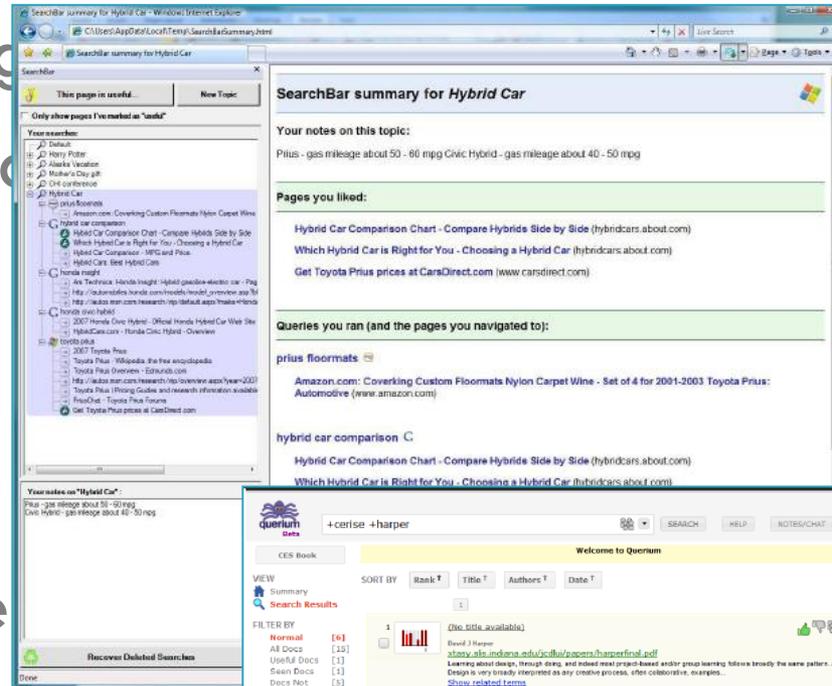
How to Support Tasks

- (Better Query Suggestion, Ranking, etc.)
- History of queries and/or URLs
- Richer snippets
- Verticals
- Inline "answers"
- Custom experiences – e.g., entit
- Richer Sensemaking
- **Apps, apps and more apps**

How to Evaluate Task Support

- How to measure success
  - Explicit
    - Retrospective (By expert judges, crowd workers or individual)
    - I*n situu* (By individual e.g., Curious Browser)
  - Implicit
    - Off-line using learned models
    - On-line operational system (using controlled experiments)
- What to measure
  - Many metrics: "scorecards"
    - E.g., clicks, dwell, engagement, time on task, sustained use, etc.
  - System as whole (and system components)
    - *"there is a quantum of barriers in task integration and work task contexts"*

Summary

- Tasks are important
  - Significant time is devote to some tasks which require multiple queries, sessions and devices to complete
- Some progress in modeling and support
  - Automatic detection and prediction of tasks, using behavioral modeling, implicit feedback and machine learning
  - Support for tasks increasing within Web search engines and in stand-alone apps
- Many challenges/opportunities remain
  - Task selection: where are biggest opportunities
  - Support: broadly available vs. specific task
    - Are there reusable components that generalize across tasks?

□Thanks !

□Questions / Comments ???

(Some) References

- Bailey, *Under task understanding*. NII Shonan 2012
- Bennett et al., *Modeling the impact of short- and long-term behavior on search personalization*. SIGIR 2012
- Broder, *A taxonomy of web search*. SIGIR Forum 2002
- Donato et al., *Do you want to take notes? Identifying research missions in Yahoo! Search Pad*. WWW 2011
- Fox et al., *Evaluating implicit measures to improve the search experience*. TOIS 2005
- Jones & Klinker, *Beyond the session timeout: Automatic hierarchical segmentation of search topics in a query log*. CIKM 2008
- Kotov et al., *Modeling and analysis of cross-session search tasks*. SIGIR 2011
- Richardson, *Learning about the world through long-term query logs*. TWEB 2008
- Russell et al., *Retrospective cued recall: A method for accurately recalling previous user behaviors*. HICCS 2009