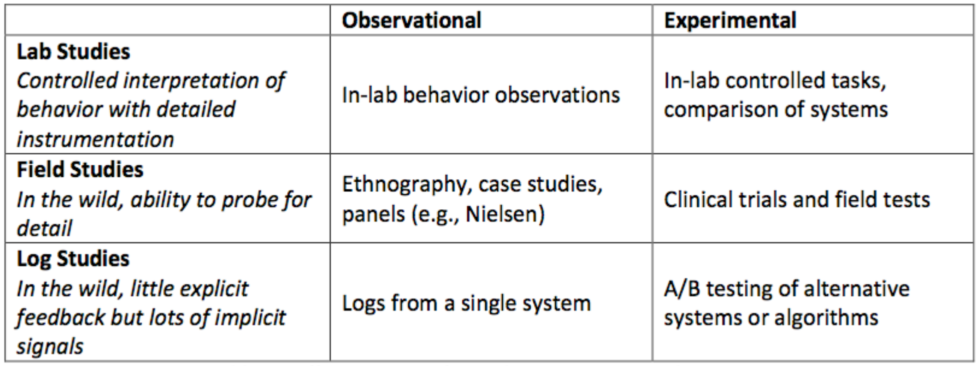
User Studies in Information Retrieval

Diane Kelly
INLS 509: Information Retrieval

April 14, 2014

Types of Studies



	Observational	**Experimental**
Lab Studies *Controlled interpretation of behavior with detailed instrumentation*	In-lab behavior observations	In-lab controlled tasks, comparison of systems
Field Studies *In the wild, ability to probe for detail*	Ethnography, case studies, panels (e.g., Nielsen)	Clinical trials and field tests
Log Studies *In the wild, little explicit feedback but lots of implicit signals*	Logs from a single system	A/B testing of alternative systems or algorithms

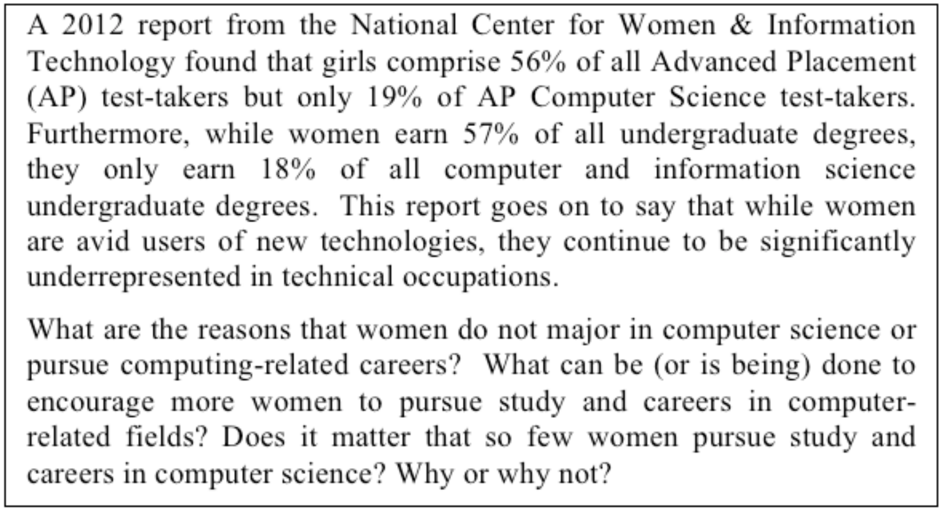
Table 1. Different types of user data in HCI research.

Dumais, S., Jeffries, R., Russell, D. M., Tang, D. & Teevan, J. (forthcoming). Understanding user behavior through log data and analysis. J.S. Olson and W. Kellogg (Eds.), *Human Computer Interaction Ways of Knowing*. New York: Springer, 2014.

Components of a User-Study

- People (i.e., users)
- Experimental "Conditions"
  - Systems/Algorithms
  - Interfaces
  - Instructions
  - ...
- Search Tasks (sometimes called Topics)
- Collection
- Data Collection Techniques
- Measures

Example Task



A 2012 report from the National Center for Women & Information Technology found that girls comprise 56% of all Advanced Placement (AP) test-takers but only 19% of AP Computer Science test-takers. Furthermore, while women earn 57% of all undergraduate degrees, they only earn 18% of all computer and information science undergraduate degrees. This report goes on to say that while women are avid users of new technologies, they continue to be significantly underrepresented in technical occupations.

What are the reasons that women do not major in computer science or pursue computing-related careers? What can be (or is being) done to encourage more women to pursue study and careers in computer-related fields? Does it matter that so few women pursue study and careers in computer science? Why or why not?

Data Collection Techniques

- Logging
  - Client-side
  - Server-side
- Observation
  - Human
  - Machine

Data Collection Techniques

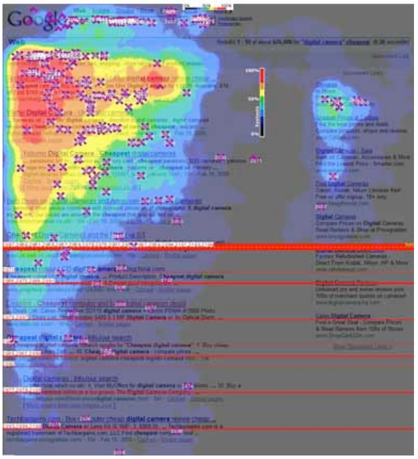
- Questionnaires
  - Demographic
  - Pre-Search
  - Post-Search
  - Post-System
  - Exit
- Other “Self-Report”
  - Individual Difference (e.g., learning style, personality)
  - Relevance

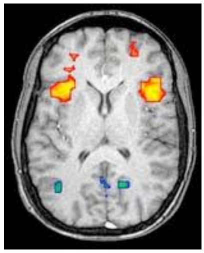
Data Collection Techniques

- Think-aloud & Stimulated Recall
- Interviews
- Evaluation of End Products

Data Collection Techniques

- Eye-Tracking
- Physiological Signals
- Brain Scans (!) (fMRI)





Note to INLS 509-01 Students:
We did not cover Slides 10-15 in class. They correspond to the some of the course readings and just list different measures and two example experimental protocols that illustrate the ‘flow’ of a typical user study.

Measures: Contextual

- Individual Differences
  - Sex, Age, etc.
  - Search Experience
  - Cognitive Ability
  - Personality
  - ...

- Information Needs
  - Task Type
  - Task Complexity
  - Task Difficulty
  - Domain Knowledge
  - ...

Measures: Interaction

- Queries
- Clicks
- Documents Viewed
- Documents Saved
- Time on Task
- ...

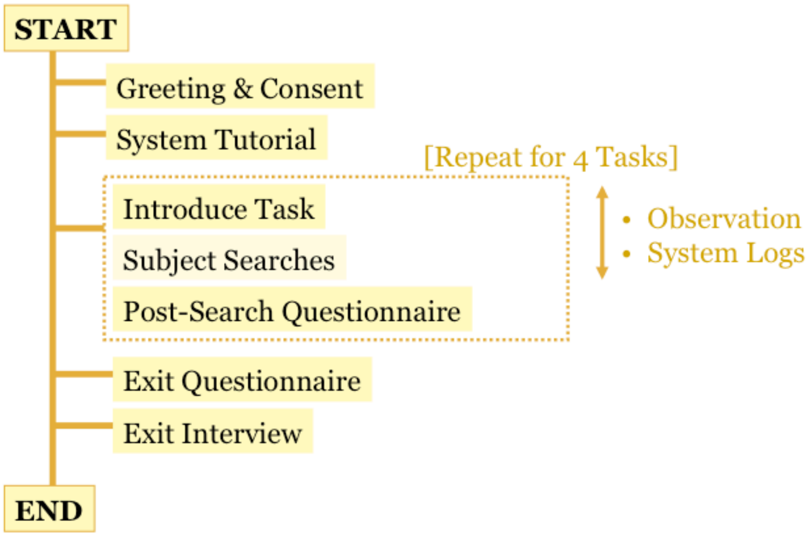
Measures: Performance

- You should have recognized a lot of these!
- Relevance
- Interactive recall and precision
- Interactive TREC precision
- Time-based Measures (e.g., Search Speed)
- Informativeness & Information Gain
- Cost and Utility
- ...

Measures: Feedback from People

- Usability
  - Effectiveness
  - Efficiency
  - Satisfaction
  - Ease of Use, Easy of Learning and Usefulness
- Preference
- Mental Effort and Cognitive Load
- Flow and Engagement (consider *entertainment* types of tasks)
- ...

Basic Protocol



START

Greeting & Consent

System Tutorial

[Repeat for 4 Tasks]

Introduce Task

Subject Searches

Post-Search Questionnaire

- Observation
- System Logs

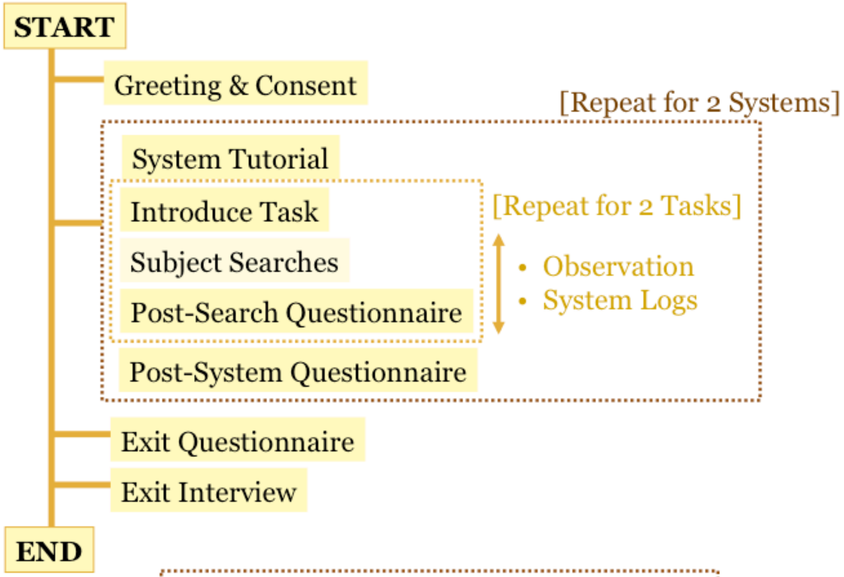
Exit Questionnaire

Exit Interview

END

Each subject uses 1 system to complete 4 tasks

Basic Protocol



START

Greeting & Consent

[Repeat for 2 Systems]

System Tutorial

Introduce Task

Subject Searches

Post-Search Questionnaire

[Repeat for 2 Tasks]

- Observation
- System Logs

Post-System Questionnaire

Exit Questionnaire

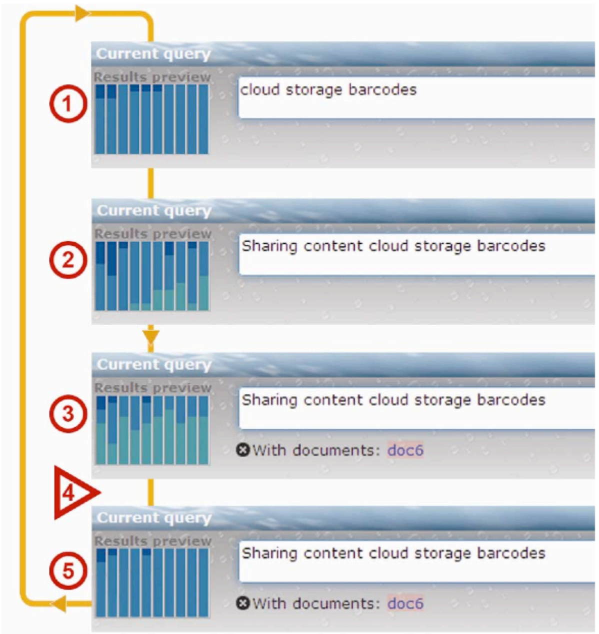
Exit Interview

END

Each subject uses 2 systems to complete 4 tasks

Looking Ahead: Query Preview in Exploratory Search

Qvarfordt, Golovchinsky, Dunnigan & Agapie
SIGIR 2013



cloud storage barcodes

Sharing content cloud storage barcodes

Sharing content cloud storage barcodes

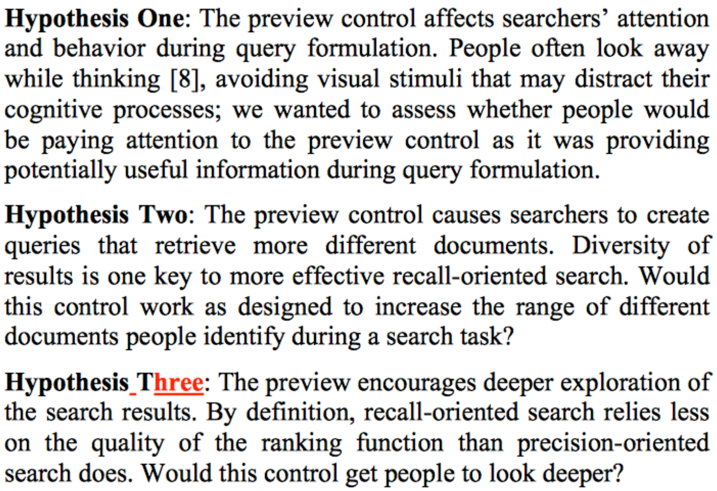
⊗With documents: doc6

Sharing content cloud storage barcodes

⊗With documents: doc6

Figure 1. Example of the preview control as the searcher adds search terms (2), selects a document for relevance feedback (3), runs the query (4), and sees the final results (5)

Hypotheses



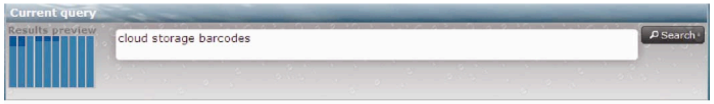
Hypothesis One: The preview control affects searchers' attention and behavior during query formulation. People often look away while thinking [8], avoiding visual stimuli that may distract their cognitive processes; we wanted to assess whether people would be paying attention to the preview control as it was providing potentially useful information during query formulation.

Hypothesis Two: The preview control causes searchers to create queries that retrieve more different documents. Diversity of results is one key to more effective recall-oriented search. Would this control work as designed to increase the range of different documents people identify during a search task?

Hypothesis Three: The preview encourages deeper exploration of the search results. By definition, recall-oriented search relies less on the quality of the ranking function than precision-oriented search does. Would this control get people to look deeper?

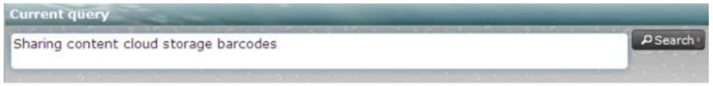
4.1. Experimental design

The experiment was a one-factor within-subjects design. It compared two interface conditions, one with the preview, and one without (see Figure 2 and Figure 3), over a total of six different search topics (three in each condition). Topics were assigned to experimental conditions in a counter-balanced manner. Each participant performed three topics in each condition; each topic was performed once by each participant. Participants were randomly assigned to the counter-balanced configuration of topics, half starting with the preview condition and half starting with the control condition. The study was divided into two sessions, one for each condition usually run on separate days.



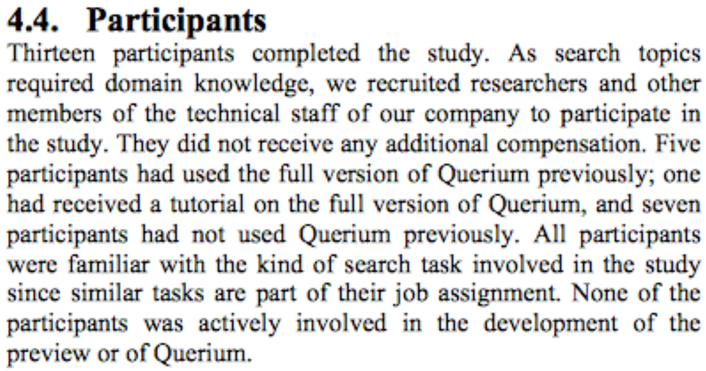
cloud storage barcodes

Figure 2. Query input area for the preview condition.



Sharing content cloud storage barcodes

Figure 3. Query input area for the control condition



4.4. Participants

Thirteen participants completed the study. As search topics required domain knowledge, we recruited researchers and other members of the technical staff of our company to participate in the study. They did not receive any additional compensation. Five participants had used the full version of Querium previously; one had received a tutorial on the full version of Querium, and seven participants had not used Querium previously. All participants were familiar with the kind of search task involved in the study since similar tasks are part of their job assignment. None of the participants was actively involved in the development of the preview or of Querium.

H1: Search Behavior (Interaction)

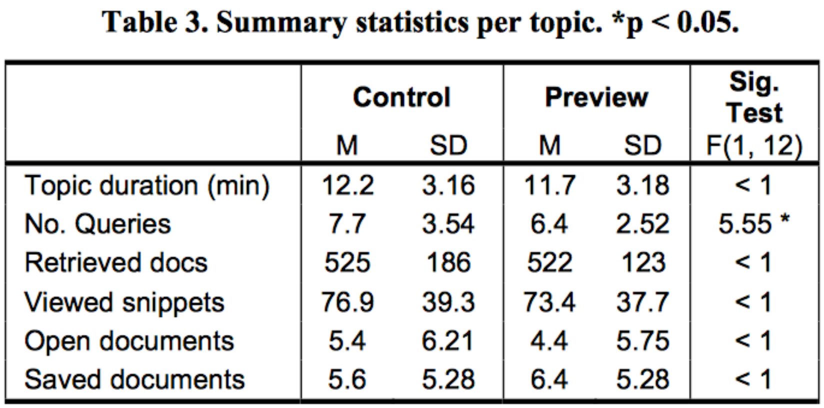
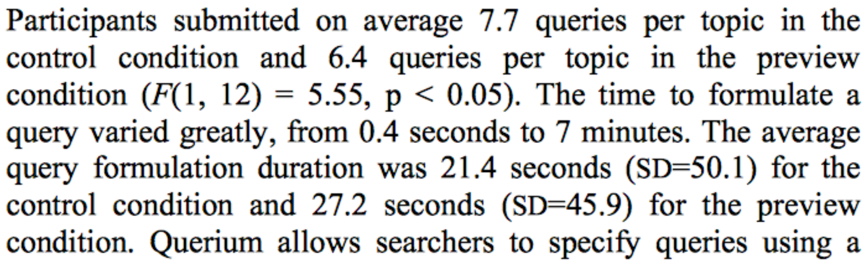


Table 3. Summary statistics per topic. *p < 0.05.

	Control		Preview		Sig. Test
	M	SD	M	SD	F(1, 12)
Topic duration (min)	12.2	3.16	11.7	3.18	< 1
No. Queries	7.7	3.54	6.4	2.52	5.55 *
Retrieved docs	525	186	522	123	< 1
Viewed snippets	76.9	39.3	73.4	37.7	< 1
Open documents	5.4	6.21	4.4	5.75	< 1
Saved documents	5.6	5.28	6.4	5.28	< 1

H1: Search Behavior (Interaction)



Participants submitted on average 7.7 queries per topic in the control condition and 6.4 queries per topic in the preview condition ($F(1, 12) = 5.55$, $p < 0.05$). The time to formulate a query varied greatly, from 0.4 seconds to 7 minutes. The average query formulation duration was 21.4 seconds (SD=50.1) for the control condition and 27.2 seconds (SD=45.9) for the preview condition. Querium allows searchers to specify queries using a

H1: Attention (Eye-Tracking)

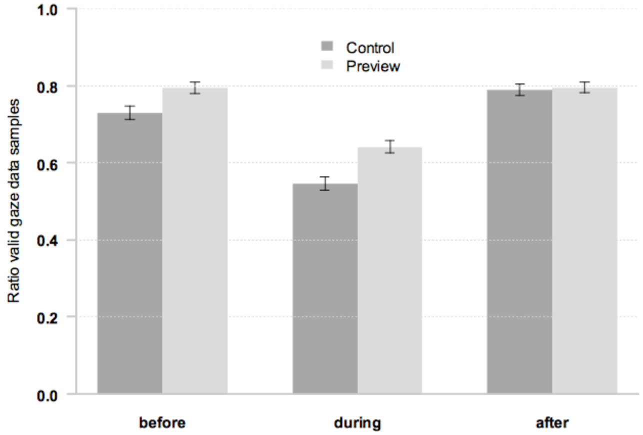


Figure 5. Ratio of valid gaze samples on the query area before, during and after query formulation.

The difference in the ratio of valid gaze samples during query formulation for the two conditions was significant ($F(1,12) = 8.18$, $p < 0.05$). These results show that participants looked at the display significantly more during query formulation when the preview control was available than in the control condition.

H1: Attention (Eye-Tracking)

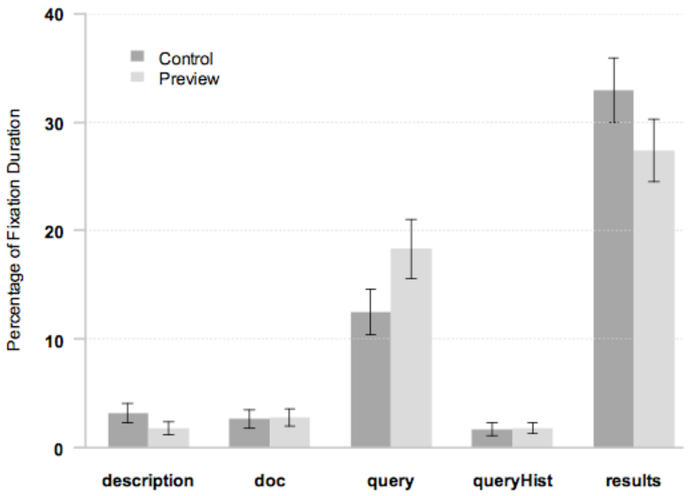


Figure 6. Percentage of attention on UI elements *during* query formulation (total fixation duration on UI element).

A Special Case of Pooling

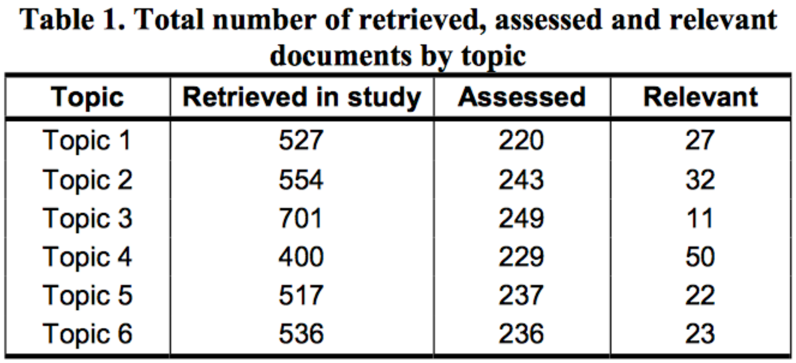


Table 1. Total number of retrieved, assessed and relevant documents by topic

Topic	Retrieved in study	Assessed	Relevant
Topic 1	527	220	27
Topic 2	554	243	32
Topic 3	701	249	11
Topic 4	400	229	50
Topic 5	517	237	22
Topic 6	536	236	23

H2: Retrieval Diversity

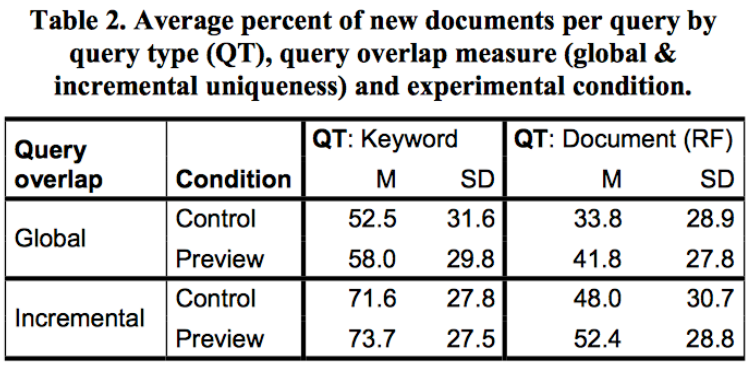


Table 2. Average percent of new documents per query by query type (QT), query overlap measure (global & incremental uniqueness) and experimental condition.

Query overlap	**Condition**	**QT**: Keyword		**QT**: Document (RF)	
		M	SD	M	SD
Global	Control	52.5	31.6	33.8	28.9
	Preview	58.0	29.8	41.8	27.8
Incremental	Control	71.6	27.8	48.0	30.7
	Preview	73.7	27.5	52.4	28.8

H3: Going Deeper

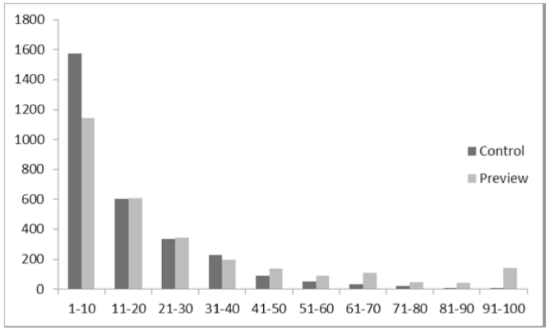


Figure 7. Distribution of viewed snippets by retrieval rank.

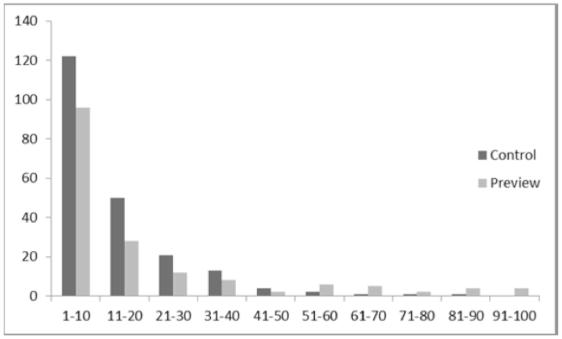


Figure 8. Distribution of opened documents by retrieval rank.

H3: Going Deeper

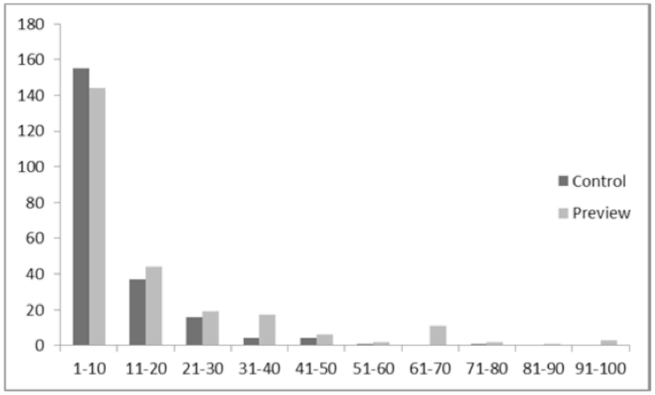


Figure 9. Distribution of saved documents by rank.

H3: Search Performance

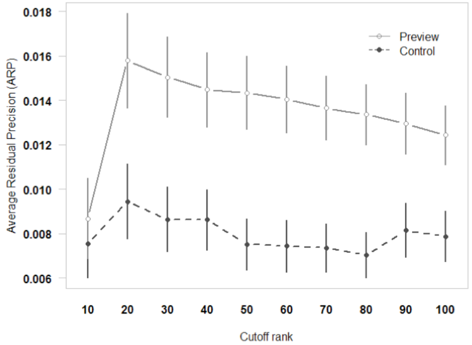


Figure 10. Average Residual Precision (ARP) vs. cutoff rank.

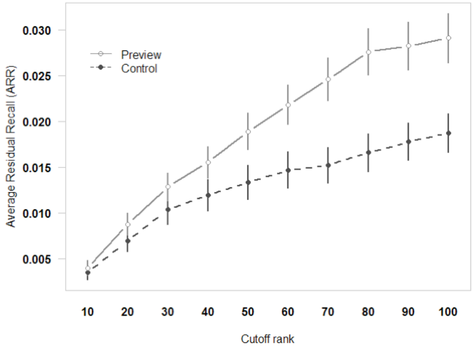


Figure 11. Average Residual Recall (ARR) vs. cutoff rank.

When comparing regular (not residual) recall and precision of the queries the two conditions, we find no statistically-significant differences (both $t(538) < 1$). We also find that the diversity of search results (the number of relevant unique documents retrieved per query) is significantly higher in the experimental condition (52 (SD=30.0) vs. 44 (SD=31.8), $t(538) = 2.7$, $p < 0.01$).

Extensions?

Pre-cursor to Query Previews

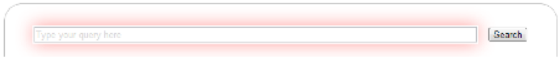

Type your query here

Figure 1. Empty query box.

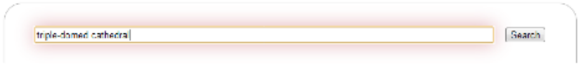

triple-domed cathedral

Figure 2. As the person starts to type, the halo changes.

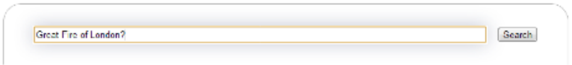

Great Fire of London?

Figure 3. A longer query with a bluer halo.

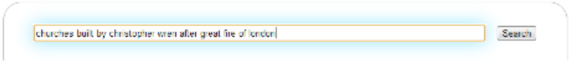

churches built by christopher wren after great fire of london

Figure 4. A long query with a bluish halo.

Agapie, E., Golovchinsky, G., & Qvarfordt, P. (2012). Encouraging behavior: A foray into persuasive computing. *Proceedings of HCIR*.