Leveraging Popular Destinations to Enhance Web Search Interaction*

RYEN W. WHITE, MIKHAIL BILENKO, and SILVIU CUCERZAN
Microsoft Research

This article presents a novel Web search interaction feature that for a given query provides links to Web sites frequently visited by other users with similar information needs. These *popular destinations* complement traditional search results, allowing direct navigation to authoritative resources for the query topic. Destinations are identified using the history of search and browsing behavior of many users over an extended time period, and their collective behavior provides a basis for computing source authority. They are drawn from the end of users' post-query browse trails, where users may cease searching once they find relevant information. We describe a user study that compared the suggestion of destinations with the previously proposed suggestion of related queries, as well as with traditional, unaided Web search. Results show that search enhanced by query suggestions outperforms other systems, in terms of subject perceptions and search effectiveness, for fact-finding search tasks. However, search enhanced by destination suggestions performs best for exploratory tasks, with best performance obtained from mining past user behavior at query-level granularity. We discuss the implications of these and other findings from our study for the design of search systems that utilize user behavior, in particular user browse trails and popular destinations.

Categories and Subject Descriptors: H5.2 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Search process*;
General Terms: Human Factors, Experimentation
Additional Key Words and Phrases: User studies, search destinations, enhanced Web search.

Authors' addresses: Microsoft Research, One Microsoft Way, Redmond, WA 98052; email: {ryenw, mbilenko,silviu}@microsoft.com

* This article expands upon the original paper: White, R.W., Bilenko, M., and Cucerzan, S. (2007). Studying the use of popular destinations to enhance Web search interaction. In *Proceedings of the 30th ACM SIGIR Conference on Research and Development in Information Retrieval*, 159-166.

1. INTRODUCTION

Information Retrieval (IR) systems help people resolve information problems. The quality of search queries submitted to these systems directly affects the quality of the retrieved search results [Croft and Thompson 1987]. However, searchers are often untrained in how to formulate search queries that are both representative of their information needs and useful for document retrieval. As such, they may issue queries that are of insufficient quality to retrieve relevant documents. The problem is potentially more acute in Web search, where a large fraction of users are not proficient in effectively using commercial search engines such as Google, Yahoo!, and Live Search. The problem of improving queries sent to IR systems has been studied extensively in IR research [e.g., Koenneman and Belkin 1996; Beaulieu 1997]. Alternative query formulations, known as *query suggestions*, can be offered to users following an initial query, allowing them to modify the specification of their needs provided to the system, and leading to improved retrieval performance. The recent popularity of Web search engines has enabled query suggestions that draw upon the query reformulation behavior of many users to make query recommendations based on previous user interactions [Beeferman and Berger 2002; Jones et al. 2006].

Leveraging the collective decision-making processes of many users for query reformulation has its roots in *adaptive indexing* [Furnas 1985]. Adaptive indexing addresses the vocabulary mismatch in human-computer communication by dynamically associating system commands with frequently-used variants (e.g., "type" and "output" may be frequently used instead of "print" in command-line environments and are thus added to the system vocabulary as alternatives for the "print" command). In recent years, the application of such techniques has become possible at a much larger scale and in a different context than proposed in early work. Click records from Web search engines provide indications of relevance based on the metadata presented to the user in the result list. Such click records can be useful as training data for learning ranking functions based on machine-learning techniques [Joachims 2002, Agichtein et al. 2006a], for ranking documents when used in isolation [Agichtein et al. 2006b] or when combined with querying information [Radlinski and Joachims 2005], for document annotation [Xue et al. 2004], image search [Craswell and Szummer 2007], and query suggestion [Beeferman and Berger 2002; Jones et al. 2006]. However, recent studies of Web-search behavior [Teevan et al. 2004; White and Drucker 2007] have shown that a significant proportion of interaction during search sessions involves pages visited beyond clicks on search engine results. Algorithms that focus solely on search engine interactions miss this potentially valuable information source, which limits their potential effectiveness. In addition, interaction-based algorithms may be less potent when the information need is exploratory, since a large proportion of user activity for such information needs may occur beyond search engine interactions [Anick 2003].

In cases where directed searching is only a fraction of users' information-seeking behavior, the utility of other users' clicks over the space of top-ranked results may be limited, as it does not cover the subsequent browsing behavior. At the same time, user navigation that follows search engine interactions provides implicit endorsement of Web resources preferred by users, which may be particularly valuable for exploratory search tasks. Thus, we propose exploiting *a combination* of past searching and browsing behavior to enhance users' Web search interactions. Since access to large volumes of interaction log data is often limited, IR researchers have generally simulated post-query behavior, e.g., to evaluate relevance feedback algorithms and their variants [White et al. 2005, Smucker and Allan 2006]. However, browser plug-ins and proxy server logs provide access to the browsing patterns of users that transcend search engine interactions. If we can leverage these patterns, then perhaps we can build better ranking algorithms than just by using search engine interactions alone. For example, Agichtein et al. [2006b] used browsing features to train a ranking algorithm and showed that search effectiveness improved, but did not consider users' entire post-query navigation trails. Bilenko and White [2008] developed ranking algorithms that utilized the complete post-query trails of many users, and demonstrated improved retrieval performance as a result.

In this article we present a user study of a technique that exploits the searching and browsing behavior of many users to suggest authoritative sources, referred to as *destinations* henceforth, in addition to the regular search results. The destinations may not be among the top-ranked results, may not contain the queried terms, or may not even be indexed by the search engine. Instead, they are Web sites or Web domains at which other users end up frequently after submitting the same or similar queries and then browsing away from initially clicked search results. We conjecture that search destinations popular across a large number of users capture "the wisdom of the crowds" for information needs, and our results support this hypothesis.

Log-based analysis of browsing patterns within particular Web sites can help understand user needs and intentions, and consequently inform the redesign of site structure to support them [Pirolli et al. 1996, Pitkow and Pirolli 1997, Anderson et al. 2001]. Browse paths followed by human "trail blazers" [Bush 1945] through information spaces can implicitly represent similarities and associations between visited items, that can be incorporated in trail recommendation systems [Chalmers et al. 1998]. The approach we describe in this article is similar in that it uses trails to infer interests, but on a much larger scale and for a different purpose (i.e., destination suggestion rather than trail recommendation).

O'Day and Jeffries [1993] identified "teleportation" as an information-seeking strategy employed by users jumping to their previously-visited information targets, while Anderson et al. [2001] applied similar principles to support the rapid navigation of Web sites on mobile devices. The very need for users to exhibit more than a trivial number of post-query interactions relates to the inability of search systems to fully understand the information needs of their users. As has been suggested already, even the "perfect" search engine, which returns exactly what is sought given a fully-specified information need, cannot address the circumstances where: (i) users are unable to specify their information needs at a level to make the system effective [Teevan et al. 2004], or (ii) they use a vocabulary that does not align with that used during document indexing [Furnas et al. 1987]. In such cases, ranking algorithms or result presentation techniques based on user interaction rather than text matching may be beneficial. Wexelblat and Maes [1999] described a system to support within-domain navigation based on the browse trails of other users. Research in collaborative filtering and recommender systems has also addressed similar issues, but in areas such as question-answering [Hickl et al. 2006], relatively small online communities [Smyth et al. 2004], and within restricted domains such as newswire [Resnick et al. 1994], music albums and artists [Shardanand and Maes 1995], or e-commerce [Sarwar et al. 2000]. However, to our best knowledge, these techniques have not been directly applied to support Web search. Perhaps the nearest instantiation is search engines' offering of several within-domain shortcuts (or "deeplinks") below the title of popular Web sites in the search results list. While these may account for user behavior on the target site, they typically save at most a few user clicks on a specific site. In contrast, our proposed approach can transport users to locations many clicks beyond the search result across multiple sites, saving time and giving them a broader perspective on the available related information adjacent to search results.

The conducted user study investigates the effectiveness of including links to popular destinations as an additional interface feature on search engine result pages. We compare two variants of this approach against the suggestion of related queries and unaided Web search, and seek answers to questions on: (i) user preference and search effectiveness for fact-finding and exploratory search tasks, and (ii) the preferred distance between query and destination used to identify popular destinations from past behavior logs. The results indicate that suggesting popular destinations to users attempting exploratory tasks provides best results in key aspects of the information-seeking experience, while providing query refinement suggestions is most desirable for fact-finding tasks.

We structure the remainder of this article as follows. In Section 2 we describe the extraction of search and browsing trails from user activity logs, and their use in identifying top destinations for new queries. Section 3 describes the design of the user study. Section 4 presents the study findings and Section 5 discusses these findings and their implications. We conclude in Section 6.

2. SEARCH TRAILS AND POPULAR DESTINATIONS

We used Web activity logs containing searching and browsing activity collected with permission from a very large number of Windows Live Toolbar[2] users over a five-month period between December 2005 and April 2006. Each log entry included an anonymous user identifier, a timestamp, a unique browser window identifier, and the URL of a visited Web page. This information was sufficient to reconstruct temporally-ordered sequences of viewed pages that we refer to as “trails”. The only limitation in using these logs was that users exhibited a higher degree of loyalty to Microsoft’s online services such as Live Search than may be expected from the average Web user. In this section, we summarize the process used to extract trails, their features, and destinations (i.e., trail end-points).

2.1 Trail Extraction

For each user, interaction logs were grouped based on browser identifier information. Within each browser instance, participant navigation was summarized as a path known as a *browser trail*, from the first to the last Web page visited in that browser. Located within some of these trails were *search trails* that originated with a query submission to a commercial search engine such as Google, Yahoo!, Live Search, and Ask. Our proposed technique uses the pages that lie at the end of these search trails to identify popular destinations for a given search query.

After originating with a query submission to a search engine, trails proceed until a point of termination where it is assumed that the user has completed their information-seeking activity. Trails must contain pages that are either: search result pages, search engine homepages, or pages connected to a search result page via a sequence of clicked hyperlinks. All page views, including cache-based browsing events, are captured by the toolbar and included in the trail.

Search trails originate with a directed search (i.e., a query issued to a search engine), and proceed until a point of termination where it is assumed that the user has completed their information-seeking activity. The following termination activities were used to determine trail end points:

- *Return to homepage:* Returning to a homepage is assumed to mark the end of a trail.
- *Check email or logon to service:* Checking Web-based e-mail, or logging-in to online services such as MySpace or del.ico.us, was used as an indicator that the search trail had terminated.
- *Type URL or visit bookmarked pages:* Entering a URL directly into the address bar of the browser, or selecting a bookmark, terminated the search trail. The only exceptions were visits to search engine homepages (e.g., http://www.google.com), which may be a necessary part of the current search activity, particularly if participants decide to switch search engines mid-trail.
- *Page timeout:* If the display time for any page exceeded 30 minutes this was assumed to mark the termination of a search trail. Similar timeouts have been used previously to demarcate sessions [Catledge and Pitkow 1995, Downey et al. 2007].
- *Close browser window*

[2] The Windows Live Toolbar is a plug-in to the Internet Explorer browser that provides additional browser functionality in return for users providing consent for their page-level interactions to be logged and used to improve their experience.

These trail termination points are determined based on the above heuristics, and thus, some may be related to the active search task, e.g., checking email to support task resolution, or running multiple searches on the same topic concurrently in different browser windows (or different tabs within the same window). However, we felt that removing potential noise from the trails outweighs the cost of possibly truncating some trails early. If a page (at step *i* in the trail) meets any of the above criteria, the trail is assumed to terminate on the previous page (i.e., step *i* – 1).

To illustrate how search trails are constructed, we present an example of how a search trail is extracted from a candidate browser trail. To simplify the exposition, we represent the browser trail as a Web behavior graph [Card et al. 2001], shown in Figure 1.[4] The graph shows user activity within a browser trail, from their homepage (*H*) through to the point at which they close the browser (*X*). The nodes of the graph represent Web pages that the user has visited: rectangles represent page views (e.g., *P3*) and rounded rectangles represent search engine queries and subsequent result pages (e.g., *S1*). Vertical lines represent backtracking to an earlier state (such as returning to a page of results in a search engine after following an unproductive link). A "back" arrow, such as that below *P4* implies that the user is about to revisit a page seen earlier in the browser trail. Time runs left to right and then from top to bottom. The region of the graph shown in gray represents a Web-based email service, in this case Microsoft's "hotmail.com".

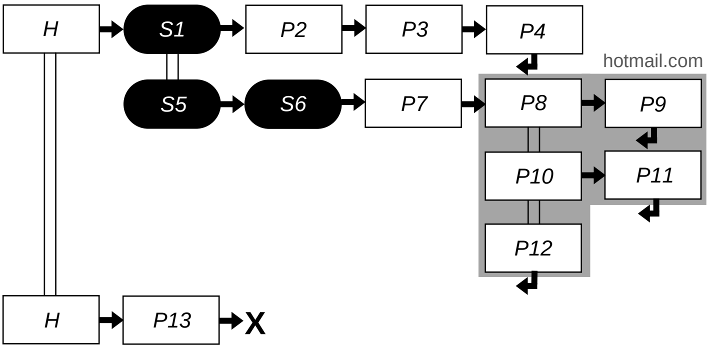


Fig. 1. Browser trail as Web behavior graph.

In the example browser trail shown above, the user is pursuing information related to their original search query (*S1*). As they navigate, they perform the following activities:

- begin at their homepage, which is a search engine (*H*);
- enter search query *S1* and browse across several pages *(P2-P4)* starting from a click on search results *(P2)*;
- enter two search queries (*S5-S6*) and browse to one search result *(P7*);
- check their Web-based email (*P8-P12*);
- return to their homepage (*H*) and browse to one linked page (*P13*);
- close the browser window (*X*).

Given this browser trail, the search trail runs from *S1* (the submission of the first query) to *P7* (the last page viewed before email checking). The visit to the Web-based email

[4] Web behavior graphs are a variant of problem behavior graphs [Newell and Simon 1972], and are useful for viewing navigation patterns.

service matches one of the five termination criteria described earlier in this section. The full search trail in the example is therefore *S1→P2→ P3→P4→S5→S6→P7*.

Since searches generally involve multiple query iterations, running trails over multiple iterations allows us to analyze richer interaction patterns than for individual queries. Given the nature of the interaction logs generated by our client-side application, we were able to extract search trails relatively easily using the approach described here, circumventing the need to rely on probabilistic models of behavior, e.g., [Pitkow and Pirolli 1999].

There are two types of search trails we consider: *session trails* and *query trails*. Session trails transcend multiple queries and terminate only when one of the five termination criteria above are satisfied. Query trails use the same termination criteria as session trails, but also terminate upon submission of a new query to a search engine. Figure 1 contains a single session trail (*S1→P2→ P3→P4→S5→S6→P7*) and three query trails (*S1→P2→ P3→P4, S5*, and *S6→P7*). The destination in the session trail is *P7* and in the two non-singular query trails are *P4* and *P7*. While alternative methodologies for trail extraction can also be considered (e.g., by accounting for query chains [Radlinski and Joachims 2007]), they fall outside the scope of the presented research and remain an interesting challenge for future work.

2.2 Trail and Destination Analysis

We extracted approximately 14 million query trails and 4 million session trails from the logs. To ensure that trails involved an information-seeking activity, all trails began with a query to a search engine. Table I presents summary statistics (the mean (M) and the standard deviation (SD)) for features of the query and session trails. Differences in user interaction between the last domain[5] on the trail (Domain *n*) and all domains visited earlier (Domains 1 to (*n* – 1)) are particularly important, because they highlight the wealth of user behavior data not captured by logs of search engine interactions. Statistics are computed across all trails with two or more steps (i.e., those trails where at least one search result was clicked). We use Web *domains* rather than Web *pages* since the log sample at our disposition was insufficient to make reliable recommendations for many pages. Aggregating page visits to domain visits resolves the data sparseness problem at the expense of reduced granularity.

Table I. Summary Statistics (Mean and Standard Deviation) for Search Trails.

Measure		Query trails		Session trails	
		M	SD	M	SD
Number of unique domains		2.0	3.4	4.3	5.1
Total page views	All domains	4.8	5.6	16.2	17.6
	Domains 1 to (*n* – 1)	1.4	2.0	10.1	12.5
	Domain *n* (destination)	3.4	2.7	6.2	7.8
Total time spent (secs)	All domains	172.6	185.4	621.8	716.9
	Domains 1 to (*n* – 1)	70.4	80.9	397.6	401.2
	Domain *n* (destination)	102.3	111.2	224.1	255.1

The statistics suggest that users generally browse far from the search results page (i.e., around five steps), and visit a range of domains during the course of their search. On

[5] In this work, we use the term *domain* to refer to both the top-level domain of a site (e.g., microsoft.com) or a lower-level domain if sufficient log data exists for pages in it (e.g., support.microsoft.com).

average, users visit two unique (non search-engine) domains per query trail, and just over four unique domains per session trail. This suggests that users often do not find all the information they seek on the first domain they visit. In query trails, users also visit more pages and spend significantly more time on the last domain in the trail compared to all previous domains combined.[6] These distinctions between the last domains in the trails and the other domains may indicate user interest, page utility, or page relevance.[7]

2.3 Destination Prediction

For frequent queries, most popular destinations identified from Web activity logs could be simply stored for future lookup at search time. However, we have found that over the six-month period covered by our dataset, 56.9% of queries are unique, while 97% queries occur 10 or fewer times, accounting for 19.8% and 66.3% of all searches respectively (these numbers are comparable to those reported in previous studies of search engine query logs [Silverstein et al. 1999, Jansen and Spink 2002]). Therefore, a lookup-based approach would prevent us from reliably suggesting destinations for a large fraction of searches. To overcome this problem, we employ a term-based destination prediction model that gives us more coverage than a query-based approach. We now describe how we score domains for a given query using the term-based destination prediction approach.

As discussed earlier, we extract two types of destinations from search trails: query destinations (i.e., domains that lie at the end of a query trail) and session destinations (i.e., domains that lie at the end of a session trail). Given that many users ultimately end up on these domains following the submission of a query and post-query browsing we regard destinations as potentially authoritative sources for the query topic. For both destination types, we obtain a corpus of query-destination pairs and use it to construct a term-vector representation of destinations that is analogous to the classic *tf.idf* document representation in traditional IR [Salton and Buckley 1988]. Then, given a new query *q* consisting of *k* terms $t_1 \ldots t_k$, we identify highest-scoring destinations using the following similarity function:

$$S(d,q) = \pi_d + \sum_{i=1:k} w_q(t_i) w_d(t_i)$$

where π_d is a smoothed inverse destination frequency, $\pi_d = \log \frac{\sum_{d' \in D} n_Q(d') + \lambda}{n_Q(d) + \lambda}$, computed using the per-query-normalized number of trails ending at destination d, $n_Q(d)$. Query and destination term weights, $w_q(t_i)$ and $w_d(t_i)$, are computed using standard *tf.idf* weighting and query- and session-normalized smoothed *tf.idf* weighting, respectively.

Analogously to the connection that exists between heuristic *tf.idf* retrieval and smoothed unigram language models [Zhai and Lafferty 2004], this scoring function is related to a probabilistic model that incorporates the smoothed unigram language model and a generative model that incorporates users, queries, terms, and destinations. We use this function to select popular destinations in two of the experimental systems described in the next section.

[6] Independent measures t-test: t(~60M) = 3.89, p < .001

[7] We tested the topical relevance of the destinations for a subset of around ten thousand queries for which we had human judgments. The average rating of most of the destinations lay between "good" and "excellent". Visual inspection of those that did not lie in this range revealed that many were either relevant but had no judgments, or were related but had indirect query association (e.g., "petfooddirect.com" for query *[dogs]*).

3. USER STUDY

To examine the usefulness of destinations, we conducted a user study investigating the perceptions and performance of 36 subjects on four Web search systems, two with destination suggestions, one with query suggestion, and one baseline.

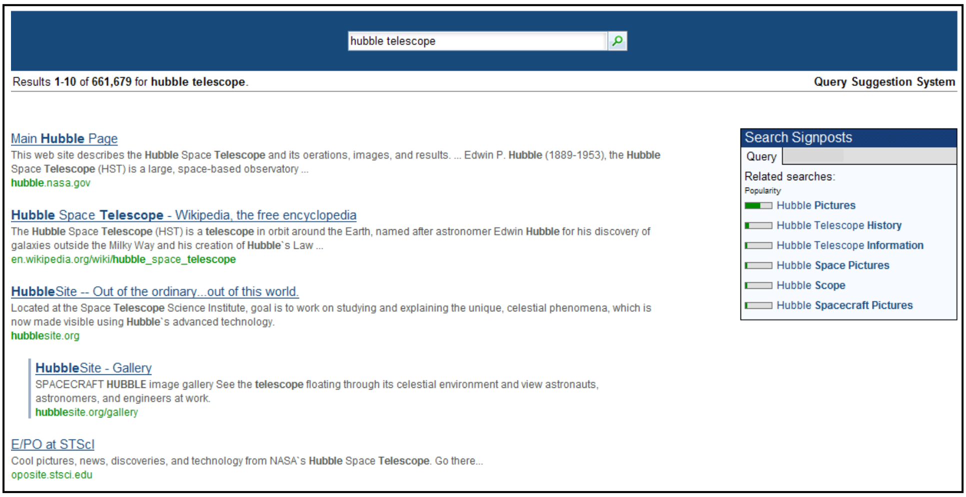
3.1 Systems

Four systems were used in this study: a baseline Web search system with no explicit support for query refinement (*Baseline*), a search system with a query suggestion component that recommends additional queries (*QuerySuggestion*), and two systems that augment baseline Web search with destination suggestions using either end-points of query trails (*QueryDestination*), or end-points of session trails (*SessionDestination*). We now describe the systems in more detail.

3.1.1 *System 1: Baseline*. To establish baseline performance for comparison with other systems, we employed a masked interface to a popular search engine (Live Search) without additional support features. This system submits the user-constructed query to the search engine and returns ten top-ranking documents retrieved by the engine. To remove potential bias that may be caused by subjects' prior perceptions, we remove all identifying information such as search engine logos and distinguishing interface features. As is standard in Web search engines, *Baseline* provides a text box on the result page for query refinement, a description of the total number of search results obtained, and short descriptions of the results along with links to them.

3.1.2 *System 2: QuerySuggestion.* In addition to the basic search functionality offered by *Baseline*, *QuerySuggestion* provides suggestions about further query refinements that searchers can make following an initial query submission. These suggestions are computed using a query log from the Live Search engine over the time period used for trail generation. For each target query, we retrieve two sets of candidate suggestions that contain the target query as a substring. One set is composed of the 100 most frequent queries with the target as a substring, while the second set contains the 100 most frequent queries that followed the target query in a search session. Each candidate query in the union of these two sets is then scored by multiplying its smoothed overall frequency by its smoothed frequency of following the target query in past search sessions, using Laplacian smoothing. Based on these scores, the six top-ranked query are employed as suggestions. If fewer than six suggestions are found, iterative back-off is performed using progressively longer suffixes of the target query, using a similar strategy to that described in Jones et al. [2006].

Suggestions were offered in a box positioned on the top-right of the result page, adjacent to the search results. Figure 2 shows the results page containing the suggestions offered for the query *[hubble telescope]*. To the left of each query suggestion is an icon similar to a progress bar that encodes its normalized popularity. Clicking a suggested query retrieves the page of search results for that query.



hubble telescope

Results 1-10 of 661,679 for hubble telescope.

Query Suggestion System

Main Hubble Page
This web site describes the Hubble Space Telescope and its oerations, images, and results. ... Edwin P. Hubble (1889-1953), the Hubble Space Telescope (HST) is a large, space-based observatory ...
hubble.nasa.gov

Hubble Space Telescope - Wikipedia, the free encyclopedia
The Hubble Space Telescope (HST) is a telescope in orbit around the Earth, named after astronomer Edwin Hubble for his discovery of galaxies outside the Milky Way and his creation of Hubble`s Law ...
en.wikipedia.org/wiki/hubble_space_telescope

HubbleSite -- Out of the ordinary...out of this world.
Located at the Space Telescope Science Institute, goal is to work on studying and explaining the unique, celestial phenomena, which is now made visible using Hubble`s advanced technology.
hubblesite.org

HubbleSite - Gallery
SPACECRAFT HUBBLE image gallery See the telescope floating through its celestial environment and view astronauts, astronomers, and engineers at work.
hubblesite.org/gallery

E/PO at STScI
Cool pictures, news, discoveries, and technology from NASA`s Hubble Space Telescope. Go there...
oposite.stsci.edu

Search Signposts

Query

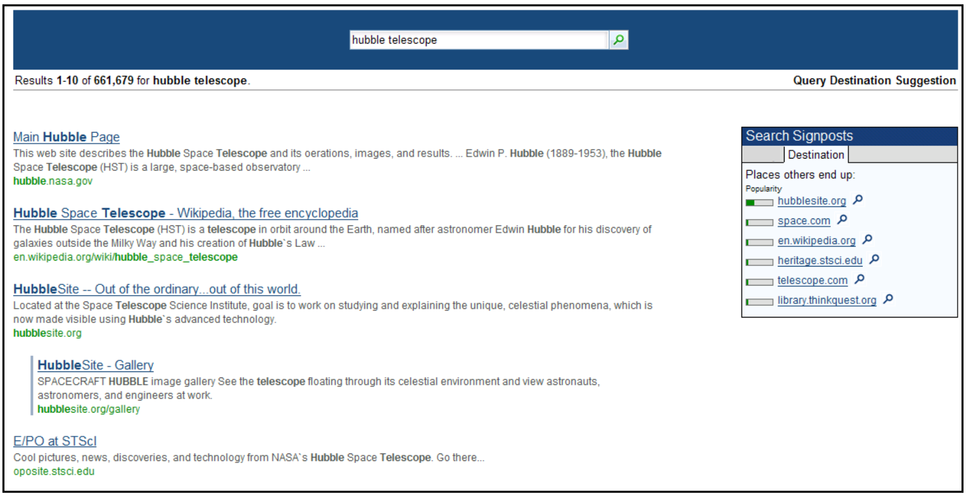
Related searches:

Popularity

- Hubble Pictures
- Hubble Telescope History
- Hubble Telescope Information
- Hubble Space Pictures
- Hubble Scope
- Hubble Spacecraft Pictures

Fig. 2. Query suggestion presentation in *QuerySuggestion*.

3.1.3 *System 3: QueryDestination*. This system uses an interface similar to *QuerySuggestion*. However, instead of showing query suggestions for the submitted query, *QueryDestination* suggests up to six popular destinations frequently visited by other users who submitted queries similar to the current one, and identified as described in the previous section.[9] Figure 3 shows the results page containing the destinations suggested for the query [*hubble telescope*].



hubble telescope

Results 1-10 of 661,679 for hubble telescope.

Query Destination Suggestion

Main Hubble Page
This web site describes the Hubble Space Telescope and its oerations, images, and results. ... Edwin P. Hubble (1889-1953), the Hubble Space Telescope (HST) is a large, space-based observatory ...
hubble.nasa.gov

Hubble Space Telescope - Wikipedia, the free encyclopedia
The Hubble Space Telescope (HST) is a telescope in orbit around the Earth, named after astronomer Edwin Hubble for his discovery of galaxies outside the Milky Way and his creation of Hubble`s Law ...
en.wikipedia.org/wiki/hubble_space_telescope

HubbleSite -- Out of the ordinary...out of this world.
Located at the Space Telescope Science Institute, goal is to work on studying and explaining the unique, celestial phenomena, which is now made visible using Hubble`s advanced technology.
hubblesite.org

HubbleSite - Gallery
SPACECRAFT HUBBLE image gallery See the telescope floating through its celestial environment and view astronauts, astronomers, and engineers at work.
hubblesite.org/gallery

E/PO at STScI
Cool pictures, news, discoveries, and technology from NASA`s Hubble Space Telescope. Go there...
oposite.stsci.edu

Search Signposts

Destination

Places others end up:

Popularity

- hubblesite.org
- space.com
- en.wikipedia.org
- heritage.stsci.edu
- telescope.com
- library.thinkquest.org

Fig. 3. Destination suggestion presentation in *QueryDestination*.

To keep the interface uncluttered, the page title of each destination is shown on hover over the page URL. Next to the destination name, there is a clickable icon that allows the user to execute a search for the current query within the destination domain displayed, by using the advanced search operator *site:*. This icon was included to address anticipated user concern about them not being directed to a particular Web page, as in normal in Web search. Clicking on the icon would return a list of Web pages in that domain that

[9] To improve reliability, in a similar way to *QuerySuggestion*, destinations are only shown if their relevance score exceeds a predefined confidence threshold.

contained the query terms. If no pages in the domain contain the query terms, the interface presents no search results for the domain search. We show destinations as a separate list, rather than increasing their search result rank, since they may topically deviate from the original query (e.g., focusing on related topics or not containing the original query terms).

3.1.4 *System 4: SessionDestination.* Interface functionality in *SessionDestination* is analogous to *QueryDestination*. The only difference between the two systems is the definition of trail end-points for queries used in computing the top destinations. *QueryDestination* directs users to the domains at which other users end up for the target query or related queries. In contrast, *SessionDestination* directs users to the domains other users visit at the end of the search session that follows the active or similar queries. This downgrades the effect of multiple query iterations (i.e., the system targets domains where users end up after submitting all queries), rather than directing searchers to potentially irrelevant domains that may precede a query reformulation.

3.2 Research Questions

We were interested in determining the value of popular destinations. To do this, we attempt to answer the following research questions:

RQ1: Are popular destinations preferable and more effective than query refinement suggestions and unaided Web search for:

a. Searches that are fact-finding?
b. Searches that are exploratory?

RQ2: Should popular destinations be taken from the end of query trails or the end of session trails?

3.3 Subjects

36 subjects (26 males and 10 females) participated in our study. They were recruited through an email announcement within Microsoft Corporation. The selected subjects held a range of roles in different divisions of the company. The average age of subjects was 34.9 years (max=62, min=27, standard deviation (SD)=6.2). All are familiar with Web search, and conduct 7.5 searches per day on average (SD=4.1). Thirty-one subjects (86.1%) reported general awareness of the query refinements offered by commercial Web search engines.

3.4 Tasks

Since the search task may influence information-seeking behavior [Beaulieu 1997], we made task type an independent variable in the study. We constructed six fact-finding tasks and six open-ended, exploratory tasks that were rotated between systems and subjects as described in the next section. Figure 4 shows examples of the two task types.

Fact-finding task
Identify three tropical storms (hurricanes and typhoons) that have caused property damage and/or loss of life.

Exploratory task
You are considering purchasing a Voice Over Internet Protocol (VoIP) telephone. You want to learn more about VoIP technology and providers that offer the service, and select the provider and telephone that best suits you.

Fig. 4. Examples of fact-finding and exploratory tasks.

The fact-finding search tasks required subjects to search for particular items of information (e.g., activities, discoveries, names) for which the target was clear. Exploratory tasks were phrased as simulated work task situations [Borlund 2003], i.e., short search scenarios that were designed to reflect real-life information needs. These tasks generally required subjects to gather background information on a topic or gather sufficient information to make an informed decision. A similar task classification has been used successfully in previous work [White and Marchionini 2007]. Tasks were taken and adapted from the Text Retrieval Conference (TREC) Interactive Track [Dumais and Belkin 2005], and questions posed on question-answering communities (Yahoo! Answers, Google Answers, and Windows Live QnA). To motivate the subjects during their searches, we invited them to select tasks that they found "more interesting". We include the task descriptions for all 12 tasks in the appendix. We allowed them to select two fact-finding and two exploratory tasks at the beginning of the experiment from the six possibilities for each category, before seeing any of the systems or having the study described to them. Prior to the experiment, all tasks were pilot tested with a small number of different subjects to help ensure that they were comparable in difficulty and "selectability" (i.e., the likelihood that a task would be chosen given the alternatives). We also verified that there was sufficient log data to generate query and destination suggestions for each of the tasks. Post-hoc analysis of the distribution of tasks selected by subjects during the full study showed no preference for any task in either category.

3.5 Design and Methodology

The study used a within-subjects (repeated measures) experimental design. System had four levels (corresponding to the four experimental systems) and search tasks had two levels (corresponding to the two task types). System and task-type order were rotated according to a Graeco-Latin square design, where each square (or block) comprised four subjects. This design allowed us to counteract learning effects and fatigue by ensuring that (i) every row and every column in the square to contain exactly one instance of each system and task, and (ii) no two cells the same ordered system-task pair.

Subjects were tested independently and each experimental session lasted up to one hour. We adhered to the following procedure:

1. Upon arrival, subjects were asked to select two fact-finding and two exploratory tasks without replacement from the six tasks of each type.
2. Subjects were given an overview of the study in written form that was read aloud to them by the experimenter.
3. Subjects completed a demographic questionnaire focusing on aspects of search experience.
4. For each of the four interface conditions:
    a. Subjects were given an explanation of interface functionality lasting around 2 minutes.
    b. Subjects were instructed to attempt the task on the assigned system searching the Web, and were allotted up to 10 minutes to do so. They were asked to record answers/notes in written form on a sheet provided by the experimenter.
    c. Upon completion of the task, subjects were asked to complete a post-search questionnaire.
    d. After completing the tasks on the four systems, subjects answered a final questionnaire comparing their experiences on the systems.
5. Subjects were thanked and compensated.

In the next section we present the findings of our study.

4. FINDINGS

In this section, we use the data obtained from the user study to address our hypotheses about query and destination suggestions, providing information on the effect of task type where appropriate. We used parametric statistical testing and set the level of significance to $p < .05$, unless otherwise stated. All Likert scales and semantic differentials used a 5-point scale where a rating closer to one signifies more agreement with the attitude statement. To reduce the number of Type I errors i.e., rejecting null hypotheses that were true, we used a Bonferroni correction to adjust the alpha level when we performed multiple tests.

4.1 Subject Perceptions

In this section, we present findings on how subjects perceived the systems that they used. Some systems contained popular destinations and others did not. Therefore, we were able to determine the perceived value of destinations to subjects by comparing subject responses to post-search (per-system) questionnaires and a final questionnaire asking them to compare all systems they had used.

4.1.1 *Search Process.* Addressing the first research question requires insight into subjects' perceptions of the search experience on each of the four systems. In the post-search questionnaires, we asked subjects to complete four 5-point semantic differentials indicating their responses to the attitude statement: "*The search we asked you to perform was*". The paired stimuli offered as responses were: "*relaxing*"/"*stressful*", "*interesting*"/ "*boring*", "*restful*"/"*tiring*", and "*easy*"/"*difficult*". The mean obtained differential values are shown in Table II for each system and each task type. The value corresponding to the differential "All" represents the mean of all four differentials, providing an overall measure of subjects' perceptions.

Table II. Subject Perceptions of Search Process (lower = better).

Differential	Fact-finding				Exploratory			
	B	QS	QD	SD	B	QS	QD	SD
Easy	2.6	**1.6**	1.7	2.3	2.5	2.6	**1.9**	2.9
Restful	2.8	**2.3**	2.4	2.6	2.8	2.8	**2.4**	2.8
Interesting	2.4	2.2	**1.7**	2.2	2.2	**1.8**	**1.8**	2
Relaxing	2.6	**1.9**	2	2.2	2.5	2.8	**2.3**	2.9
All	2.6	2	**1.9**	2.3	2.5	2.5	**2.1**	2.7

Each cell in Table II summarizes subject responses for 18 task-system pairs (18 subjects who ran a fact-finding task on *Baseline* (B), 18 subjects who ran an exploratory task on *QuerySuggestion* (QS), etc.). The most positive response across all systems for each differential-task pair is shown in bold. We applied two-way analysis of variance (ANOVA) to each differential across all four systems and two task types. Subjects found the search easier on *QuerySuggestion* and *QueryDestination* than on the other systems for fact-finding tasks.[11] For exploratory tasks, only searches conducted on *QueryDestination* were easier than on the other systems.[12] Subjects indicated that exploratory search tasks on the three non-baseline systems were more stressful (i.e., less "*relaxing*") than the fact-finding tasks.[13] As we will discuss in more detail in Section 4.1.3, subjects regarded the familiarity of *Baseline* as a strength, and may have been uncomfortable attempting a

[11] *easy*: $F(3,136) = 4.71$, $p = .0037$; Tukey *post-hoc* tests: all $p \leq .008$

[12] *easy*: $F(3,136) = 3.93$, $p = .01$; Tukey *post-hoc* tests: all $p \leq .012$

[13] *relaxing*: $F(1,136) = 6.47$, $p = .011$

more complex task while learning a new interface feature such as the query or destination suggestions.

4.1.2 *Interface Support.* We solicited subjects' opinions on the search support offered by *QuerySuggestion*, *QueryDestination*, and *SessionDestination*. The following 5-point Likert scales (ranging from "strongly agree" to "strongly disagree") and semantic differentials were used:

- Likert scale A: *"Using this system enhances my effectiveness in finding relevant information."* (Effectiveness)[14]
- Likert scale B: *"The queries/destinations suggested helped me get closer to my information goal."* (CloseToGoal)
- Likert scale C: *"I would re-use the queries/destinations suggested if I encountered a similar task in the future"* (Re-use)
- Semantic differential A: *"The queries/destinations suggested by the system were:* "*relevant*"/"*irrelevant*", "*useful*"/"*useless*", "*appropriate*"/"*inappropriate*".

We did not include these questions in the post-search questionnaire for the *Baseline* system, as they refer to interface features that *Baseline* did not offer. Table III presents the mean average responses for each of these scales and differentials, using the labels after each of the first three Likert scales in the bulleted list above. The values for the three semantic differentials are included at the bottom of the table, as is their overall average opposite "*All {1,2,3}*".

Table III. Subject Perceptions of System Support (lower = better).

Scale / Differential	Fact-finding			Exploratory		
	QS	QD	SD	QS	QD	SD
Effectiveness	2.7	**2.5**	2.6	2.8	**2.3**	2.8
CloseToGoal	2.9	**2.7**	2.8	2.7	**2.2**	3.1
Re-use	2.9	3	**2.4**	**2.5**	**2.5**	3.2
1 Relevant	2.6	**2.5**	2.8	**2.4**	**2**	3.1
2 Useful	**2.6**	2.7	2.8	2.7	**2.1**	3.1
3 Appropriate	2.6	**2.4**	2.5	2.4	**2.4**	2.6
All {1,2,3}	**2.6**	**2.6**	**2.6**	2.6	**2.3**	2.9

The results show that all three experimental systems improved subjects' perceptions of their search effectiveness over *Baseline*, although only *QueryDestination* did so significantly.[15] Further examination of the effect size (measured using Cohen's $\underline{d}$) revealed that *QueryDestination* affects search effectiveness most positively.[16] *QueryDestination* also appears to get subjects closer to their information goal (CloseToGoal) than *QuerySuggestion* or *SessionDestination*, although only for exploratory search tasks.[17] Additional comments on *QuerySuggestion* imply that subjects saw it as a convenience (to save them typing a reformulation) rather than a way to dramatically influence search outcomes. For exploratory tasks, subjects felt that they benefited more from direction to alternative information sources than from suggestions for iterative refinements of their queries. Our findings also show that our subjects felt

[14] This question was conditioned on subjects' use of *Baseline* and their previous Web search experiences. That is, subject perceptions of their search effectiveness on this system compared to their opinion of their experiences on *Baseline* and other Web search engines they have used.

[15] $\underline{F}(3,136) = 4.07$, $\underline{p} = .008$; Tukey *post-hoc* tests: all $\underline{p} \leq .002$

[16] *QS*: $\underline{d}_{(\text{Fact-finding, Exploratory})} = (.26, .52)$; *QD:* $\underline{d}_{(\text{Fact-finding, Exploratory})} = (.77, 1.50)$; *SD:* $\underline{d}_{(\text{Fact-finding, Exploratory})} = (.48, .28)$

[17] $\underline{F}(2,102) = 5.00$, $\underline{p} = .009$; Tukey *post-hoc* tests: all $\underline{p} \leq .012$

that *QueryDestination* produced more "*relevant*" and "*useful*" suggestions for exploratory tasks than the other systems.[18] All other observed differences between the systems were not statistically significant.[19] The difference between performance of *QueryDestination* and *SessionDestination* can be explained by the approach used to generate destinations (described in Section 2). *SessionDestination*'s recommendations came from the end of users' sessions that often transcend multiple queries. This increases the likelihood that topic shifts adversely affect the relevance of proposed destinations.

4.1.3 *System Ranking.* In the final questionnaire that followed completion of all tasks on all systems, subjects were asked to rank the four systems in descending order based on their preferences. Table IV presents the mean average rank assigned to each of the systems.

Table IV. Relative Ranking of Systems (lower = better).

Systems	Baseline	QuerySuggest	QueryDest	SessionDest
Ranking	2.5	2.1	**1.9**	2.3

These results indicate that subjects preferred *QuerySuggestion* and *QueryDestination* overall. However, none of the differences between systems' ratings were significant.[20] One possible explanation for these systems being rated higher could be that although the popular destination systems performed well for exploratory tasks and *QuerySuggestion* performed well for fact-finding searches, an overall ranking merges these differences. This relative ranking reflects subjects' overall perceptions, but does not separate them for each task category. Over all tasks there appeared to be a slight preference for *QueryDestination*, but as other results show, the effect of task type on subjects' perceptions is significant.

4.1.4 *Subject Comments.* The final questionnaire included open-ended questions that asked subjects to explain their system ranking, and describe what they liked and disliked about each system:

- *Baseline:* Subjects who preferred *Baseline* commented on the familiarity of the system (e.g., "*was familiar and I didn't end up using suggestions*" (S36)). Those who did not prefer this system disliked the lack of support for query formulation ("*Can be difficult if you don't pick good search terms*" (S20)) and difficulty locating relevant documents (e.g., "*Difficult to find what I was looking for*" (S13); "*Clunky current technology*" (S30)).

- *QuerySuggestion:* Subjects who rated *QuerySuggestion* highest commented on rapid support for query formulation (e.g., "*was useful in (1) saving typing (2) coming up with new ideas for query expansion*" (S12); "*helps me better phrase the search term*" (S24); "*made my next query easier*" (S21)). Those who did not prefer this system criticized suggestion quality (e.g., "*Not relevant*" (S11); "*Popular queries weren't what I was looking for*" (S18)) and the quality of results they led to (e.g., "*Results (after clicking on suggestions) were of low quality*" (S35); "*Ultimately unhelpful*" (S1)).

[18] F(2,102) = 4.01, p = .01
[19] Tukey *post-hoc* tests: all p ≥ .143
[20] One-way repeated measures ANOVA: F(3,105) = 1.50, p = .22

- *QueryDestination:* Subjects who preferred this system commented mainly on support for accessing new information sources (e.g., "*provided potentially helpful and new areas / domains to look at*" (S27)) and bypassing the need to browse to these pages ("*Useful to try to 'cut to the chase' and go where others may have found answers to the topic*" (S3)). Those who did not prefer this system commented on the lack of specificity in the suggested domains ("*Should just link to site-specific query, not site itself*" (S16); "*Sites were not very specific*" (S24); "*Too general/vague*" (S28)[21]), and the quality of the suggestions ("*Not relevant*" (S11); "*Irrelevant*" (S6)).

- *SessionDestination:* Subjects who preferred this system commented on the utility of the suggested domains ("*suggestions make an awful lot of sense in providing search assistance, and seemed to help very nicely*" (S5)). However, more subjects commented on the irrelevance of the suggestions (e.g., "*did not seem reliable, not much help*" (S30); "*Irrelevant, not my style*" (S21), and the related need to include explanations about why the suggestions were offered (e.g., "*Low-quality results, not enough information presented*" (S35)).

These comments demonstrate a diverse range of perspectives on different aspects of the experimental systems. Further research is required to improve the quality of the suggestions in all systems, but subjects seemed to identify settings when each of these systems may be useful. Even though all systems can at times offer irrelevant suggestions, subjects appeared to prefer having them rather than not (e.g., one subject remarked "*suggestions were helpful in some cases and harmless in all*" (S15)).

4.2 Search Tasks

To gain a better understanding of how subjects performed during the study, we analyze data captured on their perceptions of task completeness and task completion time.

4.2.1 *Subject Perceptions.* In the post-search questionnaire, subjects were asked to indicate on a 5-point Likert scale the extent to which they agreed with the following attitude statement: "*I believe I have succeeded in my performance of this task*" (Success). In addition, they were asked to complete three 5-point semantic differentials indicating their response to the attitude statement: "*The task we asked you to perform was:*" The paired stimuli offered as possible responses were "*clear*"/"*unclear*", "*simple*"/"*complex*", and "*familiar*"/"*unfamiliar*". Table V presents the mean average response rating to these statements for each system and task type.

Table V. Subject Perceptions of Task and Task Success (lower = better).

Scale	Fact-finding				Exploratory			
	B	QS	QD	SD	B	QS	QD	SD
Success	2.0	**1.3**	1.4	1.4	2.8	2.3	**1.4**	2.6
Clear	1.2	**1.1**	**1.1**	**1.1**	1.6	**1.5**	**1.5**	1.6
Simple	1.9	**1.4**	1.8	1.8	2.4	2.9	**2.4**	3
Familiar	2.2	**1.9**	2.0	2.2	2.6	**2.5**	2.7	2.7

Subject responses demonstrate that users felt that their searches had been more successful using *QueryDestination* for exploratory tasks than with the other three systems (i.e., there

[21] Although the destination systems provided support for search within a domain, subjects mainly chose to ignore this.

was a two-way interaction between these two variables).[23] In addition, subjects perceived a significantly greater sense of completion with fact-finding tasks than with exploratory tasks.[25] Subjects also found fact-finding tasks to be more "*simple*", "*clear*", and "*familiar*" than the exploratory tasks. [27] The results also show that the subjects' perceptions of the clarity, complexity, and familiarity of tasks matched our goals when designing the tasks and the experiment. As illustrated by the examples in Figure 4, the fact-finding tasks required subjects to retrieve a finite set of answers (e.g., "*find three interesting things to do during a weekend visit to Kyoto, Japan*"). In contrast, the exploratory tasks were multi-faceted, and required subjects to find out more about a topic or to find sufficient information to make a decision. The end-point in such tasks was less fact-finding and may have affected subjects' perceptions of when they had completed the task. Given that there was no difference in the tasks attempted on each system, theoretically the perception of the tasks' simplicity, clarity, and familiarity should have been the same for all systems. However, we observe a clear interaction effect between the system and subjects' perception of the actual tasks.

4.2.2 *Task Completion Time*. In addition to asking subjects to indicate the extent to which they felt the task was completed, we also monitored the time that it took them to indicate to the experimenter that they had finished. The elapsed time from when the subject began issuing their first query until when they indicated that they were done (or the 10-minute time limit was reached) was monitored using a stopwatch and recorded for later analysis. Figure 5 shows the average task completion time for each system and each task type.

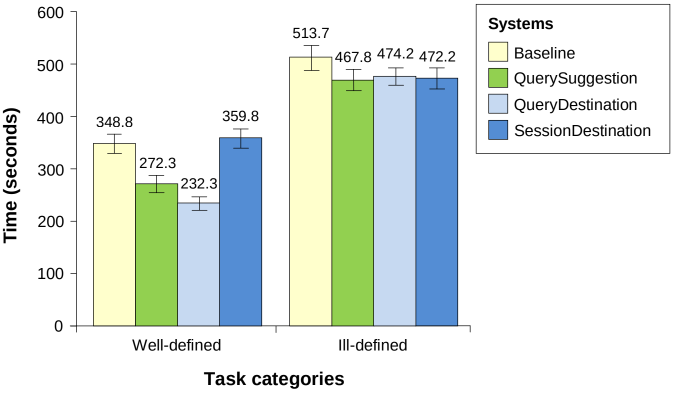


Fig. 5. Mean average task completion time (± Standard error of the mean).

As can be seen in the figure above, the task completion times for the fact-finding tasks differ greatly between systems.[28] Subjects attempting these tasks on *QueryDestination* and *QuerySuggestion* complete them in less time than subjects on *Baseline* and *SessionDestination*.[29] As discussed in the previous section, subjects were more familiar with the fact-finding tasks, and felt they were simpler and clearer. *Baseline* may have taken longer than the other systems since users had no additional support and had to

[23] F(3,136) = 6.34, p = .001

[25] F(1,136) = 18.95, p < .001

[27] F(1,136) = 6.82, p = .028

[28] F(3,136) = 4.56, p = .004

[29] Tukey *post-hoc* tests: all p ≤ .021

formulate their own queries. Subjects generally felt that the recommendations offered by *SessionDestination* were of low relevance and usefulness. Consequently, the completion time increased slightly between these two systems perhaps as the subjects assessed the value of the proposed suggestions, but reaped little benefit from them. The task completion times for the exploratory tasks were approximately equal on all four systems[30], although the time on *Baseline* was slightly higher. Since exploratory tasks had no clearly defined termination criteria other than the 10-minute time limit (i.e., the subject decided when they had gathered sufficient information), subjects generally spent longer searching, and consulted a broader range of information sources than for the fact-finding tasks.

4.3 Subject Interaction

We now focus on the observed interactions between subjects and systems. As well as eliciting feedback on each system from our subjects, we also recorded several aspects of their interaction with each system in log files. In this section, we analyze three aspects of their interaction: query iterations, search-result clicks, and subject engagement with the additional interface features offered by the three non-baseline systems.

4.3.1 *Queries and Result Clicks.* Searchers typically interact with search systems by submitting queries and clicking on search results. Therefore, we begin this section by analyzing querying and clickthrough behavior of our subjects to better understand how they conducted these core search activities, ignoring for the moment the additional interface features offered by our non-baseline systems. Table VI shows the average number of query iterations and search results clicked for each system-task pair.

Table VI. Mean Average Query Iterations and Result Clicks (per task).

Measure	Fact-finding				Exploratory			
	B	QS	QD	SD	B	QS	QD	SD
Queries	1.9	4.2	**1.5**	2.4	3.1	5.7	**2.7**	8.5
Result clicks	2.6	2	**1.7**	2.4	3.4	4.3	**2.3**	9.2

Subjects submitted fewer queries and clicked on fewer search results in *QueryDestination* than in any of the other systems for both fact-finding and exploratory tasks.[31] As discussed in the previous section, subjects using this system felt more successful in their exploratory searches yet they exhibited less of the query and result-click interactions required for search success on traditional search systems. An explanation for this – also validated in the next section – is that subjects interacted less with the system through queries and result clicks and elected to use the popular destinations instead. Across both task types, subjects issued the highest number of queries in *QuerySuggestion*, which is not surprising since this system actively encourages query refinement. To further investigate these, we look at the suggestion usage on the three non-baseline systems.

[30] $\underline{F}(3,136) = 1.06$, $\underline{p} = .37$

[31] *Queries*: $\underline{F}(3,443) = 3.99$, $\underline{p} = .008$; Tukey *post-hoc* tests: all $\underline{p} \leq .004$; *Result clicks*: $\underline{F}(3,431) = 3.63$, $\underline{p} = .013$; Tukey *post-hoc* tests: all $\underline{p} \leq .011$

4.3.2 *Suggestion Usage*. To determine whether subjects found suggestions useful, we measured the extent to which they were used when they were offered. Suggestion usage is defined as the proportion of submitted queries for which suggestions were offered and at least one suggestion was clicked. Table VII shows the average usage of suggestions for each system and task category, in terms of percentage of queries issued and percentage of experimental subjects that used them.

Table VII. Suggestion Usage.

Measure	Fact-finding			Exploratory		
	QS	QD	SD	QS	QD	SD
Percentage of queries	**35.7**	33.5	23.4	30.0	**35.2**	25.3
Percentage of subjects	**94.4**	83.3	72.2	88.9	**94.4**	88.9

The results presented in Table VII indicate that *QuerySuggestion* was used for more queries and by more subjects during fact-finding tasks than *SessionDestination*[32]; *QueryDestination* was used more than all other systems for the exploratory tasks.[34] Subjects used more destinations per query when using *QueryDestination* over *SessionDestination*.[35] As discussed earlier, these results may be explained by the lower perceived relevance and usefulness of destinations recommended by *SessionDestination*.

In the next section, we investigate change during the search session, focusing on the extent to which suggestion usage affects queries issued and domains visited.

4.3.4 *Changes Attributable to Suggestion Usage*. We envisaged that the introduction of suggestions would have a positive impact on user search interactions. To study the effect that usage of suggestions had on search behavior, we used the number of unique terms in query statements over the course of each task and the number of unique domains visited by subjects as a proxy for suggestion utility. A large number of unique query terms and/or unique domains associated with suggestion usage would imply that the systems were offering additional topic coverage to subjects.

Unique query terms: The functional objective of *QuerySuggestion* is to help users better define their information needs. Therefore, an increase in the number of unique query terms resulting from using *QuerySuggestion* may be indicative of system benefit. In contrast, the functional objective of *QueryDestination* and *SessionDestination* is to direct users to the most popular target domain (for fact-finding tasks) or broaden the set of domains they visit (for exploratory tasks). Unlike with *QuerySuggestion*, there is no direct association between the use of suggestions and query reformulation. Therefore, for us to regard the destination suggestion systems as responsible for an increase in the number of unique query terms, a previously unseen term must be present in the query iteration immediately following the use of destination suggestion.

Over the duration of each search task we monitored the total number of unique query terms issued by subjects and the usage of query and destination suggestions.[36] The

[32] $\underline{F}(2,355) \geq 4.67$, $\underline{p} \leq .01$; Tukey *post-hoc* tests: $\underline{p} \leq .006$

[34] Tukey *post-hoc* tests: all $\underline{p} \leq .027$

[35] *QD*: $\underline{M}_{\text{Fact-finding}} = 1.8$, $\underline{M}_{\text{Exploratory}} = 2.1$; *SD*: $\underline{M}_{\text{Fact-finding}} = 1.1$, $\underline{M}_{\text{Exploratory}} = 1.2$; $\underline{F}(1,231) = 5.49$, $\underline{p} = .02$; Tukey *post-hoc* tests: all $p \leq .003$.

[36] We defined usage of query suggestions as a click on the hyperlinked suggestion. We defined usage of destination suggestion as a click on a hyperlinked suggestion or a click on the "site search" option shown next to the suggestion and depicted with a magnifying glass.

number of unique terms used for each of the systems and each task type is shown in the first row of Table VIII. In addition, in the second row, we also show the percentage of unique query term and domain visitation increments that were attributable to usage of query or destination suggestions, as defined in the previous paragraph.

Table VIII. Mean Average Query Length and Query Overlap Measures.

Measure	Fact-finding				Exploratory			
	B	QS	QD	SD	B	QS	QD	SD
Unique query terms issued	5.2	**6.5**	4.3	6.1	7.4	7.8	6.5	**8.4**
% query changes attributable	–	42.2	23.1	24.6	–	27.3	27.2	25.3
Unique domains visited	1.3	1.3	1.5	**1.6**	2.1	2.3	2.7	**4.6**
% domain changes attributable	–	34.1	33.7	32.6	–	21.5	38.7	42.5

The findings show that *QuerySuggestion* increases the number of unique query terms more frequently in fact-finding tasks than *QueryDestination* and *SessionDestination*. Fror exploratory search tasks, all systems performed similarly.[37] It may be that for exploratory tasks our subjects could generate their own query refinements or, as we have conjectured already, the refinements offered by *QuerySuggestion* were less useful for such tasks (i.e., query suggestions only helped in refining the current need rather than supported exploration).

Rather than expanding users' query vocabulary, the functional objective of the destination systems was to facilitate rapid access to a broader range of authoritative sources. To account for this dimension, we also studied the number of unique domains that were visited by subjects during search tasks. We now describe the findings of that analysis.

Unique domains visited:[38] In the third and fourth rows of Table VIII we show the number of unique domains visited and proportion of unique domain visits attributable to the use of the query and destination suggestions. In order for *QuerySuggestion* to contribute to a visit to a previously unvisited domain, a subject must visit the domain during result browsing in the query iteration immediately following the use of query suggestion. The findings show that subjects attempting fact-finding tasks visited fewer unique domains on *QuerySuggestion* than either of destination suggestion systems.[39] We noted the same difference for exploratory tasks.[40]

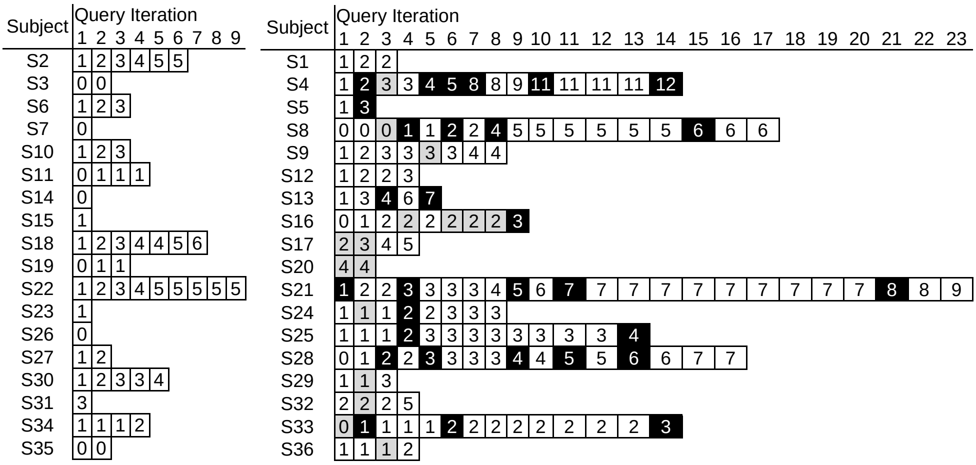
The analysis of usage data presented so far was based on aggregated statistics over all experimental subjects. However, when looking at changes in search behavior it is wise to also examine the behavior of individual users. Figure 6 depicts subject search patterns on *Baseline* and *SessionDestination* for exploratory search tasks, which were the systems/task types with the largest differences observed in querying and browsing. Each row represents a search session; each block represents a query iteration and contains a count of the number of unique domains visited in the session up until that iteration. For example, subject S5 issued two queries, visited one new domain on the first iteration and two new domains on the second iteration (for a total of three).

[37] $\underline{F}(3,136) = 3.93$; $\underline{p} = .02$; Tukey *post-hoc* tests: all $\underline{p} \leq .03$

[38] The nature of our logging meant that we could only log domains visited from the search result page, either through clicking on a search result or through selecting a destination suggestion.

[39] *Fact-finding*: $\underline{F}(3,51) = 1.48$, $\underline{p} = .23$

[40] *Exploratory*: $\underline{F}(3,51) = 4.19$, $\underline{p} = .01$, Tukey *post-hoc* tests: all $\underline{p} \leq .03$



Subject	Query Iteration 1	2	3	4	5	6	7	8	9
S2	1	2	3	4	5	5			
S3	0	0							
S6	1	2	3						
S7	0								
S10	1	2	3						
S11	0	1	1	1					
S14	0								
S15	1								
S18	1	2	3	4	4	5	6		
S19	0	1	1						
S22	1	2	3	4	5	5	5	5	5
S23	1								
S26	0								
S27	1	2							
S30	1	2	3	3	4				
S31	3								
S34	1	1	1	2					
S35	0	0							

Baseline

Subject	Query Iteration 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
S1	1	2	2																				
S4	1	2	3	3	4	5	8	8	9	11	11	11	11	12									
S5	1	3																					
S8	0	0	0	1	1	2	2	4	5	5	5	5	5	5	6	6	6						
S9	1	2	3	3	3	3	4	4															
S12	1	2	2	3																			
S13	1	3	4	6	7																		
S16	0	1	2	2	2	2	2	2	3														
S17	2	3	4	5																			
S20	4	4																					
S21	1	2	2	3	3	3	3	4	5	6	7	7	7	7	7	7	7	7	7	7	8	8	9
S24	1	1	1	2	2	3	3	3															
S25	1	1	1	2	3	3	3	3	3	3	3	3	4										
S28	0	1	2	2	3	3	3	3	4	4	5	5	6	6	7	7							
S29	1	1	3																				
S32	2	2	2	5																			
S33	0	1	1	1	1	2	2	2	2	2	2	2	2	3									
S36	1	1	1	2																			

SessionDestination

Fig. 6. Cumulative number of unique domains visited per search session. Successful destination suggestion interactions are highlighted in black, unsuccessful interactions are shown in gray.

We overlay usage information on the *SessionDestination* graphic to illustrate the role of the destination suggestions in surfacing new domains to users. Query iterations for the *SessionDestination* system that resulted in a visit to a previously unseen domain are marked in black (we infer that *SessionDestination* was successful in such cases). Usage of destination suggestions that did not lead to a new domain are marked in gray (in these cases, the site search option was selected or subjects revisited a domain using the suggestions they had already encountered in the session). Therefore, the total number of black and gray boxes equals the usage percentage for exploratory search tasks on *SessionDestination* shown in Table VII (i.e., 25.3%). Query iterations with zero unique domains (e.g., iterations 1, 2, and 3 for subject S8 on *SessionDestination*) occurred when users did not click on any search results. The differences in interaction between the two systems is striking, with subjects iterating more and visiting more unique domains on *SessionDestination*.

As can be seen from the subject listings in the figure our experimental design provided that the same subjects did not attempt an exploratory task on *Baseline* and an exploratory task on *SessionDestination*. However, since similar interaction patterns are observed across all subjects on both systems it seems that the differences are likely to be attributable to the search system rather than subject-specific searching strategies. On *SessionDestination* there were long periods where subjects refined their queries rather than visited new domains. It is interesting to note that the usage of the destination suggestions seems related to increases in the number of unique domains during the session. This implies that the suggestions are contributing towards subjects' exploration of the document space. For subjects S17, S20, S29, S32, and S36 the use of the destination suggestions did not lead to any visits to previously-unseen domains. Subjects S1 and S12 did not use the suggestions at all. While this fine-grained analysis is unsuitable to draw conclusions on usage trends given the size of our subject pool, it highlights the differences in uptake between subjects and gives an rough estimate of the proportion of new domain visits that were attributable to destination suggestion.

4.3.5 *Query Length and Query Overlap*. We continue looking at change within the search session by investigating differences in the queries submitted by subjects over the course of the search task. A better understanding of how queries evolve during search gives us insight into the role of each of the experimental systems in supporting query reformulation. In Table IX we present summary statistics on the average query length (in tokens) of all queries submitted (including those resulting from clicking on a query suggestion), and the differences in query length and overlap compared to: (i) the first query submitted for the search task, and (ii) the previous query in the session. Since term overlap, computed as the percentage of terms in a query that appear in another query, is not symmetrical (i.e., $Overlap(q_1,q_2)$ is not equal to $Overlap(q_2,q_1)$) we computed the overlap in both directions and present the mean average of the two values in Table IX.

Table IX. Mean Average Query Length and Query Overlap.

Measure	Fact-finding				Exploratory			
	B	QS	QD	SD	B	QS	QD	SD
Query length	3.0	3.0	3.3	**3.9**	2.9	3.1	3.3	**3.4**
% overlap with first query	31.5	**38.9**	29.6	30.9	26.7	**35.1**	30.2	30.3
% overlap with previous query	39.0	**43.5**	26.3	27.0	25.5	**46.3**	34.8	35.0

A two-way ANOVA revealed statistically significant differences in query length between systems and no difference between task types.[42] Further analysis of query length within each task type showed that queries on *SessionDestination* were significantly longer than *Baseline* and *QuerySuggestion* for fact-finding searches,[43] but no different from other systems for exploratory searches.[44] This may be indicative of searchers struggling to find relevant documents on *SessionDestination*, and issuing more precise query statements as a strategy to overcome that problem. Also interesting to note is that although not significantly different from the other systems, queries on *QuerySuggestion* appear to overlap most for fact-finding tasks, adding further support to our earlier claim that query suggestions were most useful for query refinement, and not for dramatically altering the search trajectory.[45] A similar trend is noticeable for overlap with the previous query rather than the base query (as shown in the last row of Table X). There is more overlap between consecutive queries with *QuerySuggestion* than with other systems (on both task types)[46]. This was to be expected given that suggestions generally contained the previous query or a prefix of the previous query as a substring, and these suggestions accounted for around 30-35% of the queries submitted.

5. DISCUSSION AND IMPLICATIONS

The study we have described in this article has shown that popular destinations can be a valuable resource for users engaged in search activities. Our findings show that subjects preferred *QuerySuggestion* for fact-finding tasks and *QueryDestination* for exploratory tasks. Analysis of subjects' perception of the search tasks and aspects of task completion showed that *QuerySuggestion* made subjects feel more successful for the fact-finding tasks. Conversely, *QueryDestination* led to heightened perceptions of search success for exploratory tasks. Query suggestions incrementally refine the original query, and

[42] *System*: $\underline{F}(3,475) = 4.48$, $\underline{p} = .004$; Tukey *post-hoc* tests: all $\underline{p} \leq .002$, *Task*: $\underline{F}(1,475) = .12$, $\underline{p} = .73$

[43] $\underline{F}(3,186) = 4.90$, $\underline{p} = .002$; Tukey *post-hoc* tests $_{\text{[SessionDest vs.[Baseline and QuerySuggest]]}}$: all $\underline{p} \leq .02$

[44] $\underline{F}(3,289) = 1.14$, $\underline{p} = .33$

[45] $\underline{F}(3,186) = 1.02$, $\underline{p} = .39$

[46] *Fact-finding*: $\underline{F}(3,186) = 2.87$, $\underline{p} = .03$; Tukey *post-hoc* tests $_{\text{[QuerySuggest vs. [QueryDest and SessionDest]]}}$: all $\underline{p} \leq .03$, *Exploratory*: $\underline{F}(3,289) = 3.85$, $\underline{p} = .01$; Tukey *post-hoc* tests: all $\underline{p} \leq .02$

therefore may be preferable for fact-finding tasks when users may have just missed their information target with their original query. However, when the task is more demanding, subjects valued destination suggestions, since these suggestions had the potential to dramatically influence the direction of a search.

Analysis of log interaction data gathered during the study indicates that although subjects submitted fewer queries and clicked fewer search results on *QueryDestination*, their engagement with suggestions was highest on this system, particularly for exploratory search tasks. Refined queries proposed by *QuerySuggestion* were used the most for the fact-finding tasks. *QuerySuggestion* led to the largest number of unique query terms issued and around 40% of new query terms came directly from the use of the query suggestions. The destination suggestion systems led to the greatest number of unique domains visited, especially for exploratory tasks, and analysis of usage statistics show that around 50% of all new domain visits were attributable to the use of the destination suggestions. There was more overlap with the initial query and the previous query with *QuerySuggestion*, implying that many of the suggestions it offered or refinements made by its users were extensions or specializations rather than dramatic changes. In previous work, it has been shown how the initial query in a search session is often used as a skeleton for refinement [White and Marchionini 2007]; it seems that *QuerySuggestion* encouraged this more than the other systems, perhaps to subjects' detriment in exploratory tasks. There appears to be a clear division between the systems: *QuerySuggestion* was preferred for fact-finding tasks, while *QueryDestination* provided most-used support for exploratory tasks. The success of popular destinations for exploratory tasks was promising given the challenge in supporting such complex activities.

The promising findings of our study suggest that systems offering popular destinations led to more successful and efficient searching compared to query suggestion and unaided Web search. Subjects seemed to prefer *QuerySuggestion* for the fact-finding tasks where the information-seeking goal was fact-finding. If the initial query does not retrieve relevant information, then subjects appreciate support in deciding what refinements to make to the query. From examination of the queries that subjects entered for the fact-finding searches across all systems, they appeared to use the initial query as a starting point, and add or subtract individual terms depending on search results. The post-search questionnaire asked subjects to select from a list of proposed explanations (or offer their own explanations) as to why they used recommended query refinements. For both fact-finding tasks and the exploratory tasks, around 40% of subjects indicated that they used a query suggestion because they "*wanted to save time typing a query*", while less than 10% of subjects did so because the suggestions "*represented new ideas*". Thus, subjects seemed to view *QuerySuggestion* as a time-saving convenience, rather than a way to dramatically impact search effectiveness.

The two variants of recommending destinations that we considered, *QueryDestination* and *SessionDestination*, offered domain suggestions that differed in their temporal proximity to the current query in previously observed user interactions. The quality of the destinations appeared to affect subjects' perceptions of them and their task performance. As discussed earlier, domains residing at the end of a complete search session (as in *SessionDestination*) are more likely to be unrelated to the current query, and thus are less likely to constitute valuable suggestions. Destination systems, in particular *QueryDestination*, performed best for the exploratory search tasks, where subjects may have benefited from exposure to additional information sources whose topical relevance to the search query is indirect. As with *QuerySuggestion*, subjects were asked to offer explanations for why they selected destinations. Over both task types they suggested that destinations were clicked because they "*grabbed their attention*" (40%), "*represented new ideas*" (25%), or users "*couldn't find what they were looking for*" (20%). The least popular responses were "*wanted to save time typing the address*" (7%) and "*the destination was popular*" (3%).

The positive response to destination suggestions from the study subjects provides interesting directions for design refinements. We were surprised to learn that subjects did not find the popularity bars useful, or hardly used the within-site search functionality, inviting re-design of these components. Subjects also remarked that they would like to see query-based summaries for each suggested destination to support more informed selection, as well as categorization of destinations with capability of drill-down for each category. Since *QuerySuggestion* and *QueryDestination* perform well in distinct task scenarios, integrating both in a single system is an interesting future direction. We hope to deploy some of these ideas on Web scale in future systems, which will allow log-based evaluation across large user pools.

It is clear that the use of query and session trails extracted from interaction logs has potential beyond providing suggestions for popular destinations following the submission of search queries. The trails can be a vehicle for better understanding search behavior [White and Drucker 2007], as a way to rank Web documents [Bilenko and White 2008], or as a way to estimate user satisfaction through patterns of interaction [Fox et al. 2005]. Although the focus in this article has been on the suggestion of popular *destinations* there are other potentially useful Web page types that might be useful to help Web searchers, particularly when the task is exploratory in nature. For example:

- <u>Interaction hubs:</u> Web pages or domains that other users *interact extensively with* following submission of a query, typically by viewing pages linked to by the hubs, and then returning to the hub and viewing more pages linked from it. Users obviously find some utility in such locations. In some respects this is similar to Kleinberg's notion of "hubs" in the HITS algorithm [Kleinberg 1998], although it is based on interaction log data rather than hyperlinks between Web pages created by page authors.

- <u>Waystations and portals:</u> Web pages or domains that other users *pass through* en route to other pages or domains. Whilst they may contain little or no relevant information they are often required to get to pages that contain such information.

Wexelblat and Maes [1999] also used navigation metaphors from the physical world (i.e., maps, paths, and signposts) in a similar way, but to describe the tools they have built rather than Web pages searchers utilize during information-seeking sessions. There is potential value in surfacing these additional types of sites as well as the destinations to support different types of information seeking activity. For example, interaction hubs may be shown for comparison shopping queries where a single, central domain or Web page is important in structuring user exploratory search behavior. As an alternative, frequently-visited links could be extracted from waystations or portal pages and offered to users as suggestions.

A limitation of this study relates to the amount of user interaction data available to us at the time the study was performed. Although *QueryDestination* was the most successful system during the experiment, this may be due to the fact because it had more trails for training than *SessionDestination.* We envisage that destinations may be even more valuable for searches with a known-target that is a significant number of clicks from the search result or even multiple queries away from the search result, e.g., users generally require more than one iteration to find relevant pages, and therefore it is possible that *SessionDestinations* variants would be successful at this. One way we can address this is by including more interaction log data that will improve coverage and give us more robust estimates on the value of a session-based destination relative to a query. Another possibility is to expand the destinations beyond domains and instead recommend particular URLs as candidate destinations. This was suggested by our subjects also, and thus, it seems like a natural enhancement to our approach given sufficient log data.

6. CONCLUSIONS

In this article, we presented a novel approach for enhancing users' Web search interaction by providing links to Web domains frequently visited by past searchers with similar information needs. So-called "popular destinations" lie at the end of many users' post-query browse trails, where information-seeking activity typically ceases once relevant information has been encountered. A user study was conducted in which we evaluated the effectiveness of the suggesting popular destinations compared with a query suggestion system and unaided Web search. Results of our study revealed that: (i) systems suggesting query refinements were preferred for fact-finding tasks, (ii) systems offering popular destinations were preferred for exploratory search tasks, and (iii) destinations should be mined from the end of query trails, not session trails. Overall, popular destination suggestions strategically influenced searches, including visits to more unique domains, in a way not achievable by query suggestion approaches, by offering a new way to resolve information problems, and enhance the information-seeking experience for many Web searchers. The promising results of employing popular destinations lead us to believe that there is value in utilizing other types of Web site contained in the search trails (e.g., interaction hubs, waystations) for search result ranking and user recommendation.

APPENDIX

Fact-finding task descriptions:

1. Identify three positive achievements of the Hubble telescope since its launch in 1991.
2. Find three hotels in Paris, France, that include a spa and health club.
3. Identify three interesting things to do during a weekend in Kyoto, Japan.
4. Find three categories of people that should not get a flu shot and why.
5. Identify three tropical storms (hurricanes and typhoons) that have caused property damage and/or loss of life.
6. Find three websites where you can buy soy milk online.

Exploratory task descriptions:

1. You have been talking to a friend about increases in size and diversity of the United States student population. You decide to find out how the student population has actually changed over the past five years.
2. A colleague has recently been diagnosed with a dust allergy. You are curious about causes of dust allergies and medications that ease the symptoms, so you decide to learn more about them.
3. You have to plan a five day vacation along the west coast of Italy. You want to find out what are the must-see sightseeing spots along the Italian west coast, and learn about Italian wine and the best vineyards in Tuscany to visit on your trip.

4. You are considering purchasing a Voice Over Internet Protocol (VoIP) telephone. You want to learn more about VoIP technology, providers that offer the service, and select the telephone and provider that best suits you.
5. You just read an article mentioning Internet music piracy. You become interested in the economics of the recording industry, and want to learn about recent performance of recording companies, losses due to piracy, and prospects for the music industry.
6. Your friend from Europe complains to you about the price of gasoline. You decide to research which costs contribute to the price of gasoline in the United States compared to Europe, and why prices seem to grow disproportionately to oil price fluctuations.

REFERENCES

AGICHTEIN, E., BRILL, E. AND DUMAIS, S. (2006). Improving Web search ranking by incorporating user behavior information. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 19-26.

AGICHTEIN, E., BRILL, E., DUMAIS, S. AND RAGNO, R. (2006). Learning user interaction models for predicting web search result preferences. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 3-10.

ANDERSON, C., DOMINGOS, P., and WELD, D.S. (2001). Adaptive Web navigation for wireless devices. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 879-884.

ANICK, P. (2003). Using terminological feedback for Web search refinement: A log-based study. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 88-95.

BEAULIEU, M. (1997). Experiments with interfaces to support query expansion. *Journal of Documentation, 53*, 1, 8-19.

BEEFERMAN, D. AND BERGER, A. (2000). Agglomerative clustering of a search engine query log. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 407-416.

BILENKO, M. AND WHITE, R.W. (2008). Mining the search trails of the surfing crowds: Identifying authoritative sources from user activity. In *Proceedings of the World Wide Web Conference*, to appear.

BORLUND, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation 56*, 1, 71-90.

BUSH, V. (1945). As we may think, *Atlantic Monthly*, 3, 2, 37-46.

CARD, S.K., PIROLLI, P., VAN DER WEGE, M., MORRISON, J., REEDER, R.W., SCHRAEDLY, P.K., AND BOSHART, J. (2001). Information scent as a driver of Web behavior graphs: results of a protocol analysis method for Web usability. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 498-505.

CATLEDGE, L.D. AND PITKOW, J.E. (1995). Characterizing browsing strategies in the world wide web. *Computer Networks and ISDN Systems*, 27, 6, 1065-1073.

CHALMERS, M., RODDEN, K., BRODBECK, D. (1998). The order of things: activity-centered information access. In *Proceedings of the World Wide Web Conference*, 359-367.

CRASWELL, N. AND SZUMMER, M. (2007). Random walks on the click graph. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 239-246.

CROFT, W.B. AND THOMPSON, R.H. (1987). I^3R: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science and Technology*, 38, 6, 389-404.

DOWNEY, D., DUMAIS, S. AND HORVITZ, E. (2007). Models of searching and browsing: languages, studies and applications. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1465-1472.

DUMAIS, S.T. AND BELKIN, N.J. (2005). *The TREC interactive tracks: putting the user into search*. In E.M. VOORHEES AND D.K. HARMAN, Eds. *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA: MIT Press, 123-153.

FOX, S., KARNAWAT, K., MYDLAND, M., DUMAIS, S. AND WHITE, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23, 2, 147-168.

FURNAS, G.W. (1985). Experience with an adaptive indexing scheme. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 131-135.

FURNAS, G.W., LANDAUER, T.K., GOMEZ, L.M. AND DUMAIS, S.T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30, 11, 964-971.

HICKL, A., WANG, P., LEHMANN, J. AND HARABAGIU, S. (2006). FERRET: Interactive question-answering for real-world environments. In *Proceedings of COLING/ACL*, 25-28.

JANSEN, B.J. AND SPINK, A. (2005). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42, 1, 248-263.

JOACHIMS, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 133-142.

JOACHIMS, T., GRANKA, L.A., PAN, B., HEMBROOKE, H. AND GAY, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 154-161.

JONES, R., REY, B., MADANI, O. AND GREINER, W. (2006). Generating query substitutions. In *Proceedings of the World Wide Web Conference*, 387-396.

KLEINBERG, J. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 668-677.

KOENEMANN, J. AND BELKIN, N. (1996). A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 205-212.

NEWELL, A. AND SIMON, H. (1972). *Human Problem Solving*. Prentice-Hall.

O'DAY, V. AND JEFFRIES, R. (1993). Orienteering in an information landscape: How information seekers get from here to there. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 438-445.

PIROLLI, P., PITKOW, J. AND RAO, R. (1996). Silk from a sow's ear: extracting usable structures from the Web. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 118-125.

PITKOW, J. AND PIROLLI, P. (1997). Life, death, and lawfulness on the electronic frontier. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 383-390.

PITKOW, J. AND PIROLLI, P. (1999). Mining longest repeating subsequences to predict World Wide Web surfing. In *Proceedings of the USENIX Symposium*, 139-150.

RADLINKSI, F. AND JOACHIMS, T. (2005). Query chains: Learning to rank from implicit feedback. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 239-248.

RESNICK, P., IACOVOU, N., SUCHAK, M., BERGSTROM, P. AND RIEDL, J. (2005). GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 175-186.

SALTON, G. AND BUCKLEY, C. (1988) Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*, 513-523.

SARWAR, B.M., KARYPIS, G., KONSTAN, J.A. AND RIEDL, J. (2000). Analysis of recommendation algorithms for e-commerce. In *Proceedings of the ACM Conference on Electronic Commerce*, 158-167.

SHARDANAND, U. AND MAES, P. (1995). Social information filtering: algorithms for automating word of mouth. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 210-217.

SILVERSTEIN, C., MARAIS, H., HENZINGER, M. AND MORICZ, M. (1999). Analysis of a very large Web search engine query log. *SIGIR Forum 33*, 1, 6-12.

SMUCKER, M. AND ALLAN, J. (2006). Find-similar: Similarity browsing as a search tool. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 461-468.

SMYTH, B., BALFE, E., FREYNE, J., BRIGGS, P., COYLE, M. and BOYDELL, O. (2004). Exploiting query repetition and regularity in an adaptive community-based Web search engine. *User Modelling and User Adapted Interaction 14*, 5, 382-423.

TEEVAN, J., ALVARADO, C., ACKERMAN, M.S. AND KARGER, D.R. (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 415-422.

WEINREICH, H., OBENDORF, H., HERDER, E. AND MAYER, M. (2006). Off the beaten tracks: Exploring three aspects of web navigation. In *Proceedings of the World Wide Web Conference*, 133-142.

WEXELBLAT, A. AND MAES, P. (1999). Footprints: history-rich tools for information foraging. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 270-277.

WHITE, R.W., RUTHVEN, I., JOSE, J.M. AND VAN RIJSBERGEN, C.J. (2005). Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems 23*, 3, 325-361.

WHITE, R.W. AND DRUCKER, S.M. (2007). Investigating behavioral variability in Web search. In *Proceedings of the World Wide Web Conference*, 21-30.

WHITE, R.W. AND MARCHIONINI, G. (2007). Examining the effectiveness of real-time query expansion. *Information Processing and Management 43*, 685-704.

WHITE, R.W., BILENKO, M. AND CUCERZAN, S. (2007). Studying the use of popular destinations to enhance Web search interaction. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 159-166.

XUE, G.-R., ZENG, H.-J., CHEN, Z., YU, Y., MA, W.Y., XI, W. AND FAN, W. (2004). Optimizing Web search using web click-through data. In *Proceedings of ACM CIKM Conference on Information and Knowledge Management*, 118-126.

ZHAI, C. AND LAFFERTY, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems 22*, 2, 179-214.