Studying the Use of Popular Destinations to Enhance Web Search Interaction

Ryen W. White
Microsoft Research
One Microsoft Way
Redmond, WA 98052
ryenw@microsoft.com

Mikhail Bilenko
Microsoft Research
One Microsoft Way
Redmond, WA 98052
mbilenko@microsoft.com

Silviu Cucerzan
Microsoft Research
One Microsoft Way
Redmond, WA 98052
silviu@microsoft.com

ABSTRACT

We present a novel Web search interaction feature which, for a given query, provides links to websites 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, whose collective behavior provides a basis for computing source authority. We describe a user study which 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 destination suggestions outperforms other systems for exploratory tasks, with best performance obtained from mining past user behavior at query-level granularity.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process.*

General Terms

Human Factors, Experimentation.

Keywords

User studies, search destinations, enhanced Web search.

1. INTRODUCTION

The problem of improving queries sent to Information Retrieval (IR) systems has been studied extensively in IR research [4][11]. 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, leading to improved retrieval performance. 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 [10].

Leveraging the decision-making processes of many users for query reformulation has its roots in *adaptive indexing* [8]. In recent years, applying such techniques has become possible at a much larger scale and in a different context than what was proposed in early work. However, interaction-based approaches to query suggestion 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. 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 user behavior to enhance users' Web search interactions.

Browser plugins and proxy server logs provide access to the browsing patterns of users that transcend search engine interactions. In previous work, such data have been used to improve search result ranking by Agichtein *et al.* [1]. However, this approach only considers page visitation statistics independently of each other, not taking into account the pages' relative positions on post-query browsing paths. Radlinski and Joachims [13] have utilized such collective user intelligence to improve retrieval accuracy by using sequences of consecutive query reformulations, yet their approach does not consider users' interactions beyond the search result page.

In this paper, we present a user study of a technique that exploits the searching and browsing behavior of many users to suggest popular Web pages, 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 pages at which other users end up frequently after submitting same or similar queries and then browsing away from initially clicked search results. We conjecture that destinations popular across a large number of users can capture the collective user experience for information needs, and our results support this hypothesis.

In prior work, O'Day and Jeffries [12] identified "teleportation" as an information-seeking strategy employed by users jumping to their previously-visited information targets, while Anderson *et al.* [2] applied similar principles to support the rapid navigation of Web sites on mobile devices. In [19], Wexelblat and Maes describe a system to support within-domain navigation based on the browse trails of other users. However, we are not aware of such principles being applied to Web search. Research in the area of recommender systems has also addressed similar issues, but in areas such as question-answering [9] and relatively small online communities [16]. Perhaps the nearest instantiation of teleportation is search engines' offering of several within-domain shortcuts below the title of a search result. While these may be based on user behavior and possibly site structure, the user saves at most one click from this feature. In contrast, our proposed approach can transport users to locations many clicks beyond the search result, saving time and giving them a broader perspective on the available related information.

SIGIR'07, July 23–27, 2007, Ámsterdam, The Netherlands.
Copyright 2007 ACM 978-1-59593-597-7/07/0007...$5.00.

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 known-item 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 known-item tasks.

The remainder of the paper is structured 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, while Sections 4 and 5 present the study findings and their discussion, respectively. We conclude in Section 6 with a summary.

2. SEARCH TRAILS AND DESTINATIONS

We used Web activity logs containing searching and browsing activity collected with permission from hundreds of thousands of 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". In this section, we summarize the extraction of trails, their features, and destinations (trail end-points). In-depth description and analysis of trail extraction are presented in [20].

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!, Windows Live Search, and Ask. It is these search trails that we use to identify popular destinations.

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. Extracting search trails using this methodology also goes some way toward handling multi-tasking, where users run multiple searches concurrently. Since users may open a new browser window (or tab) for each task [18], each task has its own browser trail, and a corresponding distinct search trail.

To reduce the amount of "noise" from pages unrelated to the active search task that may pollute our data, search trails are terminated when one of the following events occurs: (1) a user returns to their homepage, checks e-mail, logs in to an online service (e.g., MySpace or del.ico.us), types a URL or visits a bookmarked page; (2) a page is viewed for more than 30 minutes with no activity; (3) the user closes the active browser window. If a page (at step i) meets any of these criteria, the trail is assumed to terminate on the previous page (i.e., step $i - 1$).

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 three 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.

Approximately 14 million query trails and 4 million session trails were extracted from the logs. We now describe some trail features.

2.2 Trail and Destination Analysis

Table 1 presents summary statistics for the query and session trails. Differences in user interaction between the last domain 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 averages for all trails with two or more steps (i.e., those trails where at least one search result was clicked).

Table 1. Summary statistics (mean averages) for search trails.

Measure		Query trails	Session trails
Number of unique domains		2.0	4.3
Total page views	All domains	4.8	16.2
	Domains 1 to (n – 1)	1.4	10.1
	Domain n (destination)	3.4	6.2
Total time spent (secs)	All domains	172.6	621.8
	Domains 1 to (n – 1)	70.4	397.6
	Domain n (destination)	102.3	224.1

The statistics suggest that users generally browse far from the search results page (i.e., around 5 steps), and visit a range of domains during the course of their search. On average, users visit 2 unique (non search-engine) domains per query trail, and just over 4 unique domains per session trail. This suggests that users often do not find all the information they seek on the first domain they visit. For query trails, users also visit more pages, and spend significantly longer, on the last domain in the trail compared to all previous domains combined.[1] These distinctions of the last domains in the trails may indicate user interest, page utility, or page relevance.[2]

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, and 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 [15,17]). Therefore, a lookup-based approach would prevent us from reliably suggesting destinations for a large fraction of searches. To overcome this problem, we utilize a simple term-based prediction model.

As discussed above, we extract two types of destinations: query destinations and session destinations. For both destination types, we obtain a corpus of query-destination pairs and use it to construct term-vector representation of destinations that is analogous to the classic *tf.idf* document representation in traditional IR [14].

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:

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

[2] The topical relevance of the destinations was tested 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]*).

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

Where query and destination term weights, $w_q(t_i)$ and $w_d(t_i)$, are computed using standard *tf.idf* weighting and query- and user-session-normalized smoothed *tf.idf* weighting, respectively. While exploring alternative algorithms for the destination prediction task remains an interesting challenge for future work, results of the user study described in subsequent sections demonstrate that this simple approach provides robust, effective results.

3. 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.

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 method 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*).

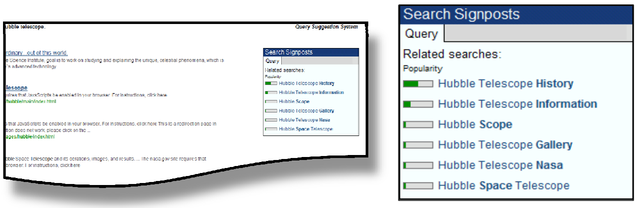
3.1.1 System 1: Baseline

To establish baseline performance against which other systems can be compared, we developed a masked interface to a popular search engine without additional support in formulating queries. This system presented the user-constructed query to the search engine and returned ten top-ranking documents retrieved by the engine. To remove potential bias that may have been caused by subjects' prior perceptions, we removed all identifying information such as search engine logos and distinguishing interface features.

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 the search engine query log over the timeframe 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 100 most frequent such queries, while the second set contains 100 most frequent queries that *followed* the target query in query logs. Each candidate query 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, six top-ranked query suggestions are returned. If fewer than six suggestions are found, iterative backoff is performed using progressively longer suffixes of the target query; a similar strategy is described in [10].

Suggestions were offered in a box positioned on the top-right of the result page, adjacent to the search results. Figure 1a shows the position of the suggestions on the page. Figure 1b shows a zoomed view of the portion of 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 suggestion retrieves new search results for that query.

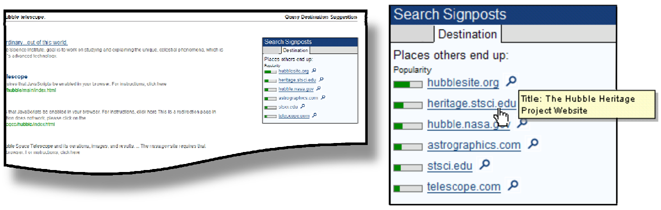


(a) Position of suggestions **(b) Zoomed suggestions**

Figure 1. Query suggestion presentation in *QuerySuggestion*.

3.1.3 System 3: QueryDestination

QueryDestination uses an interface similar to *QuerySuggestion*. However, instead of showing query refinements for the submitted query, *QueryDestination* suggests up to six destinations frequently visited by other users who submitted queries similar to the current one, and computed as described in the previous section.[3] Figure 2a shows the position of the destination suggestions on search results page. Figure 2b shows a zoomed view of the portion of the results page destinations suggested for the query *[hubble telescope]*.



(a) Position of destinations **(b) Zoomed destinations**

Figure 2. Destination presentation in *QueryDestination*.

To keep the interface uncluttered, the page title of each destination is shown on hover over the page URL (shown in Figure 2b). 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. 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., those focusing on related topics or not containing the original query terms).

3.1.4 System 4: SessionDestination

The 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 top destinations. *QueryDestination* directs users to the domains others end up at *for the active or similar 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., we only care 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:

[3] To improve reliability, in a similar way to *QuerySuggestion*, destinations are only shown if their popularity exceeds a frequency threshold.

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

a. Searches that are well-defined ("known-item" tasks)?
b. Searches that are ill-defined ("exploratory" tasks)?

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 our organization where they hold a range of positions in different divisions. The average age of subjects was 34.9 years (max=62, min=27, 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 [4], we made task type an independent variable in the study. We constructed six known-item tasks and six open-ended, exploratory tasks that were rotated between systems and subjects as described in the next section. Figure 3 shows examples of the two task types.

Known-item 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.

Figure 3. Examples of known-item and exploratory tasks.

Exploratory tasks were phrased as simulated work task situations [5], 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. The known-item search tasks required search for particular items of information (e.g., activities, discoveries, names) for which the target was well-defined. A similar task classification has been used successfully in previous work [21]. Tasks were taken and adapted from the Text Retrieval Conference (TREC) Interactive Track [7], and questions posed on question-answering communities (Yahoo! Answers, Google Answers, and Windows Live QnA). To motivate the subjects during their searches, we allowed them to select two known-item 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). 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 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 counterbalanced according to a Graeco-Latin square design.

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

1. Upon arrival, subjects were asked to select two known-item and two exploratory tasks 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.
   c. Upon completion of the task, subjects were asked to complete a post-search questionnaire.
5. After completing the tasks on the four systems, subjects answered a final questionnaire comparing their experiences on the systems.
6. Subjects were thanked and compensated.

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

4. FINDINGS

In this section we use the data derived from the experiment to address our hypotheses about query suggestions and destinations, providing information on the effect of task type and topic familiarity where appropriate. Parametric statistical testing is used in this analysis and the level of significance is set to $p < 0.05$, 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.

4.1 Subject Perceptions

In this section we present findings on how subjects perceived the systems that they used. Responses to post-search (per-system) and final questionnaires are used as the basis for our analysis.

4.1.1 Search Process

To address the first research question wanted 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 average obtained differential values are shown in Table 1 for each system and each task type. The value corresponding to the differential "*All*" represents the mean of all three differentials, providing an overall measure of subjects' feelings.

Table 1. Perceptions of search process (lower = better).

Differential	Known-item				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 1 summarizes subject responses for 18 task-system pairs (18 subjects who ran a known-item 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 the other systems for known-item tasks.[4] For exploratory tasks, only searches conducted on *QueryDestination* were easier than on the other systems.[5] Subjects indicated that exploratory tasks on the three non-baseline systems were more stressful (i.e., less "*relaxing*") than the known-item tasks.[6] As we will discuss in more detail in Section 4.1.3, subjects regarded the familiarity of *Baseline* as a strength, and may have struggled to attempt a more complex task while learning a new interface feature such as 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 Likert scales and semantic differentials were used:

- Likert scale A: *"Using this system enhances my effectiveness in finding relevant information."* (Effectiveness)[7]
- 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 in the post-search questionnaire when subjects used the *Baseline* system as they refer to interface support options that *Baseline* did not offer. Table 2 presents the 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 under "*All*".

Table 2. Perceptions of system support (lower = better).

Scale /	Known-item			Exploratory		
Differential	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.[8] Further examination of the effect size (measured using Cohen's d) revealed that *QueryDestination* affects search effectiveness most positively.[9] *QueryDestination* also appears to get subjects closer to their information goal (CloseToGoal) than *QuerySuggestion* or *SessionDestination*, although only for exploratory search tasks.[10] Additional comments on *QuerySuggestion* conveyed that subjects saw it as a convenience (to save them typing a reformulation) rather than a way to dramatically influence the outcome of their search. For exploratory searches, users benefited more from being pointed to alternative information sources than from suggestions for iterative refinements of their queries. Our findings also show that our subjects felt that *QueryDestination* produced more "*relevant*" and "*useful*" suggestions for exploratory tasks than the other systems.[11] All other observed differences between the systems were not statistically significant.[12] The difference between performance of *QueryDestination* and *SessionDestination* is explained by the approach used to generate destinations (described in Section 2). *SessionDestination*'s recommendations came from the end of users' session trails that often transcend multiple queries. This increases the likelihood that topic shifts adversely affect their relevance.

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 3 presents the mean average rank assigned to each of the systems.

Table 3. Relative ranking of systems (lower = better).

Systems	Baseline	QSuggest	QDest	SDest
Ranking	2.47	2.14	1.92	2.31

These results indicate that subjects preferred *QuerySuggestion* and *QueryDestination* overall. However, none of the differences between systems' ratings are significant.[13] One possible explanation for these systems being rated higher could be that although the popular destination systems performed well for exploratory searches while *QuerySuggestion* performed well for known-item searches, an overall ranking merges these two performances. 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.

The final questionnaire also 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*

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

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

[6] *relaxing*: $\underline{F}(1,136) = 6.47$, $\underline{p} = .011$

[7] This question was conditioned on subjects' use of *Baseline* and their previous Web search experiences.

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

[9] *QS*: $\underline{d}_{(K,E)} = (.26, .52)$; *QD:* $\underline{d}_{(K,E)} = (.77, 1.50)$; *SD:* $\underline{d}_{(K,E)} = (.48, .28)$

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

[11] $\underline{F}(2,102) = 4.01$, $\underline{p} = .01$; $\alpha = .0167$

[12] Tukey *post-hoc* tests: all $\underline{p} \geq .143$

[13] One-way repeated measures ANOVA: $\underline{F}(3,105) = 1.50$, $\underline{p} = .22$

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)).

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)[14]), 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. Work is obviously needed in improving the quality of the suggestions in all systems, but subjects seemed to distinguish the 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)).

4.1.4 Summary

The findings obtained from our study on subjects' perceptions of the four systems indicate that subjects tend to prefer *QueryDestination* for the exploratory tasks and *QuerySuggestion* for the known-item searches. Suggestions to incrementally refine the current query may be preferred by searchers on known-item tasks when they may have just missed their information target. However, when the task is more demanding, searchers appreciate suggestions that have the potential to dramatically influence the direction of a search or greatly improve topic coverage.

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 the time that it took them to complete each task.

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 4 presents the mean average response to these statements for each system and task type.

Table 4. Perceptions of task and task success (lower = better).

Scale	Known-item				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
1 Clear	1.2	**1.1**	**1.1**	**1.1**	1.6	**1.5**	**1.5**	1.6
2 Simple	1.9	**1.4**	1.8	1.8	2.4	2.9	**2.4**	3
3 Familiar	2.2	**1.9**	2.0	2.2	2.6	**2.5**	2.7	2.7
All {1,2,3}	1.8	**1.4**	1.6	1.8	**2.2**	**2.2**	**2.2**	2.3

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 was a two-way interaction between these two variables).[15] In addition, subjects perceived a significantly greater sense of completion with known-item tasks than with exploratory tasks.[16] Subjects also found known-item tasks to be more "*simple*", "*clear*", and "*familiar*". [17] These responses confirm differences in the nature of the tasks we had envisaged when planning the study. As illustrated by the examples in Figure 3, the known-item 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 well-defined 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 was monitored using a stopwatch and recorded for later analysis. A stopwatch rather than system logging was used for this since we wanted to record the time regardless of system interactions. Figure 4 shows the average task completion time for each system and each task type.

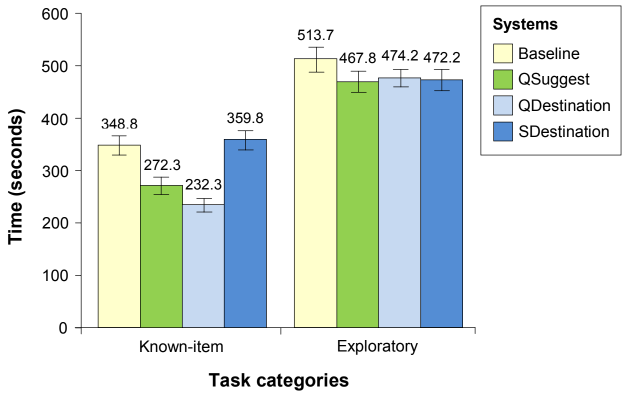


Figure 4. Mean average task completion time (± SEM).

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

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

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

[17] F(1,136) = 6.82, p = .028; Known-item tasks were also more "*simple*" on *QS* (F(3,136) = 3.93, p = .01; Tukey *post-hoc* test: p = .01); α = .167

As can be seen in the figure above, the task completion times for the known-item tasks differ greatly between systems.[18] Subjects attempting these tasks on *QueryDestination* and *QuerySuggestion* complete them in less time than subjects on *Baseline* and *SessionDestination*.[19] As discussed in the previous section, subjects were more familiar with the known-item 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 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[20], although the time on *Baseline* was slightly higher. Since these tasks had no clearly defined termination criteria (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 in the known-item tasks.

4.2.3 Summary

Analysis of subjects' perception of the search tasks and aspects of task completion shows that the *QuerySuggestion* system made subjects feel more successful (and the task more "*simple*", "*clear*", and "*familiar*") for the known-item tasks. On the other hand, *QueryDestination* was shown to lead to heightened perceptions of search success and task ease, clarity, and familiarity for the exploratory tasks. Task completion times on both systems were significantly lower than on the other systems for known-item tasks.

4.3 Subject Interaction

We now focus our analysis on the observed interactions between searchers 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 interaction aspects: 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. Although our system offers additional interface affordances, we begin this section by analyzing querying and clickthrough behavior of our subjects to better understand how they conducted core search activities. Table 5 shows the average number of query iterations and search results clicked for each system-task pair. The average value in each cell is computed for 18 subjects on each task type and system.

Table 5. Average query iterations and result clicks (per task).

Scale	Known-item				Exploratory			
	B	QS	QD	SD	B	QS	QD	SD
Queries	1.9	4.2	**1.5**	2.4	3.1	5.7	**2.7**	3.5
Result clicks	2.6	2	**1.7**	2.4	3.4	4.3	**2.3**	5.1

Subjects submitted fewer queries and clicked on fewer search results in *QueryDestination* than in any of the other systems.[21] As discussed in the previous section, subjects using this system felt more successful in their searches yet they exhibited less of the traditional query and result-click interactions required for search success on traditional search systems. It may be the case that subjects' queries on this system were more effective, but it is more likely that they interacted less with the system through these means and elected to use the popular destinations instead. Overall, subjects submitted most queries in *QuerySuggestion*, which is not surprising as this system actively encourages searchers to iteratively re-submit refined queries. Subjects interacted similarly with *Baseline* and *SessionDestination* systems, perhaps due to the low quality of the popular destinations in the latter. To investigate this and related issues, we will next analyze usage of the suggestions on the three non-baseline systems.

4.3.2 Suggestion Usage

To determine whether subjects found additional features useful, we measure the extent to which they were used when they were provided. Suggestion usage is defined as the proportion of submitted queries for which suggestions were offered and at least one suggestion was clicked. Table 6 shows the average usage for each system and task category.

Table 6. Suggestion uptake (values are percentages).

Measure	Known-item			Exploratory		
	QS	QD	SD	QS	QD	SD
Usage	**35.7**	33.5	23.4	30.0	**35.2**	25.3

Results indicate that *QuerySuggestion* was used more for known-item tasks than *SessionDestination*[22], and *QueryDestination* was used more than all other systems for the exploratory tasks.[23] For well-specified targets in known-item search, subjects appeared to use query refinement most heavily. In contrast, when subjects were exploring, they seemed to benefit most from the recommendation of additional information sources. Subjects selected almost twice as many destinations per query when using *QueryDestination* compared to *SessionDestination*.[24] As discussed earlier, this may be explained by the lower perceived relevance and usefulness of destinations recommended by *SessionDestination*.

4.3.3 Summary

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. The refined queries proposed by *QuerySuggestion* were used the most for the known-item tasks. There appears to be a clear division between the systems: *QuerySuggestion* was preferred for known-item tasks, while *QueryDestination* provided most-used support for exploratory tasks.

5. DISCUSSION AND IMPLICATIONS

The promising findings of our study suggest that systems offering popular destinations lead to more successful and efficient searching compared to query suggestion and unaided Web search.

Subjects seemed to prefer *QuerySuggestion* for the known-item tasks where the information-seeking goal was well-defined. If the initial query does not retrieve relevant information, then subjects

[18] $\underline{F}(3,136) = 4.56$, $\underline{p} = .004$

[19] Tukey *post-hoc* tests: all $\underline{p} \leq .021$

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

[21] *Queries*: F(3,443) = 3.99; p = .008; Tukey *post-hoc* tests: all $p \leq .004$; *Systems*: F(3,431) = 3.63, p = .013; Tukey *post-hoc* tests: all $p \leq .011$

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

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

[24] *QD*: $\underline{M}_K = 1.8$, $\underline{M}_E = 2.1$; *SD*: $\underline{M}_K = 1.1$, $\underline{M}_E = 1.2$; $\underline{F}(1,231) = 5.49$, $\underline{p} = .02$; Tukey *post-hoc* tests: all $p \leq .003$; ($\underline{M}$ represents mean average).

appreciate support in deciding what refinements to make to the query. From examination of the queries that subjects entered for the known-item 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 known-item tasks and the exploratory tasks, around 40% of subjects indicated that they selected 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 suggestions that differed in their temporal proximity to the current query. 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.

6. CONCLUSIONS

We presented a novel approach for enhancing users' Web search interaction by providing links to websites frequently visited by past searchers with similar information needs. A user study was conducted in which we evaluated the effectiveness of the proposed technique compared with a query refinement system and unaided Web search. Results of our study revealed that: (i) systems suggesting query refinements were preferred for known-item 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 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.

7. REFERENCES

[1] Agichtein, E., Brill, E. & Dumais, S. (2006). Improving Web search ranking by incorporating user behavior information. In *Proc. SIGIR*, 19-26.

[2] Anderson, C. *et al.* (2001). Adaptive Web navigation for wireless devices. In *Proc. IJCAI*, 879-884.

[3] Anick, P. (2003). Using terminological feedback for Web search refinement: A log-based study. In *Proc. SIGIR*, 88-95.

[4] Beaulieu, M. (1997). Experiments with interfaces to support query expansion. *J. Doc. 53*, 1, 8-19.

[5] Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems*. J. Doc. 56*, 1, 71-90.

[6] Downey *et al*. (2007). Models of searching and browsing: languages, studies and applications. In *Proc. IJCAI*, 1465-72.

[7] Dumais, S.T. & Belkin, N.J. (2005). *The TREC interactive tracks: putting the user into search*. In Voorhees, E.M. and Harman, D.K. (eds.) *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA: MIT Press, 123-153.

[8] Furnas, G. W. (1985). Experience with an adaptive indexing scheme. In *Proc. CHI*, 131-135.

[9] Hickl, A. *et al.* (2006). FERRET: Interactive question-answering for real-world environments. In *Proc. of COLING/ACL*, 25-28.

[10] Jones, R., *et al.* (2006). Generating query substitutions. In *Proc. WWW*, 387-396.

[11] Koenemann, J. & Belkin, N. (1996). A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *Proc. CHI*, 205-212.

[12] O'Day, V. & Jeffries, R. (1993). Orienteering in an information landscape: how information seekers get from here to there. In *Proc. CHI*, 438-445.

[13] Radlinski, F. & Joachims, T. (2005). Query chains: Learning to rank from implicit feedback. In *Proc. KDD*, 239-248.

[14] Salton, G. & Buckley, C. (1988) Term-weighting approaches in automatic text retrieval. *Inf. Proc. Manage. 24*, 513-523.

[15] Silverstein, C. *et al.* (1999). Analysis of a very large Web search engine query log. *SIGIR Forum 33*, 1, 6-12.

[16] Smyth, B. *et al*. (2004). Exploiting query repetition and regularity in an adaptive community-based Web search engine. *User Mod. User Adapt. Int. 14*, 5, 382-423.

[17] Spink, A. *et al.* (2002). U.S. versus European Web searching trends. *SIGIR Forum 36*, 2, 32-38.

[18] Spink, A., *et al.* (2006). Multitasking during Web search sessions. *Inf. Proc. Manage.*, *42*, 1, 264-275.

[19] Wexelblat, A. & Maes, P. (1999). Footprints: history-rich tools for information foraging. In *Proc. CHI*, 270-277.

[20] White, R.W. & Drucker, S.M. (2007). Investigating behavioral variability in Web search. In *Proc. WWW*, 21-30.

[21] White, R.W. & Marchionini, G. (2007). Examining the effectiveness of real-time query expansion. *Inf. Proc. Manage. 43*, 685-704.