Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs

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

Jeff Huang
University of Washington
Seattle, WA 98195 USA
sigir@jeffhuang.com

ABSTRACT

Search trails mined from browser or toolbar logs comprise queries and the post-query pages that users visit. Implicit endorsements from many trails can be useful for search result ranking, where the presence of a page on a trail increases its query relevance. Following a search trail requires user effort, yet little is known about the benefit that users obtain from this activity versus, say, sticking with the clicked search result or jumping directly to the destination page at the end of the trail. In this paper, we present a log-based study estimating the user value of trail following. We compare the relevance, topic coverage, topic diversity, novelty, and utility of full trails over that provided by sub-trails, trail origins (landing pages), and trail destinations (pages where trails end). Our findings demonstrate significant value to users in following trails, especially for certain query types. The findings have implications for the design of search systems, including trail recommendation systems that display trails on search result pages.

Categories and Subject Descriptors

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

General Terms

Experimentation, Human Factors, Measurement

Keywords

Search trails, trail following, log analysis

1. INTRODUCTION

Web search engines afford keyword access to Web content. In response to search queries, these engines return lists of Web pages ranked based on their predicted relevance. For decades, the information retrieval (IR) research community has worked extensively on algorithmic techniques to effectively rank documents (c.f. [22]). However, research in areas such as information foraging [18], berrypicking [2], and orienteering [17], suggests that individual items may be insufficient for vague or complex information needs. In such circumstances, search results may only serve as the starting points for exploration [24].

Search trails are a series of Web pages starting with a search query and terminating with an event such as session inactivity [33]. Although the traversal of trails following a query is common, little is known about how much value users derive from following the trail versus sticking with the origin (the clicked search result) or jumping to the destination page at the end of the trail [32]. In this paper we present a log-based study estimating the value to users of traversing multi-page search trails. Our primary aim is to estimate the benefit that *trail following* brings to users under different metrics versus viewing only the origin and/or destination pages. Significant differences in the performance of trails over origins and destinations would suggest that users benefit from the journey as well as the origin and the destination. Knowing if and when this is the case could help us build more effective search systems centered around trails, e.g., full trails could be shown to users directly on the results page. We estimate the value of trails, sub-trails, origins, and destinations (collectively called trail *sources*) based on the relevance, completeness, diversity, novelty, and utility of the information they contain. We conduct this study using a log-based methodology since logs contain evidence of real user behaviors at scale and provide coverage of many types of information needs. Information need coverage is important since differences in source performance may not hold for all search tasks.

The remainder of this paper is structured as follows. Section 2 presents related work on trails. Section 3 describes the primary data source used in our study, as well as the extraction and labeling of search trails, and trail statistics. Section 4 describes the experiment performed to estimate the value of trails or sub-trails, including a comparison with trail origins and trail destinations. Section 5 describes the findings of our study for all queries and different query types. Findings are discussed along with their implications in Section 6. We conclude in Section 7.

2. RELATED WORK

Vannevar Bush first introduced the concept of trails when he envisioned the *memex*, a theoretical proto-hypertext system to extend human memory [4]. Bush foresaw "a new profession of trail blazers, those who find delight in the task of establishing useful trails through the enormous mass of the common record." Associative trails explicitly created by trailblazing users form links between stored materials that can help others navigate. Interaction logging via browsers and toolbars has made us all (implicit) trail blazers.

A search trail consists of an origin page, intermediate pages, and a destination page. Origin pages are the search results that start a trail, and may be referred to as landing pages in other literature. The use of query and origin pages from search engine click logs has been shown to be useful for improving search result relevance [1][14]. Teevan et al. [24] studied users jumping directly to destination pages and introduced the concept of *teleportation* when they observed users issuing sophisticated queries in an attempt to navigate to a page they knew existed deep in a Web site. White et al. [32] incorporated destination pages corresponding to Web search queries into search interface prototypes and presented them to user study participants. Most users found destination pages useful when shown on the search results page after the query was submitted. Bilenko and White [3] studied full trails, including the origin, intermediate, and destination pages. They found that treating the pages in these trails as endorsements improved ranking in

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

search engines. Individual pages in full trails have been shown to improve search results, destination pages have been shown to benefit users, and origin pages have been studied extensively in search relevance. We are the first to study the value of trails to users and directly compare trails to origins and destinations.

Trails have been studied in domains outside of IR. Wexelblat and Maes [29] introduced annotations in Web browsers called "footprints," which are trails through a Website assembled by the Website designer. Their evaluation found that users required significantly less steps to find information using their system. Freyne et al. [12] add a second dimension to footprints by displaying icons with links to offer visual cues to the user. These cues are gathered from past users and include popularity, recency, and user-generated annotations. More recent work by Wang and Zhai [28] continues the footprint metaphor in a topic map. This topic map allows the user to navigate horizontally to related queries, and vertically to queries of different specificity. Simulated users with a predefined strategy benefited from such maps. Pirolli and Card [18] developed a sophisticated model of user behavior called *information foraging* derived from how animals forage for food in the wild. They use a foraging metaphor to discuss how information foragers could use cues left by previous visitors to find "patches" of information in a collection and consume patch information to satisfy information needs. Fu and Pirolli [13] developed and validated computational cognitive models of Web navigation behavior based on information foraging theory.

ScentTrails [16] combines browsing and searching into a single interface by highlighting potentially valuable hyperlinks. Olston and Chi perform user studies with different interfaces incorporating "scents" of trails in the search results. Users could find information faster and more successfully using *ScentTrails* than by either searching or browsing alone. O'Day and Jeffries [17] propose the orienteering analogy for understanding users' information-seeking strategies. Their qualitative study relates to ours in describing the benefits of building a system that considers the entirety of users' paths. Similarly, Bates's *berrypicking* [2] discusses users moving between information sources due to dynamic information needs. Search trails are extensions of these ideas into Web search, showing the routes with information to harvest, and orienting them towards the winding paths others have taken. As with orienteering and berrypicking, the origin and destination are important but the route taken in-between is also important; in this study we estimate how much benefit users gain from this journey.

Trigg [26] introduced the concept of *guided tours*, whereby authors could construct sequences of pages that may be useful to others. Reich et al. [19] discuss tours and trails as tools for helping hypertext users by showing where others have gone. Tours and trails in hypertext differ; trails are marked by users at each step while tours are typically authored beforehand and may have a hierarchical structure. Reich et al. also propose following users with similar interests as they move around the collection. Beyond hypertext, Chalmers et al. [6] present a system where people who are "recommenders" manually construct Web navigation paths. These recommenders share their paths with others. Wheeldon and Levene [30] propose an algorithm for generating trails to assist in Web navigation. Trails are presented in a tree interface attached to the browser. User study participants expressed satisfaction with the trails, noting that seeing the relationship between links helped, and found trails to be useful as a navigational aid.

The study described in this paper differs from previous work in that *we are focused on estimating the value that trail following brings to users*, rather than describing existing trail traversal behavior, modeling user behavior, or using trails or computational models to recommend future actions. If findings show that users benefit from trail following, likely post-query trails could be considered in search system design and even as units of retrieval [23].

3. SEARCH TRAILS

In this section we describe the logs, trail mining from the logs, automatic classification of trail pages, and summary trail statistics.

3.1 Log Data

The primary source of data for this study was the anonymized logs of URLs visited by users who opted in to provide data through a widely-distributed browser toolbar. These log entries include a unique identifier for the user, a timestamp for each page view, a unique browser window identifier (to resolve ambiguities in determining which browser a page was viewed), and the URL of the Web page visited. Intranet and secure (https) URL visits were excluded at the source to maintain user privacy. In order to remove variability caused by geographic and linguistic variation in search behavior, we only include entries generated in the English speaking United States locale. The results described in this paper are based on a sample of URL visits during a three-month period from March 2009 through May 2009, representing millions of URL visits from 100,000 unique users. The user sample was selected at random from a larger set of twelve million users after we had pre-filtered the data to remove several thousand extremely-active outlier users, all of whom issued over one thousand queries per day on average across the three-month period. These high-volume users were likely automated traffic. For each user, we required an adequate number of Web page visits to create their long-term search history that was used to evaluate source novelty (described in more detail later). Therefore, in addition to removing outliers, we also only selected users who issued at least 30 queries per month from March 2009 to May 2009 inclusive.

3.2 Trail Mining

We mined tens of millions of search trails from the May 2009 logs, referred to hereafter as T_x. As defined by White and Drucker [33], search trails consist of a temporally-ordered sequence of URLs beginning with a search engine query and terminating with either: (i) another query, (ii) a period of user inactivity of 30 or more minutes, or (iii) the termination of the browser instance or tab. The 30-minute inactivity timeout is commonly used to demarcate sessions in Web log analyses (e.g., [9]). We chose to use search trails rather than session trails (which comprise multiple queries) to lessen the likelihood of query skew, where user intent shifts over the course of the session, making it challenging to associate visited pages to the original query. Figure 1 illustrates a search trail, expressed as a Web behavior graph [5]. The trail starts with a search engine query ($Q1$) (which also includes the search-engine result page (SERP)) and comprises a set of pages visited until the trail terminates with a new query or an inactivity timeout. The nodes of the graph represent Web pages that the user has visited: rectangles represent page views and rounded rectangles represent search engine result pages. Vertical lines represent backtracking to an earlier state (e.g., returning to a page of results in a search engine after following an unproductive link). A "back" arrow, such as that below $P4$, indicates that the user has requested to visit a page seen earlier in the search trail. Time runs left to right and then from top to bottom. In addition to the complete trail, also marked on Figure 1 are the origin (the search result,

$P2$), the destination (the trail's terminal page, $P5$), and the pages between origin and destination (in this case $\{P3, P4, P3, P2\}$).

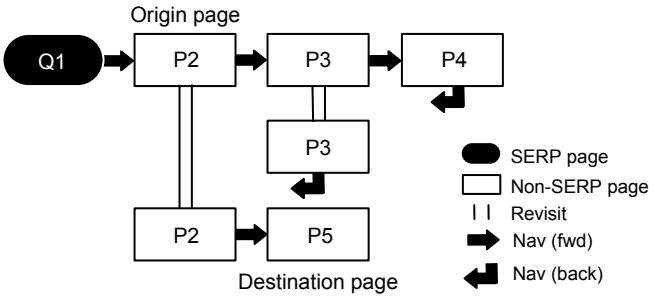

Figure 1. Web behavior graph illustrating a search trail.

3.3 Trail Labeling

Three of the five evaluation metrics used in our study—coverage, diversity, and novelty—use information about page topicality. Millions of unique URLs were present in the set of all trails mined from the toolbar logs. This made the evaluation of coverage, novelty, and diversity challenging as it was impractical to download all pages and comparisons based on URLs would be severely limited. To address this challenge, we classified the Web pages sourced from each context into the topical hierarchy from a popular Web directory, the Open Directory Project (ODP) (dmoz.org). Given the large number of pages involved, we used automatic classification. Our classifier assigned labels to pages based on the ODP using a similar approach to Shen et al. [21]. Classification began with URLs present in the ODP and incrementally pruned non-present URLs one path level at a time until a *match* was found or *miss* declared. Similar to [21], we excluded the "Regional" and "World" top-level ODP categories since they are typically uninformative for building interest models.

3.4 Trail Statistics

There were 15 million search trails followed by the 100,000 toolbar users in our sample during May 2009. The median (Med) number of trails followed per user was 91 (mean (M) was 160, standard deviation (SD) was 228). The median number of steps in the trails was two (M=5.3, SD=12.2), (i.e., the search engine result page and a single result click), but around one third of the trails were abandoned following the query, and around one third of trails contained three or more pages. The median time spent on trails was 81 seconds (M=308s, SD=615s), and around 20% of trails contained backtracking to a site already visited in the trail.

Interestingly, around 19.3% of trails with three steps or more (i.e., had pages between the origin and destination) had at least one site with a different ODP label to the origin and destination pages. Analysis of the queries on the remaining 80.7% of trails revealed that their original queries were generally navigational (e.g., *[delta airlines]*) or directed informational (e.g., *[what is daylight savings time?]*). For other types of informational query, such as undirected, advice, locate or list [20], intermediate pages may be valuable to users. The extent of this value is estimated in our study.

4. STUDY

We devised an experiment to determine the value of search trails compared to search results and destinations. In this section we outline the research questions that drove our study, describe the experimental variants, summarize the trail data preparation, and present the metrics used to compare sources.

4.1 Research Questions

Our study answers a number of research questions. Specifically, of the four sources (origin, destination, sub-trail, and full-trail), which: (i) provide more relevant information? ($RQ1$); (ii) provide more topic coverage? ($RQ2$); (iii) provide more topic diversity? ($RQ3$); (iv) provide more novel information? ($RQ4$), and; (v) provide more useful information? ($RQ5$). Answers to these questions help us understand the value of trail (or sub-trail) traversal compared to viewing only the origin and/or destination pages.

4.2 Trail Sources

To determine the value of trail traversal we experiment with a number of trail sources. They are as follows:

Origin: The first page in the trail after the SERP, visited by clicking on a search result hyperlink. This is regarded as a baseline in this study since current search engines show this source alone in search results. $P2$ is the origin in Figure 1.

Destination: The last page in the trail, visited prior to trail termination through a follow-up query or inactivity timeout. Destinations are defined similarly to the *popular destinations* from White et al. [32]. We include them here for comparison with that earlier work. $P5$ is the destination in Figure 1.

Sub-trail: All pages in the trail except for destination, including all post-SERP pages. $\{P2, P3, P4, P3, P2\}$ is the sub-trail in Figure 1.

Full-trail: The complete trail, including all post-SERP pages.

We mine these sources from each trail in T_x and compute the value of each source in terms of relevance, coverage, diversity, novelty, and utility across all queries and divided by query type. We elected not to study intermediate pages directly (i.e., pages in the trail that lie between the origin and destination) since a trail must contain an origin page in our current definition. The value of the intermediate pages over the origin can be estimated by comparing the performance differences between *origins* and *sub-trails*.

4.3 Trail Data Preparation

To help ensure experimental integrity, we did not use all search trails in T_x. Instead, we filtered T_x based on the following criteria:

- Queries originating the trails were normalized to facilitate comparability between trails, and between the trails and other resources (as described in the next section). Normalization involved the removal of punctuation, lowercasing, trimming extraneous whitespace, and ordering terms alphabetically.
- Trails were required to contain at least three pages: an origin page, a destination page, and at least one intermediate page. It was important to have these sources in all trails used since we wanted to compare their value.
- To ensure that origin pages were reached through a SERP click, we required that the first non-SERP page in the trail be connected to the SERP with a hyperlink click (i.e., the referrer of the origin page must be a SERP). Trail pages thereafter were not required to be joined via a hyperlink click.
- The coverage of our ODP classifier with URL back-off was approximately 65%. A missing label may have skewed the distribution of labels for or against a particular source. We therefore required that all selected trails be fully labeled.
- To prevent sample bias from highly-active users, we selected at most 10 search trails that met the above criteria from each user.

The application of these criteria reduced T_x to one quarter of its original size, but yielded a high-quality data set for our study.

4.4 Metrics

We used five metrics to compare the different trail sources: relevance, coverage, diversity, novelty and utility. These metrics were chosen to capture many important elements of information seeking, as highlighted by the wealth of relevant research in the IR community (e.g., [7][8]). The use of multiple metrics allowed us to compare the value of the different sources in different ways. For example, a trail destination page may be less relevant than sub-trail, but may provide additional information not in the sub-trail. We now describe each metric and its implementation.

4.4.1 Relevance

The first metric used to compare the sources was relevance to the query that initiated the trail. In addition to the trail data used during the course of this study, we also obtained human relevance judgments for over twenty thousand queries that were randomly sampled by frequency from the query logs of the Bing search engine; they were normalized per the description in Section 4.3, and were present in T_x. Trained judges assigned relevance labels on a six-point scale—*Bad*, *Poor*, *Fair*, *Good*, *Excellent* and *Perfect*—to top-ranked pooled Web search results for each query from the Google, Yahoo!, and Bing search engines as part of a separate search engine assessment activity. This provided hundreds of relevance judgments for each query. These judgments allowed us to estimate the relevance of information encountered at different parts of the trails. For each trail in T_x, we computed the average relevance judgment score for each source. Each page in the trail was used at most once in relevance score calculations, even if it appeared multiple times in the trail. This discounted revisitation, since diminishing returns from each repeat visit to a page in the same trail were likely. In this analysis we only used trails for which we had a relevance judgment for the origin page, the destination page, and at least one intermediate page. Trails for 8,712 queries, comprising a query set R and initiating around two million trails, afforded a detailed comparison of source relevance.

4.4.2 Coverage

Another aspect that we studied was topic coverage, meant to reflect the value of each trail source in providing access to the central themes of the query topic. To estimate the coverage of each trail source, we first constructed a set of *query interest models* representing the dominant intents associated with each query in R. These models served as the ground truth for our estimates of coverage (in this subsection) and diversity (in the next subsection). Each constructed query interest model is assumed to contain most of the significant themes for the query. A query's interest model comprises the ODP category labels assigned to the URLs in the union of the top-200 search results for that query from Google, Yahoo! and Bing. ODP category labels are grouped and their frequency values are normalized such that across all labels they sum to one. For example, the highest-weighted labels in the query interest model for *[solar system discoveries]*, and their associated normalized frequencies (w_l), are shown in Figure 2.

Label	w_l
$Top/Science/Technology/Space/NASA$	0.64
$Top/Science/Technology/Space/News\_and\_Media$	0.18
$Top/Reference/Encyclopedias$	0.16

Figure 2. Top ODP categories for *[solar system discoveries]*.

To improve the reliability of our coverage estimates, we selected a set of query interest models, Q_x, that were required to be based on at least 100 fully-labeled search results (i.e., were not missing a label and did not have a label from an ignored ODP category) and were based only on labels with a frequency count of at least five (to reduce label noise). T_x was modified to include only trails originating from queries with interest models in Q_x. For each trail t in T_x, we created a *source interest model* comprising ODP category labels and associated frequencies for *origin*, *destination*, *sub-trail*, or *full-trail*. We then compute the coverage of each source s in t (denoted t_s) using:

$$Coverage(t_s) = \sum_{l \in (s \cap q_x)} w_l \quad (1)$$

Where l is ODP category label and w_l represents the normalized frequency weight of that label in the corresponding interest model for the current query, denoted as q_x.

4.4.3 Diversity

Another aspect studied was topic diversity, which estimates the fraction of unique query-relevant concepts surfaced by a given trail source. Exposure to different perspectives and ideas may help users with complex or exploratory search tasks. Indeed, existing search engines already consider diversity in the search results they present to satisfy more users with the first few results.

To estimate the diversity of information provided by each trail source we use an approach similar to our coverage estimation. We generate trail interest models for each trail source and compare those with the relevant query interest model to estimate diversity. The main difference between how the estimates of coverage and diversity lies in whether normalized label frequency is considered. When estimating coverage we want to establish the fraction of q_x appearing in t_s (i.e., label frequency is used). In contrast, when we estimate diversity, we only count the number of unique category labels from q_x that appear in t_s (i.e., frequency is ignored).

For each trail t in T_x originating with one of the queries in Q_x, we created a source interest model comprising ODP category labels and associated frequencies for *origin*, *destination*, *sub-trail*, and *full-trail.* We computed diversity for each t_s using:

$$Diversity(t_s) = \sum_{l \in (s \cap q_x)} \frac{1}{|q_x|} \quad (2)$$

Where l is ODP label and $|q_x|$ is the number of unique q_x labels.

4.4.4 Novelty

Another aspect that we studied was the amount of new query-relevant information from each trail source. Novel information may help users learn about a new subject area or broaden their understanding of an area with which they are already familiar.

Trails with novelty contain information that users have not encountered for a query. Unlike coverage and diversity, the novelty provided by a trail source may depend on both the query and the user. For example, what is new topic-related information for one individual may not be new information for another. Therefore, to estimate the novelty of the information provided by each trail source, we first had to construct a model of each user's general interest in the query topic based on historic data. To do this, we leveraged users' search trails for the two-month period from March to April 2009 inclusive (referred to hereafter as T_h), and constructed *historic interest models* H, for all user-query pairs. Each interest model h_x, whose query was present in Q_x, comprised a distribution of ODP category labels (and associated nor-

malized frequencies) similar to those used in earlier coverage and diversity estimates. Only labels appearing in the query interest model q_x are included in h_x. The historic interest model is therefore a subset of q_x focused on a given user's history with that query. White et al. [31] used a similar approach to depict long-term user interests. We estimate the novelty of each trail source relative to the historic interest model for the user and the query.

For each trail t in T_x, we built source interest models to estimate the source novelty based on whether it contained topic-related information *not in* h_x. The novelty of each t_s is estimated using:

$$Novelty(t_s) = \sum_{l \in (s \cap q_x) \wedge l \notin h_x} \frac{1}{|q_x|} \quad (3)$$

Where l represents an ODP category label present in s and q_x but not in h_x, and $|q_x|$ represents the number of unique q_x labels.

4.4.5 Utility

The final aspect that we studied was the utility of each of the trail sources, estimated for the purposes of this study using page dwell time (i.e., the amount of time spent on a particular page by a user). Dwelling on a page for a significant amount of time implies that a user may be deriving utility from it. Indeed, prior research has shown that during search activity, a dwell time of 30 seconds or more on a Web page can be indicative of page utility [11]. We apply this threshold in our analysis and across all trails in T_x, we estimate the fraction of page views from the *origin*, *destination*, *sub-trail*, and *full-trail* that exceed this dwell time threshold.

In all metrics used in this study, a higher value is regarded as a more positive outcome. The metrics are computed for each trail, then micro-averaged within each query, and then macro-averaged across all queries to obtain a single value for each source-metric pair. This procedure ensures that all queries are treated equally in the analysis and popular queries are not allowed to dominate the aggregated metric values for each source. Although we might expect *sub-trails* and *full-trails* to have higher metric scores than *origins* or *destinations* (simply because they have more pages), it is the extent that the metrics' values increase from these sources that lets us estimate the additional value of trails and sub-trails. This is reasonable since we plan to show *full-trails* and *sub-trails* directly to users on the search engine result page.

4.5 Methodology

In this section so far we have described the research questions, the four trail sources evaluated, trail data preparation procedures, and the metrics used to evaluate the sources. The methodology employed during our experiments comprised the following steps:

1. Construct the set of query interest models Q_x based on the set of queries for which we have human relevance judgments (R).
2. Construct historic interest models (H) for each user-query pair in T_h, filtered to only include queries appearing in Q_x.

The data sets created during the first two steps are used to evaluate each of the four trail sources.

3. For each search trail t in T_x:
   a. Assign ODP labels to pages all pages in t.
   b. Build source interest models for the *origin*, *destination*, *sub-trail* and *full-trail* sources.
   c. Compute relevance, coverage, diversity, novelty and utility using the methods described in Section 4.4.
4. Compute the average values for each metric per query, and then average across all queries (to treat all queries equally), breaking out the findings by query type as appropriate.

In the next section we report on the findings from our study.

5. FINDINGS

We first present findings over all queries; then divided by query type, varying query popularity and query re-finding behavior, both of which have been shown to influence search interaction in previous work [8][10]. Since our data were shown to be normally distributed, we use parametric statistical testing, with $\alpha = .05$.

5.1 All Queries

We computed the five metrics across all trails in T_x and now report on source performance.

Relevance: We begin our analysis by reporting on the relevance of the information encountered at the *origin*, *destination*, *sub-trail* and *full-trail*, determined using human relevance judgments. As noted in the previous section, the judgments were captured for query-URL pairs on a six-point scale, ranging from 0 (*Bad*) to 5 (*Perfect*). Sources that provide more relevant information would be expected to have a higher average relevance score. In the "All" column of Table 1 (shaded) we report on the mean average relevance score obtained from each of the sources across all trails in T_x. Also reported are the percentage differences between the relevance score obtained for each of the non-origin sources and *origins* (Δ%) to estimate the additional value obtained from full or partial trail traversal, or from teleporting directly to destinations. We do not show standard deviations to avoid crowding findings.

The findings show that the relevance scores for all sources were generally positive (around three or *Good*). An independent-measures analysis of variance (ANOVA) computed between the relevance scores obtained from all four sources revealed no significant differences in the relevance of the origin page versus information encountered on the trail ($F(3,8708) = 1.5$, $p = 0.21$). However, as is apparent in the table, trends in the findings suggest that the relevance scores for non-origin sources were slightly lower than those of the origin pages (e.g., 3.3 versus 2.9-3.0). This may be related to a combination of the distance between non-origin sources and the original queries, and the effect of dynamism in information needs as users traverse search trails [33]. Since non-origin sources are further from the query than origin pages, they may be less query relevant as user needs evolve.

Coverage: We also studied the extent that each trail source covered the query interest models representing the dominant themes for each query. The coverage estimate of each source for each trail was computed using Equation 1. The average coverage scores for each metric are reported in the "All" column of Table 1, along with the percentage difference between each of the sources and *origin*. The findings show that on average, around 40% of the total mass of the query interest models can be covered by *origins* and *destinations*, and around 50% are covered by *sub-trails* and *full-trails* (coverage gains of 20-30% from traversing trails). Analysis of the findings using a one-way independent measures ANOVA revealed statistically significant differences between the sources ($F(3,8708) = 5.5$, $p < .001$). Post-hoc testing, performed using Tukey tests, revealed that on average across all queries, *sub-trails* and *full-trails* covered more of the query interest models in Q_x than the *origins* or *destinations* alone (all $p < 0.01$).

Table 1. Metric scores across all queries and broken down by query popularity and query history. Statistically-significant differences between non-origin trail sources and the origin *within each metric* are shown in **bold** ($p \leq .05$) and ***bold-italic*** ($p \leq .01$).

	Source	All		Query breakdown											
				Query Popularity (per query)						Query History (per user-query pair)					
				Low		Medium		High		None		Some		Lots	
		N=8,712		N=211		N=6,421		N=2,080		N=1,022,874		N=1,081,895		N=18,220	
		M	Δ%	M	Δ%	M	Δ%	M	Δ%	M	Δ%	M	Δ%	M	Δ%
Relevance	Origin	3.3		2.9		3.2		3.4		3.1		3.3		3.4	
	Destination	2.9	–12	2.6	–10	2.9	–9	3.1	–9	2.7	–13	3.0	–9	3.0	–12
	Sub-trail	3.0	–9	2.8	–3	3.0	–6	3.1	–9	2.8	–10	2.9	–12	3.0	–12
	Full trail	3.0	–9	2.8	–3	3.1	–3	3.2	–5	2.8	–10	3.0	–9	3.1	–9
Coverage	Origin	0.377		0.355		0.374		0.389		0.382		0.374		0.373	
	Destination	0.372	–1	0.349	–2	0.369	+1	0.385	–1	0.385	+1	0.371	–1	0.367	–2
	Sub-trail	***0.455***	+21	***0.454***	+28	***0.455***	+22	***0.456***	+17	***0.472***	+24	***0.439***	+17	***0.410***	+10
	Full trail	***0.489***	+30	***0.485***	+37	***0.488***	+30	***0.492***	+26	***0.502***	+31	***0.476***	+27	***0.457***	+23
Diversity	Origin	0.291		0.287		0.290		0.293		0.293		0.290		0.290	
	Destination	0.307	+5	0.296	+3	0.307	+6	0.308	+5	0.311	+6	0.305	+5	0.304	+5
	Sub-trail	***0.384***	+32	***0.369***	+29	***0.384***	+32	***0.385***	+31	***0.398***	+36	***0.370***	+28	***0.339***	+17
	Full trail	***0.412***	+42	***0.407***	+42	***0.413***	+42	***0.413***	+41	***0.433***	+48	***0.394***	+36	***0.361***	+24
Novelty	Origin	0.034		0.031		0.034		0.036		n/a	n/a	0.034		0.010	
	Destination	0.045	+32	0.043	+39	0.044	+29	0.046	+28	n/a	n/a	0.046	+35	0.012	+20
	Sub-trail	***0.127***	+273	***0.125***	+303	***0.126***	+271	***0.129***	+258	n/a	n/a	***0.129***	+279	***0.040***	+300
	Full trail	***0.159***	+367	***0.156***	+403	***0.159***	+368	***0.162***	+350	n/a	n/a	***0.161***	+374	***0.066***	+560
Utility	Origin	0.473		0.473		0.473		0.473		0.440		0.468		0.492	
	Destination	**0.498**	+5	**0.493**	+4	**0.497**	+5	**0.502**	+6	**0.461**	+4	**0.489**	+4	**0.523**	+6
	Sub-trail	***0.624***	+32	***0.617***	+30	***0.623***	+32	***0.629***	+33	***0.599***	+36	***0.645***	+38	***0.664***	+35
	Full trail	***0.653***	+38	***0.649***	+37	***0.653***	+38	***0.656***	+39	***0.626***	+42	***0.676***	+44	***0.689***	+40

Diversity: To estimate the extent that each trail source covers *different* aspects of the query interest model, we calculated their diversity using Equation 2. Increased diversity may be useful to users engaged in search tasks with multiple sub-tasks, such as planning a vacation. The average coverage scores for each source across all trails in T_x are reported in the "All" column of Table 1. The findings show that approximately one-third of the central themes for a query can be captured by each trail source, with more topic diversity coming from the trail-based sources (diversity gains of 30-40% from traversing trails). Statistical analysis of the findings reveals significant differences between the levels of topic diversity provided by each source ($F(3,8708) = 7.0$, $p < 0.001$). Post-hoc testing revealed that *sub-trails* and *full-trails* provide more diversity than *origins* (all $p < 0.01$). The increase in diversity for *destinations* over *origins* was not statistically significant.

Novelty: Novelty calculations estimate the amount of new query-relevant information provided to users by each of the trail sources. Unlike the other metrics in this study, novelty is specific to both user and query; one user's experience with a query may differ from another's. As described previously, novelty is computed based on the number of new query-relevant ODP category labels added to a user's query interest models compared with historic data. In Table 1 ("All" column) we report on the average novelty score and the percentage differences between all non-origin sources and the origin only. The findings show modest increases in the amount of new information obtained from all sources, but seemingly larger gains from the non-origin trail sources (0.13-0.16 versus 0.03). Statistical analysis of our findings revealed differences among the sources ($F(3,8708) = 3.0$, $p = .01$). Post-hoc testing revealed significant differences between *sub-trails* / *full-trails* and *origins* / *destinations* (all $p < 0.01$). On average, trails provide more novel information than *origins* or *destinations*. In turn, *destinations* provide slightly more novel information than *origins*, but differences were not significant ($p = 0.12$).

Utility: We also studied the utility of each trail source. To estimate utility for a given Web page from the logs, we used a 30-second page dwell time threshold selected based on previous work [11]. For each of the sources across all trails in T_x, we computed the fraction of trails for which each source contained a useful page (i.e., a page with a dwell time equaled or exceeded the 30-second threshold). These values are shown in the "All" column of Table 1. Also shown are the percentage differences between non-origin

sources and *origins*. The findings show that just under half of *origins* and *destinations* are useful, around 60% of *sub-trails* have useful pages, and almost two-thirds of *full-trails* contain useful pages. Statistical analysis of the findings revealed significant differences between the sources in terms of their estimated utility ($F(3,8708) = 3.3$, $p = .01$). Post-hoc testing revealed that all sources differed from trail origins (*destinations*: $p = .03$; *sub-trails*: $p < .01$; *full-trails*: $p < .01$). It seems that users find non-origin pages more useful than origin pages. This may be because origin pages are search results and may only be the starting points for a search task or sub-task [24].

One important factor that may cause variation in the effectiveness of search trails is the nature of the search query. Downey et al. [10] showed that user behavior following a query varied significantly with query popularity. Teevan et al. showed that the frequency with which a query is reissued by a given user over a period of time (so-called "re-finding" behavior) affects that user's search interactions for that particular query [25]. To test whether such factors influenced the source value we varied query popularity and history as part of our experimental design. In the remainder of this section we report on the findings of this analysis.

5.2 Effect of Query Popularity

To study the effect of query popularity on source value, we created a tripartite division of queries in T_x, grouping them into *low*, *medium*, and *high*, based on user frequency in T_h. Low popularity queries were issued by at most one user in T_h, medium popularity queries were issued by between 1 and 100 users in T_h, and high popularity queries were issued by over 100 users in T_h.

Table 1 presents findings on the effect of query popularity on source performance for each of the five metrics we study. On all metrics, we observe a trend that as query popularity increases, each of the metric values also increases. The relative ordering and percentage gains from the trail sources remain consistent across all five metrics. However, within each metric, differences in the values obtained for the three query popularity groupings are not significant using a two-way independent measures ANOVA with source and query popularity group as the factors (source (rows): all $p \leq .02$; popularity (columns): all $p \leq .13$). F-statistics for all performed ANOVA are not reported to avoid crowding the paper. Small increases in coverage as query popularity increases may be attributable to the dominance of the intent associated with the query. More popular queries are more likely to have a single dominant intent, giving the category label for that intent a high weight (w_l from Equation 1). Since coverage derives from w_l, we are likely to observe increases in coverage as a dominant intent with a high w_l. Improvements in search engine performance as query frequency increases (already noted in [10]) may account for some of the slight increases in relevance and utility with popularity (Table 1).

5.3 Effect of Query History

We also studied the effect of query history on the value of each of the four trail sources. We divided queries into three groups—*none*, *some*, and *lots*—based on the number of times they were issued by a particular user in T_h. Queries in *none* appeared in T_x but did not appear in T_h, queries in *some* appeared in T_x and were issued by a particular user 30 times or less in T_h (i.e., on average less than once every two days), and queries in *lots* appeared in T_x and were those issued by a particular user more than 30 times in T_h (i.e., on average more than once per two days).

Table 1 presents findings on the effect of query history on source performance for each of the five metrics. From the findings, it seems that as query history increases, there is a mixed effect on the five metrics. However, within each metric all differences between sources and between query history groupings are significant, as shown by a two-way independent measures ANOVA with source and query history grouping as the factors (source (rows): all $p \leq .001$; history (columns): all $p \leq .001$). We found that relevance and utility rise across all sources given increased re-finding behavior. This is perhaps because users are more familiar with the query topic and are more able to identify relevant information. Similar findings have been reported in previous work on topic familiarity (e.g., [15]). In contrast, coverage, diversity, and novelty decrease, perhaps as a result of a reduced variance in the pages visited. Such consistency in interaction behavior for queries with high re-finding rates has been reported previously [27].

6. DISCUSSION AND IMPLICATIONS

We have demonstrated that following search trails provides users with significant additional benefit in terms of coverage, diversity, novelty, and utility over origins and destinations. Although more work is required to supplement the methodology used in our study and further understand the impact of experimental decisions such as only studying search trails that could be fully-labeled using ODP lookup, our log analysis helps establish the value of trails to users and inform search system design.

We showed that *full-trails* and *sub-trails* provided significantly more coverage, diversity, novelty, and utility, versus trail origins and destinations. The one metric for which we did not obtain significant differences between origin and non-origin sources was relevance. Trends in the findings suggest that trails were less relevant than origins. This may be related to the definition of relevance in this study. Our relevance judgments are assigned to pairs of queries and search results. However, during the session, user intent may shift and the relevance to the initial query is dynamic [2]. Pages encountered on the trails may be relevant but not appear so due to these shifts. More work is required on how relevance changes during browsing and to understand the relevance benefit from trails. Enhancements include studying the cumulative relevance of trail information, considering relevance changes, and devising proxies for relevance similar to that used for utility.

Destinations were more useful and led to a slight novelty increase over origins. This confirms some of the findings of White et al. [32], who showed in a user study that destinations were a useful addition to the results interface. In retrospect, this agrees with expectations that users will give more attention, and hence dwell on the destination page. In information foraging theory [18] where this corresponds to a food patch, users satisfy some or all of their need and do not pursue the information scent further.

While destinations were useful in one metric, adding intermediate pages contributed to gains in several metrics, notably in novelty, diversity, and utility, where the differences between *origins* and *sub-trails* are substantial. The success of *sub-trails* suggests that users may not need to traverse *full-trails* to derive significant value from post-query navigation. As expected, *full-trails* provide even more benefit than *sub-trails*; *full-trails* are *sub-trails* plus *destinations*. Although the findings of our study appear to support trail recommendation, they also suggest that the nature of the query is important. For some queries, the trails might be useful in supporting exploration, but for other queries, especially for focused tasks, presenting trail information might be a hindrance.

Questions remain about how to select trails and how to integrate trails into the SERP. Popular search trails are typically short and obvious, so we need to consider diverse and unexpected trails, perhaps leveraging popular sub-trails as well as full trails in trail selection algorithms. Trail selection methods could discount trails with numerous cases of rapid backtracking or maximize relevance, coverage, diversity, novelty, and utility with the shortest path. Alternatively, we can personalize trail recommendation by weighting trails based on the extent of the current user's re-finding behavior or perform *a priori* trail analysis to recommend trails when the destination is unclear (i.e., users end up on many pages), and present trail destinations when the destination is clear (i.e., many users end up at the same page). Trails can be presented as an alternative to result lists, as instant answers above result lists, in pop-ups shown after hovering over a result, below each result along with the snippet and URL, or even on the click trail a user is following. Follow-up user studies and large-scale flights will further analyze trail appropriateness for different queries and compare trail selection algorithms and trail presentation methods.

7. CONCLUSIONS

In this paper we have presented a study estimating the value of search trails to users. Our log-based methodology has allowed us to systematically compare the estimated value of trails to other trail components: trail origins (clicked search results), the trail destinations (terminal trail pages), and sub-trails comprising the origin plus intermediate pages. We studied the relevance, coverage, diversity, novelty, and utility of each of the four sources using metrics devised for this purpose, human relevance judgments, historic log data, and URL classification where appropriate. When we varied the query by overall popularity, the values of each metric increased with query frequency. The evaluation showed that *full-trails* and *sub-trails* provide users with significantly more topic coverage, topic diversity, and novelty than trail origins, and slightly more useful but slightly less relevant information than the origins. Our findings show that there is value in the trail (the scenic route), as well as the origin and the destination. These findings vary slightly by query popularity over all users and significantly by the level of re-finding performed by a user for a given query. The next steps are to investigate best-trail selection for query-origin pairs and add trails to search engine result pages.

REFERENCES

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

[2] Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5): 407-424.

[3] Bilenko, M. & White, R.W. (2008). Mining the search trails of surfing crowds: identifying relevant websites from user activity. *Proc. WWW*, 51-60.

[4] Bush, V. (1945) As we may think. *Atlantic Monthly*, 3(2): 37-46.

[5] Card, S.K. et al. (2001). Information scent as a driver of web behavior graphs: results of a protocol analysis method for web usability. *Proc. SIGCHI*, 498-505.

[6] Chalmers, M., Rodden, K. & Brodbeck, D. (1998). The order of things: activity-centered information access. *Proc. WWW*, 359-367.

[7] Clarke, C.L.A. et al. (2008). Novelty and diversity in information retrieval evaluation. *Proc. SIGIR*, 659-666.

[8] Cole, M. et al. (2009) Usefulness as the criterion for evaluation of interactive information retrieval. *Proc. HCIR*, 1-4.

[9] Downey, D., Dumais, S. & Horvitz, E. (2007). Models of searching and browsing: languages, studies, and application. *Proc. IJCAI*, 2740-2747.

[10] Downey, D. et al. (2008). Understanding the relationship between searchers' queries and information goals. *Proc. CIKM,* 449-458.

[11] Fox, S. et al. (2005). Evaluating implicit measures to improve the search experience. *ACM TOIS,* 23(2): 147-168.

[12] Freyne, J. et al. (2007). Collecting community wisdom: integrating social search and social navigation. *Proc. IUI*, 52-61.

[13] Fu, W.-T. & Pirolli, P. (2007). SNIF-ACT: A cognitive model of user navigation on the world wide web. *Human-Computer Interaction*, 22(4): 355-412.

[14] Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proc. SIGKDD*, 133-142.

[15] Kelly, D. & Cool, C. (2002). The effects of topic familiarity on information search behavior. *Proc. JCDL*, 74-75.

[16] Olston, C. & Chi, E.H. (2003). ScentTrails: integrating browsing and searching on the web. *ACM TOCHI*, 10(3).

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

[18] Pirolli, P. & Card, S.K. (1999). Information foraging. *Psychological Review*, 106(4): 643-675.

[19] Reich, S. et al. (1999). Where have you been from here? Trails in hypertext systems. *ACM Computing Surveys*, 31(4).

[20] Rose, D.E. & Levinson, D. (2004). Understanding user goals in web search. *Proc. WWW*, 13-19.

[21] Shen, X., Dumais, S. & Horvitz, E. (2005). Analysis of topic dynamics in web search. *Proc. WWW*, 1102-1103.

[22] Singhal, A. (2001). Modern information retrieval: a brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4): 35-43.

[23] Singla, A., White, R.W. & Huang, J. (2010). Studying trail-finding algorithms for enhanced web search. *Proc. SIGIR*.

[24] Teevan, J. et al. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. *Proc. SIGCHI,* 415-422.

[25] Teevan, J. et al. (2007) Information re-retrieval: repeat queries in yahoo's logs. *Proc. SIGIR*, 151-158.

[26] Trigg, R.H. (1988). Guided tours and tabletops: tools for communicating in a hypertext environment. *ACM TOIS*, 6(4).

[27] Tyler, S.K. & Teevan, J. (2010). Large scale query log analysis of re-finding. *Proc. WSDM,* 191-200.

[28] Wang, X. & Zhai, C. (2009). Beyond hyperlinks: organizing information footprints in search logs to support effective browsing. *Proc. CIKM*, 1237-1246.

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

[30] Wheeldon, R. & Levene, M. (2003). The best trail algorithm for assisted navigation of web sites. *Proc. LA-WEB,* 166.

[31] White, R.W., Bailey, P. & Chen, L. (2009). Predicting user interests from contextual information. *Proc. SIGIR*, 363-370.

[32] White, R.W., Bilenko, M. & Cucerzan, S. (2007). Studying the use of popular destinations to enhance web search interaction. *Proc. SIGIR*, 159-166.

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