Path Following in Social Web Search

Jared Lorince[1]*, Debora Donato[2], and Peter M. Todd[1]

[1] Cognitive Science Program
Department of Psychological & Brain Sciences
Indiana University, Bloomington, Indiana 47405, USA
{jlorince,pmtodd}@indiana.edu
[2] StumbleUpon, Inc.
San Francisco CA 94107, USA
debora@stumbleupon.com

Abstract. Many organisms, human and otherwise, engage in path following in physical environments across a wide variety of contexts. Inspired by evidence that spatial search and information search share cognitive underpinnings, we explored whether path information could also be useful in a Web search context. We developed a prototype interface for presenting a user with the "search path" (sequence of clicks and queries) of another user, and ran a user study in which participants performed a series of search tasks while having access to search path information. Results suggest that path information can be a useful search aid, but that better path representations are needed. This application highlights the benefits of a cognitive science-based search perspective for the design of Web search systems and the need for further work on aggregating and presenting search trajectories in a Web search context.

Keywords: Search Paths, User Study, Path Following, Social Search

1 Introduction

Path following is ubiquitous among social species in natural environments, be it mediated by stigmergic pheromone trails of ants and termites [1], emerging from crowd dynamics [2], or evidenced by the the reinforcement of worn paths through grass and snow on college campuses [3]. On the Web, too, we follow paths — albeit implicitly — when the search results we encounter, videos we watch, and products suggested to us all depend on the interaction patterns of the users who went before us. In this paper we take inspiration from work on path following in physical environments to explore whether sharing explicit search paths between Web searchers can be a useful search aid.

Research in cognitive science suggests that goal-directed cognition is an evolutionary descendant of spatial foraging capacities [4], and an increasing number of studies show that the way humans search in information spaces is deeply linked to the way we search in spatial environments [5, 6]. This conclusion is also bolstered by the embodied view of cognition, which highlights the connections between information processing and bodily movement in space [7].

With this in mind, we developed and tested a prototype interface for applying the notion of path following to a Web search environment. A path, by

* The first author was an intern at, and the second employed by, Yahoo! Research during study development. Data collection/analyses were done at Indiana University.

definition, carries a special kind of information typically lacking from Web-based recommendations and other search tools: It provides not only a destination, but a route from one's current location to that destination, thus delineating what lies between. This may seem a simple point, but the vast majority of tools for guiding information search on the Web — from product recommendations on Amazon to "Also try:" suggestions on Yahoo! search — are pointillistic: You should issue *this* query, or buy *this* product. This is not to say that such recommendation systems are not utilizing path-like data "under the hood" to generate suggestions, but to our knowledge there exist no systems in regular use that explicitly share paths — sequences of activity extended over time — between users.

In a Web context, sharing path information creates opportunities for serendipitous discoveries by exposing the user to content that would be missed by "teleporting" directly to a recommended resource. When these paths are relevant to the current search context, they can provide windows on how to approach a search task that the user might not consider otherwise, and that likely could not be readily communicated via pointillistic recommendations. This of course holds little value in cases where a user's query has a clear, discrete answer ("What is the capital of North Dakota?"). Many search tasks we engage in, however, are simultaneously more complex and less explicitly defined ("What car should I buy?", "What is fun to do in North Dakota?"). In these cases, paths can capitalize on modern Web users' interest in shared social content and propensity for social copying. We hypothesized that the incorporation of path information into the search interface would lead to increased levels of (1) user engagement and (2) satisfaction with solutions to assigned search tasks. To explore our hypotheses, we developed a custom search engine interface that incorporated path information. Study participants were assigned a series of search tasks, and presented with the paths taken by previous users performing the same task.

A full understanding of search path use requires work at three levels: The cognitive-behavioral (what is the theoretical case for using path information in search and how do people respond to it), the algorithmic (how can search paths be generated and coherently aggregated across multiple users), and design-centric (how should such paths be presented to users). Here we address the first level, as a preliminary attempt to explore how path-like information can be translated to a Web search context. While some of our positive results are suggestive of the power of this approach, our other negative results also indicate that it will be crucial to determine better ways of presenting path information if it is to be helpful to users. Thus a principal goal of this paper is to encourage future work that explores methods for creating and presenting useful path information to individuals searching the Web an other information spaces.

2 Related work

Cognitive science: Goldstone and colleagues' work in collective behavior [8–10] exemplifies an expanding interest in the emergent dynamics of human groups within the cognitive science community. Of particular interest for the present study is their work on group path formation [11, 3]. In experiments where human participants controlled avatars in a virtual environment, the authors studied the trail systems that emerged as the participants moved from one destination to another while attempting to minimize their travel costs. Emulating physical environments, where people's movement along a path facilitates future movement along that path (e.g. paths through the snow, or the shortcuts often found

through the lawns of houses on corner lots), the authors reduced travel costs through any particular traversed point in the virtual environment by making them inversely proportional to the number of individuals who had previously move through that point. Analysis of the paths showed they were well approximated by a simple "active walker model" [12].

hough the case of movement through a snow-covered university campus bears little surface similarity to a Web searcher's "movement" on the Internet, there is a valuable analogy here. Just as one's passage through snow or grass is facilitated by others having forged a path before, the activity of Web searchers leaves behind valuable signals that can facilitate future users' search efforts. These signals are utilized by many modern search systems, both when they are left behind explicitly (as in collaborative tagging or when people share links on a social network) and, more commonly, when they are implicit. These implicit signals, formed as users issue queries and click on results, are integral to intelligent query suggestions and to the ranking of results on modern Web search engines. Our interest, however, lies in determining if this similarity can be more than a helpful analogy. There is evidence that information search and physical search are, from a cognitive perspective, not as disparate as they might seem.

In the mid-1990s, Pirolli applied optimal foraging theory — a theory of how organisms search for resources in a physical environment — to Web search with considerable success [13, 14]. More recent work [5] found that participants could be primed by a spatial search task to behave in predictably different ways on a subsequent mental search task. Participants first completed a simulated spatial foraging task on a computer, where hidden resources were either clumpy (distributed in dense clusters or patches) or diffuse (distributed uniformly in the environment), and then performed a "Scrabble" task where they were asked to form as many words as possible from a sequence of random sets of letters, with each successive letter set corresponding to a clump or patch of resources. Participants in the diffuse condition spent relatively little time with each set (patch) of letters before moving on to the next, while participants in the clumpy condition spent longer searching through each letter set. That is, being primed by search in a patchy spatial environment led to longer search within each patch in a mental environment, indicating that the two types of search, for physical resources and for information, may share deep underlying mechanisms. The authors hypothesized the existence of generalized cognitive search processes, and molecular and behavioral evidence [4, 15, 6] supports the hypothesis that evolved capacities for spatial search deeply influence the way we search in other domains. This suggests the usefulness of spatially-inspired data representations, like paths, for information search.

Path-based web search: Recently a few works [16–19] have studied algorithms inspired by physical spatial search to improve web search engine performance, modifying page ranking by enriching link data with collective intelligence information. For each page, the information about Web trails taken by other users (often called Web pheromones) is accumulated and used to modify the global rank of the page. This differs from our approach of showing the paths used by others, but leaving page ranks unchanged. In terms of methodology, only one other study [19] conducted a controlled experiment on real users as we did, but again, participants were not directly presented with search paths. These works differ from ours both in methodology and objectives. We are not presenting an algorithm inspirited by physical spatial searches for improving the relevance of results. Instead we present the user with real paths in order to eval-

uate how the social awareness of other users' search trails changes perception of difficulty of the tasks, utility of the system and ultimately engagement. ***Social web search*** While search systems in use today do not share search paths explicitly between users, the closely related idea of social Web search is a robust topic of study, covering a variety of approaches. "Social" can for instance refer to the fact that the answer returned by the search engine originates within the user's acquaintance network. An example is Google's Social Search, a service that allows users to customize search results based upon the people in their social network. In other cases the term "social" refers to sharing information with other individuals: Evans et al. [20] conducted a survey of 150 users on Amazon's Mechanical Turk service, and found that social interactions play an important role throughout the search process. They presented a canonical social model of user activities before, during, and after search, suggesting points in the search process where both explicitly and implicitly shared information may be valuable to individual searchers. Recently, Horowitz et al. [21] described Aardvark, a social search engine able to route a user's queries to the person in that user's extended social network most likely to be able to answer that question. We use the term "social" is a similar way: our tool presents users with the activity of the most skilled users with respect to each of the tasks to be completed during the study.

Search tool evaluation: Social search tools can be evaluated via two main criteria: effectiveness (and hence user satisfaction) and elicited engagement [22–25]. Often, shorter time to completion (i.e. the time spent on a search task) is used to assess effectiveness. In a social setting, however, time to completion is not always a good metric: Social interactions can lead to increased engagement, which can in turn increase time to completion, such as through distortions in the subjective perception of time [26, 22]. Since evaluations of social search tools depend on subjective measures, they are typically tested with user studies [22–24], which are limited in number of participants and constrained by the need for extended experience with a new tool [27]. Despite these problems, there is typically no viable alternative for testing users' subjective responses to search tools.

3 Methodology

Participants completed a sequence of search tasks either with social search information (BestSearcher paths condition) or without (baseline condition). We ran the baseline condition first, and used data from those participants to generate the search path information for the experimental BestSearcher condition. All participants completed the same set of search tasks (in randomized order) for one condition or the other. The study was administered in a web browser modified with the HCIbrowser extension[28], which allowed for display of path information, presentation of search tasks, collection of task responses, survey administration, and clickstream logging. The HCIbrowser extension incorporates an integrated display of search tasks in an in-browser toolbar, buttons for submitting responses and response URLs for each task, specification of the number of responses allowed for any task, and presentation of pre- and post- task questionnaires. Thus we could display a task description, prompt users to complete it and submit responses, and then administer a post-task questionnaire for each task. The extension also logs users' interactions with the browser, allowing us to later

reconstruct their search trajectories for analysis. The questionnaire capacities of the extension did not meet our needs, so we made source-code modifications such that the questionnaire interface displayed SurveyGizmo[1] questionnaires we developed separately.

Baseline condition participants used a standard version of the Yahoo! SERP, except for the removal of ads, "also try" suggestions, and other extraneous visual elements. For the experimental condition, in addition to these changes, we implemented a sidebar displaying the social search path data, and ran custom javascript that provided contextual highlighting of search results and data from the sidebar. We discuss how the paths were generated in section 3.1, but they all consisted of a sequence of queries and clicked URLs from a previous user or aggregation of users, with the URLs indented beneath the corresponding query. Participants in the social search condition could click any of the queries to issue that query in the search engine, or click a listed URL to go directly to that page. Further, any time a clicked URL appeared both in the current results page and the sidebar, it would be highlighted in both locations. Similarly, whenever the participant's current query (i.e. the query issued to the search engine and for which results are currently displayed) matched one from the sidebar, it would be highlighted (this occurred whether the user typed the query or clicked on it in the sidebar). These mechanisms provided clear indicators to participants of when they had encountered resources explored by a previous searcher. Figure 1 gives an example view of the interface.

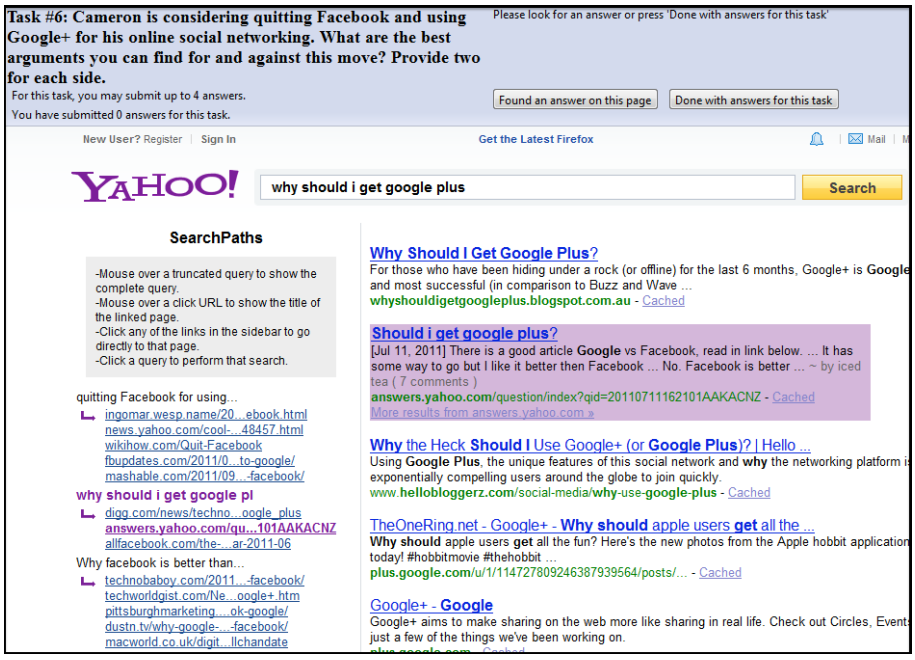

Fig. 1: Screenshot of experiment interface

[1] http://www.surveygizmo.com/

Table 1: Search tasks

"austria": You're on a backpacking tour of Europe, and will be stoping in Innsbruck, Austria, but unfortunately you'll only have a few hours to spend there. Find the two most interesting activities that could both be done in the 4 hours you'll have.
"boots": You're looking for a new pair of high-quality hiking boots for an upcoming backpacking trip. Narrow your options down to three pairs that you'll go try on at the store
"disney": Tammy is planning a two-day trip to Disneyland with her three-year-old daughter (who loves princesses) and is looking for the must-see attractions. She's already been to disneyland.com, and had little luck, so find three appropriate pages to help her in her trip planning.
"facebook": Cameron is considering quitting Facebook and using Google+ for his online social networking. What are the best arguments you can find for and against this move? Provide two for each side.
"indiana": Your assignment editor at the Indiana Daily Student asks you to write a news story about whether state budget cuts in Indiana are affecting financial aid for college and university students. Find the names of two people with appropriate expertise that you are going to interview for this story and save just the pages or sources that describe their expertise and how to contact them.
"linux": You want to try out the Linux operating system on your new computer, but don't know which of the many options to choose. Find the two best pages or articles that you can to help you decide which version would be best for you.
"metal": A friend wants to take up metal detecting as a hobby. Find the three best resources (books, online tutorials, videos, etc.) you can to get her started.
"tommy": Tommy has gained weight suddenly, feels fatigue, and has difficulty dealing with cold temperatures. Provide three possible diagnoses that explain his symptoms.

Participants were given search tasks that we deliberately selected to be both complex and minimally specific; that is, none of them had a particular set of "correct" answers. The goal was to use questions that would enable participants to utilize social information to aid their searches, without the social information leading to a single best answer for any question. Thus all tasks incorporated a level of subjectivity ("find the best...", etc.) and required multiple answers. Each task had a maximum number of allowable responses. Table 1 shows the tasks used. All participants completed each of the eight tasks. Participants also completed one free choice task, in which they could specify and complete a search task of their own choosing.

3.1 Generating search paths

To generate the social search path data displayed in the sidebar, we had to extract meaningful paths from users' search activity in the baseline condition. After comparing several options, we settled on "BestSearcher" paths for this study, which show the complete path (sequence of queries and clicks) of the "best" participant from the baseline condition for each task — requiring a measure of query success. Intuitively, a "best searcher" is one able to formulate many successful queries; that is, queries that allow the searcher to successfully complete the task in a short period of time. D. Ciemiewicz et al. [29] analyzed the quality of clicked documents as a function of dwell-time (i.e. the time a user spends reading the content of the page pointed to by the URL). The study shows that the probability that results are considered high quality increases with

the dwell-time, and that the probability of perceived "good" results increases to 77% for clicked documents with a dwell-time greater than 100 seconds. The time elapsed from the beginning of the search to the first long dwell-time click is thus a good indicator of how long it take the user to find a high-quality page. With this in mind, we ranked, on a task-by-task basis, the paths generated in the baseline condition according to a metric that takes into account the total number of queries (since the tasks require multiple answers, issuing more queries should increase the probability of finding more unique pages), the total number of long dwell-time clicks per query, and the inverse of the time required to reach the first long dwell-time result.

Because the average dwell-time for all clicks collected was 30 seconds, we considered the number of clicks with dwell-time greater than 30, 60, and 100 seconds. We then ranked the searchers according the following "BestSearcher Score":

$$Score = Q * \left(\sum_{i \in \{30,60,100\}} \frac{n_i * \bar{nc}_i * itf_i}{C} \right)$$

The score is given by the weighted sum of the total number of clicks n_i with dwell-time greater or equal to, respectively, 30,60, or 100 seconds, normalized by the total number of clicks C.[2] Thus clicks with greater dwell times are weighed more heavily.[3] The path followed by the baseline condition participant with the greatest score on this metric for each task was then used as the BestSearcher path for that task, such that all participants in the experimental condition saw that same path for any given task (but the source "best" searcher for paths varied from task to task). We also manually corrected for any typing errors in the displayed paths. Given that error correction is a standard feature of modern search engines, we did not want errors in the paths to degrade their perceived reliability or otherwise affect our results.

Though we ultimately did not utilize them, we also considered a variety of other possible paths that could be generated from the baseline data. In particular, we were curious if we could generate useful aggregate search paths from multiple participants' baseline data for display in the experimental condition. Two options we explored were PageRank-based [30] paths and top-queries paths. To generate PageRank-based paths, we built bipartite graphs using all the issued queries and clicked URLs for each task, generating paths built from sequences of queries and clicks ranked by PageRank score. Top-queries paths, on the other hand, consisted of the most popular queries and clicks across users for each task, ordered by their average position in baseline search sequences. We then used the PageRank [30] implementation available in the LAW library[4] to rank both queries and clicks by their PageRank scores, and generated paths consisting of

[2] Thus the number of 30-second dwell-time clicks includes all 60- and 100-second dwell-time clicks, and the 60-second dwell-time clicks include all 100-second dwell-time clicks.

[3] More specifically, the weights are given by the product of the average number $\bar{nc}_i$ of clicks preceding, and the inverse time itf_i to, the first long dwell-time click for $i \in 30, 60, 100$. Q is the total number of queries this searcher issued for the task. In this way, clicks with dwell-time greater than or equal to 30 but less than 60 seconds are weighted by $w_{30} = \bar{nq}_{30} * itf_{30}$. Clicks with dwell-time between 60 and 100 seconds are weighted by $w_{30} + w_{60}$. Clicks longer than 100 seconds are instead weighted by $w_{30} + w_{60} + w_{100}$.

[4] http://law.di.unimi.it/software/download/

the 10 queries with the highest PageRank scores (ranked by those scores in descending order) and their associated clicked URLs. For the top-queries paths, we simply determined the 10 most commonly issued queries by participants in the baseline condition for each task, and then generated an aggregate path consisting of those queries, ordered by their average position in the baseline participants' search sequences, and their associated clicks (also ordered by average position in the baseline search sequences).

However, we do not include our investigations of these aggregate paths here, as they did not capture the notion of "path" that we are focusing on. Even if such aggregate representations of user search histories did prove useful to later searchers, we cannot reasonably argue that those later searchers had been presented with a path (in the sense of the term motivating this paper), given the way the paths were constructed; none of these aggregated paths was guaranteed to delineate a path that any single earlier user actually traversed. Further work is necessary to develop methodologies to extract and represent aggregated path information from large numbers of search engine users.

3.2 Participants & Procedure

Participants were Indiana University undergraduates compensated with course credit. 26 female and 42 male students (12 female and 12 male in the baseline condition, 14 female and 30 male in the BestSearcher condition) participated. All were between 18 and 24 years old. Participants in the baseline condition were informed that they would be presented with a series of questions for which they should search the web for answers. Those in the experimental condition were given the same instructions, but were also told that they would have "access to information about how previous IU students have completed the search tasks." Participants first completed a practice trial to get familiar with the interface, then eight experimental search tasks (in randomized order). They then rated task difficulty, satisfaction with results, engagement with the task, and, for the BestSearcher condition, the usefulness of the search path information.

4 Results & Discussion

We focus here on determining if participants found the social path data engaging and/or helpful. Analyses discussed below reflect only participants who utilized the social path information (by clicking a query or URL) at least once (33 of 44).

Demographic and Web experience data All demographic and Web experience questions were answered on a 7-point Likert scale. Of interest here are participants' self-reported evaluations of their search expertise and experience, as captured by the following questions: "How much do you agree with the statement, 'I consider myself an expert search engine user?' " (1=strongly disagree, 7=strongly agree), "How often, on average, do you use search engines to find information online?"(1=never, 7=Five or more times per day), and "Do you consider yourself more of a producer or consumer of social media?" (1=Almost completely a consumer, 7=Almost completely a producer).

On average, participants rated themselves roughly in the middle of the expertise scale (mean 4.79), and a two-sample t-test showed no significant difference in expertise between the baseline and experimental conditions ($t(49) = .40, p =$

.69). Participants in the experimental condition did show a small but significant difference ($t(39) = 2.81, p < .01$) in how often they used search engines (mean 6.09) versus the baseline participants (mean 5.29). In light of the non-significant difference in expertise, and the fact that any response of 5 or greater on the response scale indicates using a search engine on a daily basis, these results do not appear problematic. The last question was meant to capture any between-group differences in propensity to consume social media that might bias our results, as individuals who identify more as social media consumers might be more likely to utilize the search path data. Participants on average slightly favored being social media consumers than producers (mean 3.18), and there was no significant difference between conditions ($t(45) = .85, p = .40$). There were no other systematic variations with respect to demographic data.

Subjective Measures: In all conditions, participants completed a questionnaire after each task consisting of five general subjective evaluations of the task, plus four task evaluations specific to the search path information in the BestSearcher condition (discussed below). All questions were on a 7-point Likert scale. For the general questions, participants rated how engaged they felt during the task, the task difficulty, how satisfied they were with the search results they found, and how realistic the task seemed both for themselves and for "people in general." Participants were also able to provide freeform comments on each of the tasks. Figure 2 summarizes the responses to the general subjective measures across both conditions and all tasks.

Unsurprisingly, we found a general pattern of anticorrelation between task difficulty and satisfaction (baseline: ($r(209) = -.59, p < .0001$), experimental: ($r(317) = -.60, p < .0001$), as well as weak but significant correlation between engagement and search satisfaction (baseline: $r(209) = .35, p < .0001$, experimental: $r(317) = .26, p < .0001$) across both conditions. As subjective difficulty went up, engagement went down in the Baseline condition ($r(209) = -.29, p < .0001$), but not in the BestSearcher condition. This suggests that social facilitation did ameliorate the negative effect of task difficulty on engagement. In contrast to our initial predictions, we found no significant difference in mean satisfaction or engagement between conditions. Problematically, participants did not report the experimental tasks to be of strong personal relevance, rating them on average below the midpoint of a Likert scale (i.e. disagreeing with the statement "This is a realistic search task for you in particular.").

There were relatively few significant differences between the conditions on a task-by-task basis, but some observations are noteworthy. First, there is clear heterogeneity in perception of the different tasks (particularly the "indiana" task, which was significantly more difficult than the others, and resulted in minimal satisfaction with search results.) The "free" (free-choice) task showed the opposite pattern, with low reported difficulty and high satisfaction. This suggests the search tasks we gave participants were more difficult than the tasks they would typically pick for themselves (though note that the free-choice task was always the last performed, so fatigue effects may contribute to the choice of relatively easy searches). These observations are consistent with a general pattern of anticorrelation between task difficulty and satisfaction in both the baseline ($r(209) = -.59, p < .0001$) and BestSearcher ($r(317) = -.60, p < .0001$) conditions.

With respect to engagement, the "free" task remains (unsurprisingly) an outlier, and we find a weak but significant correlation between engagement and search satisfaction across conditions (Baseline: $r(209) = .35, p < .0001$, Best

Searcher: $r(317) = .26, p < .0001$). There was a weak but significant anticorrelation between engagement and difficulty in the Baseline condition ($r(209) = -.29, p < .0001$), but no such significant correlation on the BestSearcher condition. This may suggest that social facilitation did ameliorate the negative effect of task difficulty on engagement.

The "linux" task is a distinctive case in the engagement plot, with participants showing significantly greater engagement in the BestSearcher condition compared to the baseline condition. It is not entirely clear why this is this the case, but one possible account is that the participants in the BestSearcher condition are more experienced Web users on average (in terms of how often they use search engines and may have been more likely to find the content of the "linux" task of interest. Consistent with this explanation is the fact that BestSearcher participants were more satisfied with their results on the "linux" task, and also picked more difficult tasks for the "free" tasks (perhaps indicating a general tendency to select more difficult and technical search tasks). Also possible is that most users across both conditions had little familiarity with the Linux operating system (anecdotally supported by participant comments like "I'm not sure what a Linux operating system is..." and "I don't know much about computer software."). Thus the search path data provided information aiding in a task that otherwise would have been challenging for people with no background knowledge on the topic ("I actually decided to use it [the SearchPaths tool] this time because it was harder to find the information that I was looking for.").

In contrast to our initial predictions, we found no significant difference in mean satisfaction or engagement between conditions. This motivated a finer-grained analysis of these measures, to be discussed in Section 4.1.

Behavioral measures: We also collected a number of behavioral measures from participants, summarized in Figure 4. All analyses were performed by parsing the raw log data from the HCIbrowser extension and converting it to a "search mission" format (similar to that used in, e.g., [31]). This consists of an ordered sequence of clicks and queries, along with the time in seconds between each event. For example:

```
Q:indiana university financial aid office 2
C:p=1:http://www.indiana.edu/~sfa/ 21
C:p=1:http://www.indiana.edu/~sfa/office/... 88
Q:indiana financial aid cuts 3 ...
```

The key measures we extracted from these logs were the time to complete each task (see Figure 4a), the mean dwell time on each task (i.e. the average time spent on each clicked page, 4b), proportion of trials successfully completed on each task (4c), and total number of search events (i.e sum of queries and clicked URLs for each task, 4d). To account for the fact that different tasks required participants to submit different numbers of answers, we normalized the total task time and total number of search events (4a and 4d) by the number of required submissions (see Table 1). The proportion of successful trials reported in Figure 4c was determined manually, by simply verifying that participants both (a) submitted as many answers as were required for the task, and (b) submitted answers that were relevant to the task (relevance was determined at a basic level, only discounting clear "garbage" submissions with no relevance to the task). Each participant thus received a score equal to the proportion of required submissions that were submitted and relevant to the task.

As with the subjective measures, we found notable heterogeneity across tasks, presumably owing to variance in their difficulty and interestingness (Figure 2),

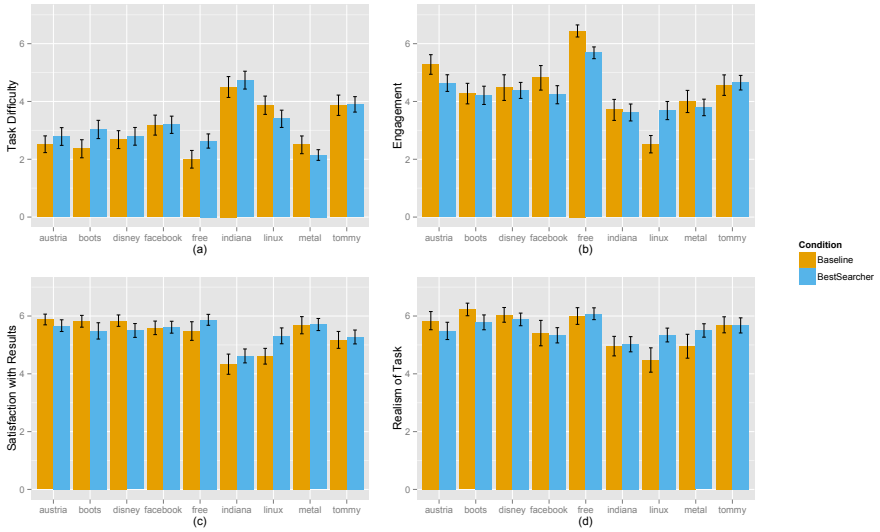

Fig. 2: Summary of subjective measures by condition and task. All ratings were made by participants on a 7-point Likert scale. Shown are responses to questions: "This was a difficult search task." (a, 1=strongly disagree, 7=strongly agree), "This was an engaging/interesting search task." (b, 1=strongly disagree, 7=strongly agree), "How satisfied were you with the answers you submitted for this search task?" (c, 1=very dissatisfied, 7=very satisfied), and "This seemed like a realistic search task for you in particular. That is, it is a task you can imagine yourself doing." (d, 1=strongly disagree, 7=strongly agree). Error bars show +/- 1 standard error. Responses to "This seemed like a realistic search task. That is, it is the kind of task you can imagine people in general actually performing." are omitted, as they did not figure into further analyses.

but little in the way of between-conditions differences. The data suggest a trend towards faster completion times and fewer total search events when path information was available, with the notable exception of the "indiana" task, which required significantly more time and search events in the BestSearcher condition compared to baseline. This may stem from the difficulty of the task (highest subjective difficulty rating), along with the possibility that the information in the sidebar was not particularly useful, but participants still explored the social data in detail in an effort to solve the difficult task. This task also had the greatest proportion of activity originating from the sidebar (i.e. the total number of times participants clicked URLs or queries in the sidebar, divided by the their total number of search events for that task, see Figure 3) across participants. There is, in fact, a weak but significant ($r(209) = -.29, p < .0001$) correlation between the proportion of activity originating from the sidebar and the perceived difficulty of tasks, indicating that participants relied more on socially available data when search tasks were more challenging..

There was a weak but significant correlation between time to completion (Figure 3a) and task difficulty (baseline: $r(184) = .34, p < .0001$, BestSearcher: $r(279) = .25, p < .0001$), but no other notable correlations between the behavioral and subjective measures. The "facebook" and "tommy" tasks appear to be outliers with respect to mean dwell time (Figure 3b), for reasons that are not completely clear. One possible account is that these tasks required more evalua-

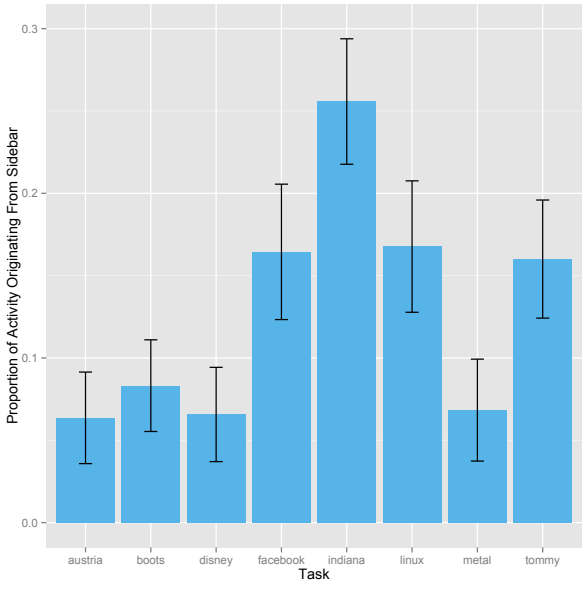

Fig. 3: Mean proportion of total search activity (i.e. sum of clicked URLs and issued queries) originating from the SearchPaths sidebar by task in the BestSearcher Condition. Error bars show +/- 1 standard error.

tion of content than the other tasks (the "facebook" tasks required finding two arguments each in favor of opposing sides of a debate, and the "tommy" task required finding three unique diagnoses for an illness).

Evaluation of search paths: Participants in the BestSearcher condition, in addition to the questions outlined above, responded to four statements evaluating the SearchPaths tool during the post-task questionnaire, all on a 7-point Likert scale (1=strongly disagree, 7=strongly agree). These statements rated the usefulness of the tool, whether or not it made the the task more interesting/engaging than it would have been otherwise, whether or not participants actually used the tool, and whether it allowed them to complete the task more quickly than they would have otherwise. Responses to these four questions are summarized in Figure 5. Responses hovered around the middle of the response scale on average, indicating that participants found the search paths moderately helpful overall. There appears to be a general pattern of paths being more positively evaluated on the more difficult tasks, though the only measure here that correlates (weakly) with difficulty in a statistically reliable way is participants' reported usage levels ($r(258) = .23, p < .001$). These responses were not particularly strongly aligned with the respective behavioral measures we collected, though; ratings of how much participants actually used the paths, for instance, had only a weak correlation ($r(258) = .22, p < .001$) with their total sidebar activity (i.e. sum of clicked queries and URLs from the sidebar) and their ratings of how much more interesting/engaging the sidebar data made the task were only weakly correlated ($r(253) = .27, p < .0001$) with their general ratings of how interesting or engaging the task was. The greatest reported benefit is of perceived improvement in task completion time (m=4.09), while the weakest effect is of improvement in engagement (m=3.55).

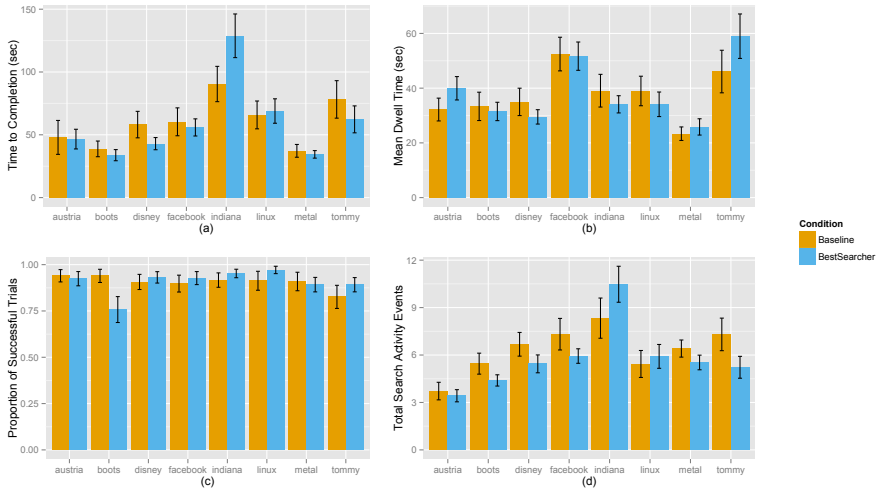

Fig. 4: Summary of behavioral measures by condition and task. (a): Mean time to complete task (seconds). (b): Mean dwell time (seconds). (c): Mean proportion of successfully completed trials. (d): Total search activity (number of clicks + number of queries). (a),(b), and (d) normalized by the number of responses required for each task. Error bars show +/- 1 standard error.

Notable here is that all responses to the search paths evaluation questions were moderately to highly inter-correlated ($r > .6, p < .0001$ in all pairwise correlations). So, even though their subjective responses may not correlate well with their behavioral patterns, these results indicate (consistent with our hypotheses) that a useful search tool is one that enhances both engagement and the speed with which a user can achieve his or her search goal. Unexpected here is the weak correlation between how much participants reported using the sidebar by actually clicking queries and URLs and the true use of the sidebar we logged experimentally. The unexpected low correlation between perceived and actual use could have come about because participants had an inflated sense of how much they used the sidebar when they found the sidebar path information to be useful. Nevertheless, these data suggest that the SearchPaths tool may have been of help to participants in ways not apparent from our collected behavioral measures.

These responses were not particularly strongly aligned with the respective behavioral measures we collected, though; their ratings of how much they actually used the tool, for instance, had only a weak correlation ($r(258) = .22, p < .001$) with their total sidebar activity – sum of clicked queries and URLs from the sidebar – and their ratings of how much more interesting/engaging the sidebar data made the task were only weakly correlated ($r(253) = .27, p < .0001$) with their general ratings of how interesting or engaging the task was. An analysis of variance indicates no differences between search tasks with respect to the engagement questionnaire item ($F(7, 247) = 1.05, p = .40$). The other questionnaire items showed greater variance between tasks, but no immediately apparent systematic variation with other variables. The one exception was a weak but significant correlation between how much participants reported using the sidebar and the difficulty of the task ($r(258) = .23, p < .001$).

	TI	UL	E	U
Time Imp.	-	0.77(254)	0.68(253)	0.83(253)
Usage Level	0.77(254)	-	0.61(254)	0.75(257)
Engagement	0.68(253)	0.61(254)	-	0.71(255)
Usefulness	0.83(253)	0.75(257)	0.71(255)	-

Table 2: Correlation table for sidebar evaluation measures. Pearson's R is reported for each pairwise correlation (degrees of freedom in parentheses). For all cases, $p < .0001$. Note that "Engagement" here corresponds to the questionnaire item on the engagement due to the SearchPaths tool, not general engagement with the search task. Abbreviated column labels correspond to the row labels (e.g. "UL" = "Usage Level").

Path-following measures Participants definitely utilized the SearchPaths tool, but did they do so in a manner indicative of path-following? We tested two different measures to explore this, both based upon the sequence of clicks and queries issued by a participant and the sequence visible to him or her in the sidebar. The first (the "path score") counted all instances of sequential events (clicks or queries) from the participant's path that mirrored the sidebar path, such that any time the participant performed two or more actions in the same uninterrupted sequence as in the sidebar, her path score was incremented. The second measure, a modified Spearman's rank coefficient calculation between issued and sidebar-listed actions, was slightly more lenient (because it could capture instances of interrupted sequences of path following, e.g., issuing two queries from the sidebar in the listed order, but clicking other results in between that were not visible in the sidebar). To calculate this, we took the sequences of events from the sidebar and from the participant's search mission, and removed all unique items, leaving only the overlapping search events in each sequence while maintaing their original, respective ordering. We then calculated a modified Spearman's rank coefficient on the two sequences, weighted by the total length of the overlapping sequence.

Scores were calculated for each participant and task in the BestSearcher condition. Both scores indicated that participants were not using path data in the same order that they found it in the sidebar (98% of scores for the path score measure, and 91% of scores for the Spearman's rank measure, were zero). While this result suggests that participants did not follow search trajectories that mirrored the paths presented to them, more sensitive measures of ordered path following still need to be developed, and it is not in itself a reflection on the usefulness of path-based data overall.

This hypothesis is supported by the fact that the "indiana" task had the largest number of discrete search events (sum of queries and clicked URLs) of all tasks (Figure 4d), and the largest proportion of search events originating from the sidebar (i.e. the total number of times participants clicked URLs or queries in the sidebar, divided by the their total number of search events for that task, see Figure 3).

4.1 Regression Analysis

Not having found many clear patterns while looking at simple correlations and between-conditions comparisons, we elected to perform a more in-depth multiple

regression analysis to explore how engagement and satisfaction were related to the measures discussed above.

Before proceeding to the analysis we checked whether the data collected from the user study were sufficient to obtain meaningful results. After cleaning and filtering, we extracted 233 records for modeling engagement and 245 records for modeling satisfaction, where each record corresponds to a particular participant's data for a particular search task (the different numbers arose from a minority of participants who either completed only some of the search tasks, or who encountered technical errors in the recording of their survey responses). We obtained a minimum required sample size of 201 for the multiple regression analysis, given a desired probability level $p = .05$, 15 predictors, a priori statistical power level of 0.8 and a medium effect size $f^2 = 0.10$ [32].

The analysis was done using the behavioral measures previously described, several measures defining properties of the tasks, and two binary variables dif and $reaP$ (see Table 3). These final two variables are binarized versions of participants' subjective ratings of task difficulty and task realism (our previous analyses suggested that these subjective measures were the clearest predictors of satisfaction and engagement). To binarize the values, response values greater than 4 on the Likert scale for the corresponding questions were coded as ones, and values of 4 or less as zeroes. The predicted values of engagement and satisfaction were obtained in a similar way: we consider the user satisfied (or engaged) if he declared a value between 5 and 7, unsatisfied (or not engaged) otherwise.

Predictor	Description
SBsubs	Number of URLS submitted as answers originating from the sidebar
index	Sequence order of task (e.g. 3 = third task completed)
meanDwell	Mean dwell time across all clicked URLs
nC	Total number of clicked URLs
nQ	Total number of queries issues
pathScore	Path score
rank	Spearman rank score
sbC	Total number of clicks originating from the sidebar
sbQ	Total number of queries originating from the sidebar
subs	Total number of task answers submitted
success	Proportion of required submissions submitted
taskLength	Total number of items (queries and clicks) displayed in sidebar
totalTime	Total time in seconds spent on task
dif	Binarized subjective difficulty rating
reaP	Binarized subjective realism/interestingness rating

Table 3: Summary of predictors used in regression analysis.

Table 4 reports the results of the satisfaction analysis. As measure of the model fit, we computed the significance of the overall model. Chi-square analysis demonstrates that the model as a whole fits significantly better than an empty model ($\chi^2(15, N = 245) = 75.4, p < .0001$). The results indicate that the fifteen independent variables explained 45% of the variance. However, only four variables contributed significantly to the prediction of satisfaction: the difficulty of the task dif ($\beta = -.79, p < .0001$), the proportion of the successful

trials *success* ($\beta = 1.63, p < .01$), the number of queries *sbQ* selected from the SearchPaths tool ($\beta = -.887, p = .03$) and the sequential number of the task *index* ($\beta = .21, p = .06$). Not surprisingly, having a difficult task, versus an easy one, decreases the log odds of being satisfied by 0.79, even if for the three most difficult tasks (facebook, indiana and tommy), users in the Best Searcher condition are slightly more satisfied than users in the baseline. Satisfaction obviously increases with the fraction of relevant answers submitted ($\beta = 1.663$). For every one unit change in the total number of queries originating from the sidebar *sbQ*, the log odds of satisfaction decreases by 0.887. As we hypothesized, for hard tasks, it may be the case that information in the sidebar was not particularly useful, but that participants still explored the social data in detail in an effort to complete the more difficult tasks. It is not clear why satisfaction increases with the sequence order of the task *index*. We can hypothesize that after using the tool for a while, users become more familiar with it and hence are able to quickly find results. There is indeed a weak but significant anticorrelation between $totalTime$ and *index* ($r(280) = -.26, p < .0001$).

Predictor	β	z value	$Pr(>\|z\|)$
(Intercept)	2.2002044	1.810	0.0703 .
SBsubs	-0.1111520	-0.345	0.7302
index	0.2132609	1.849	0.0644 .
meanDwell	0.0128208	0.941	0.3468
nC	0.0150186	0.074	0.9413
nQ	0.2547072	1.537	0.1242
pathScore	0.7192973	0.784	0.4331
rank	0.4172694	1.122	0.2620
sbC	0.0415489	0.220	0.8260
sbQ	-0.8871162	-2.189	0.0286*
subs	-0.0857400	-0.345	0.7304
success	1.6631313	1.968	0.0490 *
taskLength	-0.0633546	-1.388	0.1652
totalTime	-0.0007219	-0.177	0.8598
dif	-0.7904588	-5.155	2.54e-07***
reaP	0.1831004	1.550	0.1212

Signif. codes: *** 0.001, ** 0.01, * 0.05, . 0.1

Table 4: Multiple regression analysis for satisfaction

Table 5 reports the results of the engagement analysis. The model as a whole is again significantly better than a null model ($\chi^2(15, N = 233) = 61.34, p < 0.0001$).. The fifteen independent variables explained 31% of the variance. In this case only two variables contributed significantly to the prediction of satisfaction: the difficulty of the task dif ($\beta = 0.25, p < .01$), the interestingness $reaP$ ($\beta = .50, p < .001$). The results suggest that engagement increases with the difficulty of the task; having a difficult task, versus an easy one, increase the log odds of being engaged by 0.25. As suggested by previous work [23, 24], interestingness is a proxy for engagement; having a realistic task increases the log odds of being engaged by 0.5. We performed a similar regression analysis (not reported) predicting users' engagement with the SearchPaths tool (responses to "The SearchPaths information made this task more interesting/engaging than

it would have been otherwise"), and found a similar pattern of significant predictors.

Predictor	β	z value	$Pr(>\|z\|)$
(Intercept)	-3.1529103	-3.182	0.00146 **
SBsubs	0.3639378	1.504	0.13269
index	-0.0380183	-0.474	0.63569
meanDwell	0.0056154	0.772	0.44027
nC	-0.0146414	-0.122	0.90323
nQ	0.0241263	0.194	0.84580
pathScore	0.9583210	1.477	0.13956
rank	0.0299222	0.110	0.91235
sbC	-0.2205454	-1.559	0.11897
sbQ	-0.2126343	-0.599	0.54899
subs	-0.1528504	-0.847	0.39682
success	1.3448507 s	1.714	0.08661 .
taskLength	-0.0127488	-0.395	0.69309
totalTime	0.0003533	0.154	0.87793
dif	0.2550790	2.620	0.00879 **
reaP	0.5052121	5.754	8.7e-09 ***

Signif. codes: *** 0.001, ** 0.01, * 0.05, . 0.1

Table 5: Multiple regression analysis for engagement

5 Conclusions

We have made a theoretical case for leveraging cognitive science research linking spatial and information search in the development of social search aids, specifically in the context of sharing search paths between users. We also presented a preliminary effort at designing and testing a simple system with such social functionality. In the end, our empirical results do not allow for strong conclusions to be drawn from our user study, but our methods will likely be useful in future comparative work that considers other path-based search tools.

Multiple regression analysis on our data reveals that use of the sidebar (with respect to queries, but not to clicks) was a significant predictor of satisfaction with participants' search results, but not engagement. This validates the usefulness of the SearchPaths tool to some extent, and the overall modest results with respect to engagement and satisfaction may be in part due to the sidebar facilitating participants' searches implicitly (i.e. by providing useful information, even when they did not directly interact with it).

Regression analysis also highlighted the importance of task relevance ("This seemed like a realistic search task for you in particular.") to participants. This was the strongest predictor of engagement, and draws attention to the importance of using search tasks that participants find personally relevant in research of this kind. We of course were unable to generate search path data "on the fly" that would be relevant to any arbitrary search goal, but exploring automated methods for doing so is an important future direction of research. It is a complex problem, as it involves both identifying a user's search goal (in a broader

sense than can typically be determined from any given query) and culling and aggregating search log data so as to generate search paths that are relevant to that search goal. Privacy is another consideration here, as sharing complete search paths can reveal more personally identifiable information than most current search assistance technologies do. Future work along these lines will likely take inspiration from opt-in models like Microsoft Research's "SearchTogether" [33] or explicit sharing models like the Bing search engine's Facebook integration and Google's social search.

Our study faced a number of limitations, many stemming from its relatively small scale. While our hypothesis that path information should be helpful for moderately complex search tasks like those we assigned may hold true, we doubt such an effect can be clearly measured when study participants are presented with tasks in which they have little intrinsic interest or stake in the outcome. Subsequent work on such search tools must ensure that participants are provided with tasks that capture their interest in an ecologically valid manner. Further, larger-scale work is also required to determine how to aggregate path information from many searchers and how to effectively present that information to users.

Our study does nonetheless suggest that path information may be useful to Web searchers. Research in cognitive science has revealed that human search mechanisms in non-physical environments remain deeply connected to evolved foraging and spatial search processes, and work of this nature thus can inform both our understanding of how individuals interact with information search systems, and the design of tools to facilitate search in such environments. Our study focused on one particular application, namely applying notions of spatial path following to a Web search environment. Our hope is that this work can serve as inspiration for further exploration of how path information can be leveraged in Web search, and for applications of cognitive science research about search behavior to the improvement of online information search systems more generally.

References

1. Theraulaz, G., Bonabeau, E.: A brief history of stigmergy. Artificial life **5**(2) (1999) 97–116
2. Helbing, D., Molnar, P., Farkas, I.J., Bolay, K.: Self-organizing pedestrian movement. Environment and Planning B: Planning and Design **28**(3) (2001) 361–384
3. Goldstone, R.L., Roberts, M.E.: Self-organized trail systems in groups of humans. Complexity **11**(6) (2006) 43
4. Hills, T.T.: Animal foraging and the evolution of goal-directed cognition. Cognitive Science: A Multidisciplinary Journal **30**(1) (2006) 3–41
5. Hills, T.T., Todd, P.M., Goldstone, R.L.: Search in external and internal spaces: Evidence for generalized cognitive search processes. Psychological Science **19**(8) (2008) 802–808
6. Todd, P.M., Hills, T.T., Robbins, T.W., eds.: Cognitive Search: Evolution, Algorithms, and the Brain. MIT Press (2012)
7. Glenberg, A.M.: Embodiment as a unifying perspective for psychology. Wiley Interdisciplinary Reviews: Cognitive Science **1** (2010) 586–596
8. Goldstone, R.L., Gureckis, T.M.: Collective behavior. Topics in Cognitive Science **1**(3) (2009) 412–438
9. Goldstone, R.L., Roberts, M.E., Mason, W., Gureckis, T.: Collective search in concrete and abstract spaces. Decision Modeling and Behavior in Complex and Uncertain Environments (2008) 277–308
10. Goldstone, R.L., Roberts, M.E., Gureckis, T.M.: Emergent processes in group behavior. Current Directions in Psychological Science **17**(1) (2008) 10

11. Goldstone, R.L., Jones, A., Roberts, M.E.: Group path formation. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans **36**(3) (2006) 611–620
12. Helbing, D., Schweitzer, F., Keltsch, J., Molnar, P.: Active walker model for the formation of human and animal trail systems. Physical Review E **56**(3) (1997) 2527–2539
13. Pirolli, P., Card, S.: Information foraging in information access environments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. (1995) 51–58
14. Pirolli, P.: Information foraging theory: Adaptive interaction with information. Oxford University Press, USA (2007)
15. Hills, T.T., Todd, P.M., Goldstone, R.L.: The central executive as a search process: priming exploration and exploitation across domains. Journal of Experimental Psychology: General **139**(4) (2010) 590
16. Wu, J., Aberer, K.: Swarm intelligent surfing in the web. In: Web Engineering. Springer (2003) 431–440
17. Kantor, P.B., Boros, E., Melamed, B., Menkov, V.: The information quest: A dynamic model of user's information needs. In: Proc. of the American Society For Information Science Annual Meeting. Volume 36. (1999) 536–545
18. Furmanski, C., Payton, D., Daily, M.: Quantitative evaluation methodology for dynamic, web-based collaboration tools. In: System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on, IEEE (2004) 10–pp
19. Gayo-Avello, D., Brenes, D.J.: Making the road by searching - a search engine based on swarm information foraging. CoRR **abs/0911.3979** (2009)
20. Evans, B., Chi, E.: Towards a model of understanding social search. In: Proceedings of the 2008 ACM conference on Computer supported cooperative work, ACM (2008) 485–494
21. Horowitz, D., Kamvar, S.: The anatomy of a large-scale social search engine. In: Proceedings of the 19th international conference on World wide web, ACM (2010) 431–440
22. Attfield, S., Kazai, G., Lalmas, M., Piwowarski, B.: Towards a science of user engagement (position paper). In: WSDM Workshop on User Modelling for Web Applications. (2011)
23. McCay-Peet, L., Lalmas, M., Navalpakkam, V.: On saliency, affect and focused attention. In: Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, ACM (2012) 541–550
24. O'Brien, H.: Exploring user engagement in online news interactions. Proceedings of the American Society for Information Science and Technology **48**(1) (2011) 1–10
25. O'Brien, H.L., Toms, E.G.: Is there a universal instrument for measuring interactive information retrieval?: The case of the user engagement scale. In: Proceedings of the Third Symposium on Information Interaction in Context. (2010) 335–340
26. O'Brien, H., Toms, E.: The development and evaluation of a survey to measure user engagement. Journal of the American Society for Information Science and Technology **61**(1) (2009) 50–69
27. Hoeber, O., Yang, X.D.: A comparative user study of web search interfaces: HotMap, Concept Highlighter, and Google. In: IEEE/WIC/ACM International Conference on Web Intelligence. (2006) 866–874
28. Capra, R.: HCI Browser: A Tool for Administration and Data Collection for Studies of Web Search Behaviors. Design, User Experience, and Usability. Theory, Methods, Tools and Practice (2011) 259–268
29. Ciemiewicz, D., Kanungo, T., A., L., M., S.: On the use of long dwell time clicks for measuring user satisfaction - with application to web summarization. Technical report, Yahoo! Labs (2010)
30. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer networks and ISDN systems **30**(1-7) (1998) 107–117

31. Donato, D., Bonchi, F., Chi, T., Maarek, Y.: Do you want to take notes?: Identifying research missions in Yahoo! search pad. In: Proceedings of the 19th international conference on World wide web. (2010) 321–330
32. J, C.: Statistical Power Analysis for the Behavioral Sciences (2nd Edition). Lawrence Earlbaum, Hillsdale, NJ. (1988)
33. Morris, M., Horvitz, E.: SearchTogether: an interface for collaborative web search. In: Proceedings of the 20th annual ACM symposium on User interface software and technology. (2007) 3–12

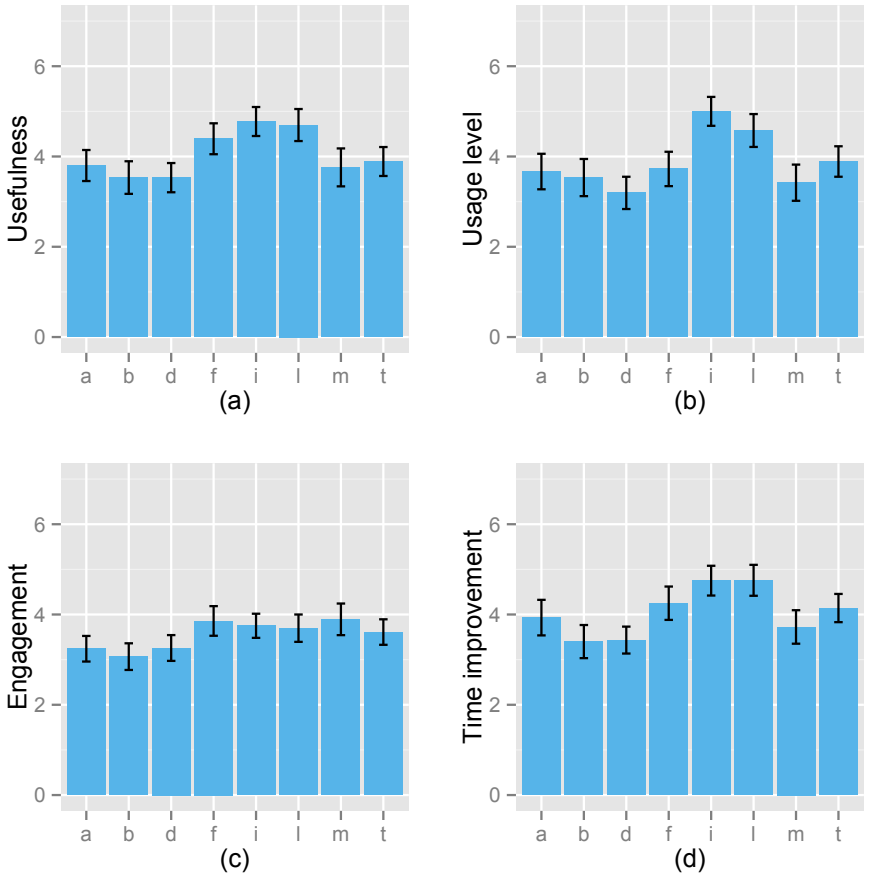

Fig. 5: Summary of subjective evaluations of SearchPath data in the BestSearcher condition. All ratings were made by participants on a 7-point Likert scale. Shown are responses to questions: "The SearchPaths information was useful to me while completing this task." (a), "I actually utilized the SearchPaths information by clicking on queries and links from the sidebar." (b), "The SearchPaths information made this task more interesting/engaging than it would have been otherwise" (c), and "The SearchPaths information helped me to complete the task in a shorter time than I would have been able to otherwise" (d). Error bars show +/- 1 standard error. Task codes are abbreviated here to the first letter only.