Using Trails to Support Users with Tasks of Varying Scope

Robert Capra and Jaime Arguello
University of North Carolina at Chapel Hill
{rcapra,jarguello}@unc.edu

ABSTRACT

A *search trail* is an interactive visualization of how a previous searcher approached a related task. Using search trails to assist users requires understanding aspects of the task, user, and trails. In this paper, we examine two questions. First, what are task characteristics that influence a user's ability to gain benefits from others' trails? Second, what is the impact of a "mismatch" between a current user's task and previous user's task which originated the trail? We report on a study that investigated the influence of two factors on participants' perceptions and behaviors while using search trails to complete tasks. Our first factor, *task scope*, focused on the scope of the task assigned to the participant (broad to narrow). Our manipulation of this factor involved varying the number of constraints associated with tasks. Our second factor, *trail scope*, focused on the scope of the task that originated the search trails given to participants. We investigated how task scope and trail scope affected participants' (RQ1) pre-task perceptions, (RQ2) post-task perceptions, and (RQ3) search behaviors. We discuss implications of our results for systems that use search trails to provide assistance.

ACM Reference Format:
Robert Capra and Jaime Arguello. 2019. Using Trails to Support Users with Tasks of Varying Scope. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331290

1 INTRODUCTION

A *search trail* is an interactive visualization that conveys how a previous searcher approached the same (or a related) search task. A search trail may contain a previous searcher's queries, pages visited, pages bookmarked, and annotations. Prior work has found that search trails can provide a wide range of benefits, including helping users learn new terminology, discover useful resources, discover new search strategies, and confirm information found on their own [3]. Prior work on search trails has focused on a wide range of issues, such as investigating factors that influence search trail quality [11], developing techniques for predicting search trails for a current user [5, 7], and studying the challenges involved with benefiting from other people's search trails [3].

While search trails have been found to provide benefits, open questions remain. For example, what are task characteristics that may influence a user's willingness to engage with search trails and their ability to gain benefits from them? In this paper, we investigate the effects of the *task scope* on a user's perceptions and behaviors while interacting with search trails. Additionally, using search trails to support users requires deciding *which* trails to display. Ideally, the system should display trails representing alternative approaches to the *same* search task. In practice, however, it is unlikely for two search tasks to have all of the same characteristics. Thus, an important question is: What is the impact of a "mismatch" between a user's current task and the task which originated a trail displayed to provide support. In this paper, we investigate "mismatches" related to the *task scope*. Specifically, are users able to gain benefits from trails on the same underlying topic, but different scope?

We report on a large-scale crowdsourced study that investigated the influence of two factors (task scope and trail scope) on participants' pre-/post-task perceptions and behaviors while interacting with search trails. Participants were given search tasks that required them to find information and construct a written response. To find information, participants were *only* given access to a tool referred to as the SearchGuide (SG). The SearchGuide tool displayed the search trails from three previous searchers (from a previous study [4]) who completed the same task or a related task (i.e., same topic, different scope). Each search trail displayed the previous searcher's queries, clicked results, and bookmarked pages.

To control for other task characteristics, participants completed *comparative* tasks, which involve comparing items (e.g., models of cars) across different dimensions (e.g., price, gas mileage, reliability). Our manipulation of task scope (4 conditions) involved specifying exact items and/or dimensions for participants to consider. Our broadest tasks specified no items nor dimensions, and our most specific tasks specified two items and one dimension. Our manipulation of trail scope (2 conditions) involved manipulating which search trails were included in the SG. In the [broad | narrow] trail condition, the trails came from previous searchers who completed the [broadest | narrowest] task version. Our study investigated three research questions:

RQ1: What is the effect of the *task scope* on participants' pre-task perceptions? We focus on perceptions about the task specificity, expected difficulty, and expectations about the SearchGuide.

RQ2: What is the effect of the *task scope* and *trail scope* on participants' post-task perceptions? We focus on perceptions about the quality and familiarity of the search trails and the task difficulty.

RQ3: What is the effect of the *task scope* and *trail scope* on participants' behaviors while interacting with the search trails provided in the SearchGuide?

2 RELATED WORK

Using Trails to Support Users: Early work by Bilenko and White [1] explored the potential information value of search trails. Using learning-to-rank, the authors found improvements in retrieval performance by generating training data from *full trails* versus only

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331290

SERP-level interactions (i.e., clicks and skips). White et al. [8] developed a "popular destinations" tool, which displayed pages that were frequently the trail end-point for similar queries. In a study, users reported benefits from using the tool during exploratory search tasks [8]. White and Huang [9] compared the usefulness of *full* search trails versus only portions of a trail, and found that full trails provided greater topical coverage, diversity, and novelty.

Using search trails to support users requires predicting trail quality and relevance. With respect to quality, Yuan and White [11] found that search trails created by domain experts (vs. novices) contained more relevant information, more factual (vs. subjective) information, and a more logical transition from general to specific searches. With respect to relevance, Singla et al. [7] evaluated different algorithms for ranking search trails in response to an initial query-click pair. Hendahewa and Shah [5] proposed a simple trail-matching algorithm based on query edit-distance.

Task Complexity and Scope: Much research has focused on characterizing search tasks along different dimensions [6], including complexity [10]. One influential view of task complexity is through the lens of *a priori* determinability [2]. In this respect, a complex task is one with great uncertainty about the form of the solution, requirements, and processes involved.

In a previous study [4], we investigated the relationship between task complexity and scope. As in this paper, we manipulated the scope of comparative tasks by specifying items and/or dimensions for participants to consider. Initially, we expected narrower, more well-defined tasks to be more determinable (less complex). However, narrowing the task by specifying items had a *different* effect than by specifying dimensions—specifying two items made the task less complex and specifying one dimension made the task more complex. An analysis of participants' queries [4] suggests that searching for items is easier than dimensions, for several reasons. First, items tend to be concrete nouns, whereas dimensions tend to be abstract concepts (e.g., durability) with more varied language. Secondly, dimensions tend to be subjective criteria (e.g., ease of use), requiring users to synthesize opinions. Searching for dimensions introduced uncertainty into the task, increasing its complexity.

3 TASK SCOPE MANIPULATION

Our task manipulation involved narrowing/broadening the scope of comparative tasks by including/excluding specific items and dimensions for participants to consider. We developed 17 task topics and 4 task versions per topic (68 task descriptions). Each task description included a backstory and a final information request that was manipulated based on our four task versions:

Unspecified (U): no items or dimensions specified. "Your sister has started gardening recently and has asked you to help her choose the right fertilizer for her garden. What are different types of garden fertilizers and how do they differ?"

Dimension (D): specified one dimension to consider, but no items. "Your sister... What are different types of garden fertilizers and how do they differ in terms of their **nutrient content**?"

Items (I): specified two items to compare, but no dimension. "Your sister... How do **organic** fertilizers differ from **chemical** fertilizers for garden use?"

Both (B): specified two items to compare and one dimension. "Your sister... How do **organic** fertilizers differ from **chemical** fertilizers for garden use in terms of their **nutrient content**?"

Figure 1: SearchGuide

4 SEARCHGUIDE (SG)

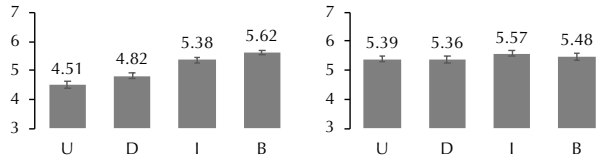
Study participants were only given access to the SearchGuide (SG) tool to find information. The main interface contained four elements: (1) the search task description, (2) the SG, (3) a textbox for the participant's response, and (4) a "done with task" button. The SG (Figure 1) displayed three search trails on different tabs, labeled "Path 1" to "Path 3". Each trail displayed the queries issued by the previous participant in chronological order. Clicking on a query expanded an accordion control that displayed the search results clicked and pages bookmarked for that query (marked with a "thumbs up" symbol). Each bookmark included a "reason bookmarked" provided by the previous searcher.

In the current study, there were 34 versions of the SG (17 task topics × 2 trail versions per topic). The search trails displayed in the SG came from participants from a previous study [4]. In that study, participants were assigned task versions U and B and asked to find and bookmark pages using a BingAPI-based search engine.

5 USER STUDY

To investigate our three research questions, we conducted a crowdsourced study using Amazon Mechanical Turk (MTurk). Participants in the study were given comparative search tasks that required finding information and constructing a written response. To find information, participants were instructed to only use the SearchGuide (SG). We manipulated two experimental variables: task version (four conditions) and trail version (two conditions). The task version variable manipulated the *scope* of the comparative task: U, I, D, or B. The trail version variable manipulated the scope of the search trails included in the SG. In the broad trail condition, the trails came from participants in a previous study [4] who completed task version U (no items or dimensions) for the same task topic. In the narrow trail condition, the trails came from participants who completed task version B (two items, one dimension).

In total, the study had 136 experimental conditions: 17 task topics × 4 task versions per topic × 2 trail version conditions. For each experimental condition, we posted 10 redundant HITs (1,360 total). Our HITs were implemented as "external HITs", allowing us to control the assignment of experimental units to participants. Experimental units were assigned randomly, except that participants were not allowed to do multiple HITs for the same task topic. While we published 1,360 total HITs, we stopped data collection once each experimental condition had at least 8 redundant HITs completed. Ultimately, we gathered data for 1,234 HITs (from 557 workers). MTurk workers spent on average 16 minutes working on each task and were paid US$1.25 per HIT.



(a) Specificity (U < D < I < B) **(b) SG Expectations (D < I)**

Figure 2: The effects of task version on pre-task factors.

Upon accepting the MTurk HIT, participants were given instructions and a video describing the SearchGuide (SG) as an "interactive visualization that displays the search trails or *paths* followed by three searchers who completed a similar task using a web search engine." Following these instructions, participants completed the HIT in four steps: (1) search task description, (2) pre-task questionnaire, (3) main task, and (4) post-task questionnaire.

Pre-/Post-task Questionnaires: On both questionnaires, participants reported their level of agreement with statements using a 7-point scale (1-strongly disagree to 7-strongly agree). For measures of interest that involved multiple questions, we used Cronbach's α to measure internal consistency. The pre-task questionnaire included 9 items designed to measure participants' perceptions regarding: (1) the task being specific and narrowly focused (i.e., specificity) (5 items, $\alpha = .751$); (2) being confident about the information the SG trails might contain (3 items, $\alpha = .759$); and (3) the task being difficult (1 item). The post-task questionnaire included 15 items designed to measure participants' perceptions regarding: (1) gains obtained from the SG (4 items, $\alpha = .871$); (2) the SG having proper coverage of the task topic (3 items, $\alpha = .786$); (3) the SG having high-quality information (3 items, $\alpha = .848$); (4) the SG having information they expected to find (2 items, $\alpha = .818$); and (5) the task being difficult (3 items, $\alpha = .708$).

6 RESULTS

We present results in terms of our three research questions (RQ1-RQ3). We used one-way ANOVAs to analyze the effect of task version on each RQ1 measure, and two-way ANOVAs to analyze the effects of task and trail version on each RQ2 and RQ3 measure. We used Bonferroni correction in all post-hoc pairwise comparisons. In our RQ1-RQ2 results, slight variations in the F-statistic's degrees of freedom are due to missing questionnaire responses. For our pre-task (RQ1) and post-task (RQ2) measures, our MTurk participants used only a narrow portion of the 7-point scale in our questionnaires. We report on significant differences, but note that some of these are small. In Section 7, we discuss how our RQ1-RQ2 results (though small) are consistent with trends from prior studies.

Pre-task perceptions (RQ1): As shown in Figures 2a-2b, task version had a significant effect on: (1) specificity ($F(3, 1226) = 85.09$, $p < .001$; post-hoc: U < D < I < B) and (2) SG expectations ($F(3, 1226) = 2.87$, $p < .05$; post-hoc: D < I). Task version did not have a significant effect on pre-task difficulty.

Post-task perceptions (RQ2): As shown in Figures 3a-3c, task version had a significant main effect on participants' perceptions about: (1) gains obtained from the SG ($F(3, 1222) = 5.21$, $p < .005$; post-hoc: D < I,B), (2) the coverage of search trails in the SG ($F(3, 1213) = 4.12$, $p < .01$; post-hoc: D < I), and (3) the quality of information in the SG ($F(3, 1217) = 4.31$, $p < .005$; post-hoc: D < I). Additionally, trail version had a significant main effect on participants' perceptions of the quality of information in the SG (not shown in Figure 3; $F(1, 1217) = 5.52$, $p < .05$). SG information quality was significantly higher in the broad trail condition (6.00 ± 0.08) versus narrow trail condition (5.86 ± 0.08). Task and trail version did not have significant effects on post-task difficulty.

As shown in Figure 3d, task and trail version had a significant *interaction* effect on the extent to which the trails in the SG matched participants' expectations ($F(3, 1216) = 2.84$, $p < .05$). For task versions U and D, participants reported having their expectations met more in the broad versus narrow trail condition. However, participants reported no such difference for task versions I and B.

Search Behaviors (RQ3): To investigate RQ3, we used logged data to compute four measures related to the amount of SG exploration: (1) query clicks, (2) result clicks, and (3) total clicks, (4) number of tab clicks (to explore different trails).

As shown in Figures 3e-3g, task version had a main effect on: (1) number of query clicks ($F(3, 1226) = 6.88$, $p < .001$; post-hoc: U < D,B), (2) result clicks ($F(3, 1226) = 4.69$, $p < .005$; post-hoc: U,I < B), and (3) total clicks ($F(3, 1226) = 4.97$, $p < .005$; post-hoc: U,I < D).

As shown in Figure 3h, task and trail version had a significant *interaction* effect on the number of SG tab clicks ($F(3, 1226) = 6.27$, $p < .001$). For task versions U and D, participants made significantly more tab clicks with narrow vs. broad trails. Conversely, for task versions I and B, the differences in tab clicks were less pronounced.

7 DISCUSSION AND CONCLUSION

RQ1: Our RQ1 results found two important trends. First, task version had a significant effect on participants' pre-task perceptions of the task being specific and narrowly focused (i.e., specificity) (Figure 2a). These results are consistent with previous findings [4] and suggest that: (1) including either items or dimensions increased participants' perceptions of task specificity (U < D,I), (2) including items had a *stronger* effect than specifying dimensions (D < I), and (3) including items and dimensions had an additive effect (D,I < B).

Second, task version had a small, but significant effect on the extent to which participants expected to be familiar with the search trails in the SG (Figure 2b). Specifically, participants' expectations were higher when the task specified items than when the task specified a dimension (D<I). This result extends our prior work [4]. In Capra et al. [4], participants interacted with a search engine (not search trails). We found that specifying items in the task description made participants perceive the task as more determinable (i.e., less complex), and that specifying dimensions made the task *less* determinable (i.e., more complex). As previously noted, dimensions tend to be abstract concepts with varied vocabulary and are often subjective criteria, requiring assessing credibility and synthesizing opinions. Our results in the current study show that the effects of determinability *generalize* beyond interactive search, onto other forms of searching (i.e., using others' trails).

RQ2: Our RQ2 results found two main trends. First, task version had significant effects on participants' perceptions of the search trails found in the SG (Figures 3a-3c). Specifically, for tasks that included a dimension (and no items), participants reported lower gains from the SG (D < I,B), lower coverage of the task topic in the SG trails (D < I), and lower information quality in the SG trails (D < I). This result is largely consistent with our RQ1 results, and suggests that searching for dimensions using search trails is more difficult than searching for items.

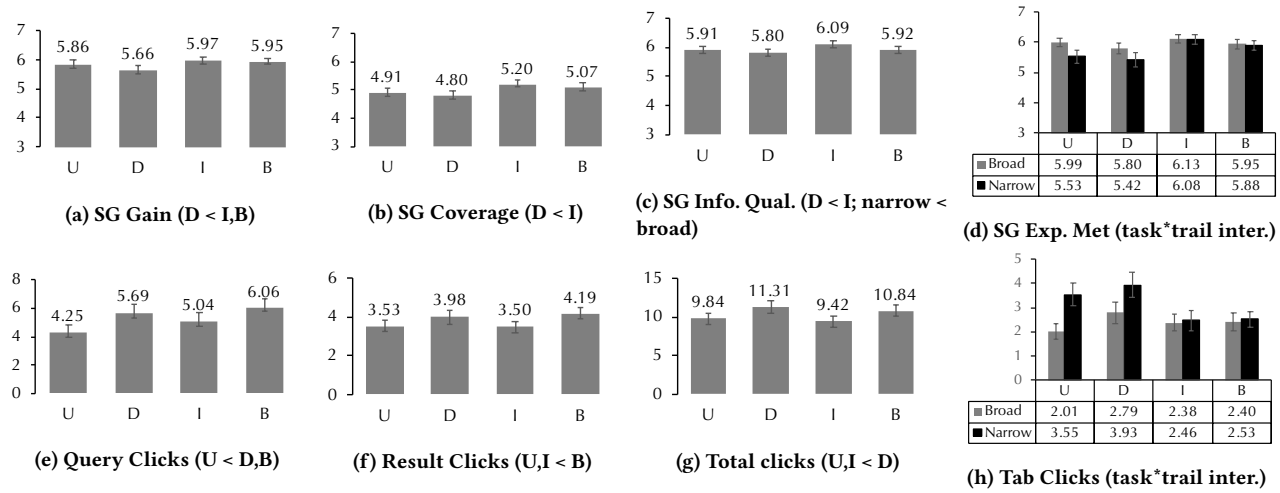


Figure 3: The effects of task and trail version on post-task factors (RQ2) and search behaviors (RQ3).

Secondly, based on participants' perceptions, broad trails outperformed narrow trails in two respects. Participants reported higher information quality from the SG trails in the broad versus narrow trail condition. Additionally, task and trail version had a significant interaction effect on the extent to which the SG trails matched participants' expectations (Figure 3d). Specifically, for tasks that were perceived to be narrowly focused (I, B based on RQ1), participants' expectations were similarly met with both trail versions. Conversely, for broadly perceived tasks (U, D based on RQ1), participants' expectations were better met by broad trails.

RQ3: Our RQ3 results reinforce some of the trends mentioned above. First, tasks that included a dimension required more interaction (i.e., more search effort) in terms of the number of SG query clicks (U < D,B), result clicks (U,I < B), and total clicks (U,I < D).

Secondly, task and trail version had a significant interaction effect on the number of tab clicks (i.e., a surrogate of number of search trails explored). For narrowly focused tasks (I,B), the number of tab clicks were similar. Conversely, for broadly focused tasks (U,D), participants had fewer tab clicks with broad versus narrow trails. Put simply, our RQ2 and R3 results suggest that participants were better able to address narrow tasks with broad trails than broad tasks with narrow trails (i.e., broad trails are better).

Implications: Using search trails to support users requires making two important decisions: (1) deciding *when* to display trails and (2) deciding *which* trails to display. To help address the first decision, one important question is: How do task characteristics impact users' expectations about the search trails provided and their ability to gain benefits? To help address the second decision, an important question is: Should the system favor narrowly or broadly focused trails on the topic of a user's current search session?

In our study, we manipulated the scope of comparative tasks by specifying two items (I), one dimension (D), or both (B). Consistent with prior work [4], our results suggest that searching for items is easier than for dimensions. Put differently, searching for items made tasks more determinable (less complex), and searching for dimensions made tasks less determinable (more complex). Our results suggest that task determinability is an important criterion for deciding whether to show trails. For more determinable tasks (e.g., I vs. D), our participants were more confident about knowing what the search trails might contain (RQ1) and reported better experiences interacting with the trails provided (RQ2). This result is somewhat paradoxical—during indeterminable (more complex) tasks, searchers may need greater support, but may be less able to gain benefits from search trails. Future work is needed to better understand this relationship. Perhaps search trails are most useful for tasks with a medium-level of determinability (not trivial, not overly complex). Alternatively, there may be trail characteristics or presentation strategies well-suited for highly indeterminable tasks.

Finally, our results have implications for deciding which trails to display. First, based on our results, the best alternative is to show trails with the same scope as the user's task—broad trails for broad tasks, and narrow trails for narrow tasks. Secondly, our results suggest that participants were better able to address narrow tasks with broad trails than broad tasks with narrow trails. As a design implication, if a system cannot infer the scope of a user's task, displaying broad trails may be the best choice.

Acknowledgements: This work was supported in part by NSF grant IIS-1718295. Study materials at: https://ils.unc.edu/searchtrails/sigir2019/

REFERENCES

[1] Mikhail Bilenko and Ryen White. 2008. Mining the Search Trails of Surfing Crowds. In *WWW*. ACM, 51–60.
[2] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information Processing & Management* 31, 2 (1995), 191–213.
[3] Robert Capra, Jaime Arguello, Anita Crescenzi, and Emily Vardell. 2015. Differences in the Use of Search Assistance for Tasks of Varying Complexity. In *SIGIR*. ACM, 23–32.
[4] Robert Capra, Jaime Arguello, Heather O'Brien, Yuan Li, and Bogeum Choi. 2018. The Effects of Manipulating Task Determinability on Search Behaviors and Outcomes. In *SIGIR*. ACM, 445–454.
[5] Chathra Hendahewa and Chirag Shah. 2017. Evaluating User Search Trails in Exploratory Search Tasks. *IP&M* 53, 4 (2017), 905–922.
[6] Yuelin Li and Nicholas J. Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *IP&M* 44, 6 (2008), 1822 – 1837.
[7] Adish Singla, Ryen White, and Jeff Huang. 2010. Studying Trailfinding Algorithms for Enhanced Web Search. In *SIGIR*. ACM, 443–450.
[8] Ryen White, Mikhail Bilenko, and Silviu Cucerzan. 2007. Studying the Use of Popular Destinations to Enhance Web Search Interaction. In *SIGIR*. ACM, 159–166.
[9] Ryen White and Jeff Huang. 2010. Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs. In *SIGIR*. ACM, 587–594.
[10] Barbara M. Wildemuth, Luanne Freund, and Eliane G. Toms. 2014. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *Journal of Documentation* 70, 6 (2014), 1118–1140.
[11] Xiaojun Yuan and Ryen White. 2012. Building the Trail Best Traveled: Effects of Domain Knowledge on Web Search Trailblazing. In *SIGIR*. ACM, 1795–1804.