The Search Dashboard: How Reflection and Comparison Impact Search Behavior

Scott Bateman
University of Saskatchewan
Saskatoon, SK, Canada
scott.bateman@usask.ca

Jaime Teevan, Ryen W. White
Microsoft Research
Redmond, WA, USA
{teevan, ryenw}@microsoft.com

ABSTRACT

Most searchers do not know how to use Web search engines as effectively as possible. This is due, in part, to search engines not providing feedback about how search behavior can be improved. Because feedback is an essential part of learning, we created the *Search Dashboard,* which provides an interface for reflection on personal search behavior. The Dashboard aggregates and presents an individual's search history and provides comparisons with that of archetypal expert profiles. Via a five-week study of 90 Search Dashboard users, we find that users are able to change aspects of their behavior to be more in line with that of the presented expert searchers. We also find that reflection can be beneficial, even without comparison, by changing participants' views about their own search skills, what is possible with search, and what aspects of their behavior may influence search success. Our findings demonstrate a new way for search engines to help users modify their search behavior for positive outcomes.

Author Keywords

Search expertise, Web search, reflection, social learning.

ACM Classification Keywords

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

General Terms

Experimentation, Human Factors.

INTRODUCTION

Although Web search engines work very well most of the time, many people still experience problems finding what they are looking for. In a recent survey, nearly five percent of users reported completely failing at their most recent attempt to search for something on the Web [11]. When search failures occur, they cost people a lot of time: searchers spend over ten minutes when they fail before giving up, as compared to needing less than five minutes when they are successful [11]. Part of the problem is that when search becomes difficult, many people are unsure of how to change strategies or how to make use of the advanced search engine functionality that could help them [1,22]. People largely use the same search behavior regardless of the situation and how successful they are [22]. Further, even when people are successful, they still may not have been as efficient as they potentially could have. This suggests that searchers have room for improvement.

The main approach for improving people's search success has been to improve search engines to deliver the best results for a given query. However, another approach is by educating users to be better search engine users [20]. Previous work has shown that the knowledge and use of particular search engine functionalities, and other characteristics of search engine use (e.g., the average number of terms per query), can predict *search performance* [2,27] – the degree to which people are successful and efficient in their search tasks. This suggests that if users knew more about how to use search engines they would be able to improve their own search performance.

The best way to help users improve their search behavior is unclear. But work in education and in persuading users to adopt new behaviors have both highlighted the importance of feedback on personal behavior for reflection and learning [7,18]. Additionally, observing other skilled practitioners can improve learning [4,19], and knowing what others do can lead to positive choices [25]. However, search engines lack feedback. Users are not able to get an idea of their own search behavior and are not able to find out what behavior works for other people. The lack of facilities for reflection on personal behavior means it can be difficult for users to learn how they can get better, and what search strategies can lead to improved performance. Further, even if people could get an accurate idea of their search behavior and how it might be changed, it is not clear whether they would be able to adjust their search behavior in any meaningful way.

In order to find out if reflecting on search feedback can have an influence on the attitudes and behaviors of searchers, we created and studied the *Search Dashboard*. Our system is the first that aims to positively influence user search behavior through reflection on personal history and comparison with the behavior of archetypal user profiles (such as search or topic experts). We performed a five-week study of 90 users and examined how reflection affects user attitudes and understanding of how search engines can be used, and can lead to observable behavior change.

Our study provides two main results. First, feedback and reflection on past search behavior *can* lead to changes in

CHI'12, May 5–10, 2012, Austin, Texas, USA.
Copyright 2012 ACM 978-1-4503-1015-4/12/05...$10.00.

attitudes and behavior in actual search engine use. Second, the presentation of archetypal profiles, which aggregate the behavior of many people, can lead to increased interest and engagement, and were critical in creating the observed behavior change.

RELATED WORK

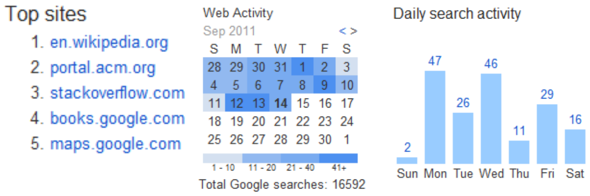
The work presented here builds on research in: (i) feedback, reflection, and persuasion, where information on people's activities are shown to them, to provide insights, change attitudes, and to promote learning and positive behavior change; and (ii) search performance, where factors that influence and predict successful searching are analyzed.

Feedback, Reflection, and Persuasion

The area of personal informatics aims to design systems that help people learn about and understand their own behavior, with the goal of providing new insights, increasing self-control, and promoting the acquisition and maintenance of desirable behavior [18]. Similarly, persuasive technologies have been described as systems that seek to change behavior or attitudes, without the use of coercion [12]. Personal informatics and persuasive systems have been created for reflecting on past behavior and promoting behavior change in several domains including physical activity [9], environmental impact [13], and webpage visitation [26].

Theories of learning have promoted reflection as an essential part of learning [7,8]. Reflection is the process "… in which people recapture their experience, think about it mull it over and evaluate it … [it is] this working with experience that is important in learning [7, p.19]." Other theories also highlight the social nature of learning, where learning occurs through observing and imitating skilled practitioners [4,19]; and in so doing, learners can derive the thought process of others [8]. Work in social psychology has also shown that providing descriptions of what other people do can "nudge" people towards particular decisions [25].

In the domain of Web search there has been little research into what, and how, data should be displayed for feedback. In one study of personal and shared web activity, participants found views containing the data of others to be more useful [26]. The Google and Bing search engines both provide "Search History" functionality that allows users to review past queries (google.com/history; bing.com/profile /history). Such systems have been shown to improve performance in re-finding and resuming search tasks [21], but it is unclear whether they would be effective tools for reflection. Google's search history includes "Trends", which displays a user's most-frequent queries, most-visited sites, most-often clicked search results, and the total number of searches executed over different time frames (see Figure 1). While search engines have recognized some value in presenting summaries of behavior, there is little information about what data should be presented, and what effect it might have on user attitudes and behavior.



Top sites

1. en.wikipedia.org
2. portal.acm.org
3. stackoverflow.com
4. books.google.com
5. maps.google.com

Web Activity

Sep 2011

S	M	T	W	T	F	S
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	1

Total Google searches: 16592

Figure 1. A user's search activity displayed in Google Trends.

Search Performance

Moraveji et al. [19] identified several factors that affect an individual's search performance, including knowledge of search engine features [14, 20] and the resources being sought [28], topical expertise [29], level of general literacy [16], and differences between task types [1].

There has been much work in characterizing the differences between search experts and novices, and in identifying characteristics that predict search performance. Several studies have used different user characteristics to determine who is an expert, including: having more than 50 hours experience on the Web [17], browsing the Web more than 5 hours/week [15], and performing searches as part of a job for at least 3 years [14]. Behavioral differences have also been noted between novices and experts, where experts tend to take more [6] or less time [23] to complete tasks, use more query terms [14,27], use advanced operators [3], and have higher [15,17] or comparable performance [6].

White and Morris [27] used a simple approach of identifying advanced search engine users by their use of four search operators (any of quotation marks, '+', '–', or 'site:'). By having external judges rate the relevance of Web pages users visited for a given query, they showed that, on average, the identified advanced users visited pages judged to be more relevant than non-advanced users (i.e., advanced users were more successful).

Recent work has also examined how search performance can be improved by increasing search engine knowledge. A controlled study showed that users who were taught advanced search engine functionality that would greatly improve their performance on a search task, were able to successfully apply the knowledge, both at the time of learning about the functionality and a week later [20]. Bing recently added a rewards program (bing.com/rewards) that provides periodic tasks, each of which introduces new search engine functionality. When tasks are completed, users earn points redeemable for merchandise. A Google A Day (agoogleaday.com) allows users to practice their search skills, by providing daily questions. Users race against a timer to search for an answer, and if they become stuck can receive a hint query that will lead to the correct answer.

Previous research has taken a number of different approaches to identify, characterize, predict and improve search expertise, but there has been no work on what effect feedback and reflection on personal search activities might have on user attitudes and behavior. We address this gap by creating a prototype system to explore the space of possible feedback data, how viewing the data of others might improve reflection on personal data, and to study if and how people make use of this new type of information.

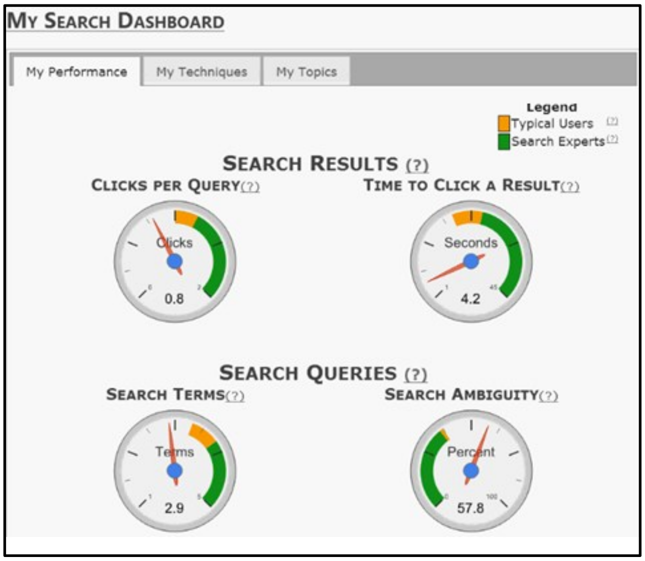


Figure 2. The Search Dashboard, displaying the Tendencies section (referred to as Performance in the study).

THE SEARCH DASHBOARD

We created the Search Dashboard system (Figure 2) to display personal search history data, with the goal of understanding what types of information are most useful and influential to searchers. The design was influenced by a 53-question survey of 75 employees at the Redmond headquarters of Microsoft. Participants were solicited by email and asked to rate their interest in being able to see different aspects of their search history (e.g., the days of the week they search on, or their average query length), and whether they might be interested in comparing that information with other individuals or groups (such as friends, colleagues, or Web search experts). Although our main research questions relate to behavior change, the initial survey focused on interest because prior research on social learning suggests people must have an interest and perceive value in an observed behavior in order to have motivation to change it [4]. Survey results are highlighted where appropriate below.

We begin by describing the types of personal search history reflected in the Search Dashboard. Because the ability to compare one's behavior to others has been shown to be a valuable learning tool, we also developed a variant of the system that supports comparison, and we next describe the comparison data we collected. Finally, we discuss how all of this information was gathered and presented to the user.

Data Types (Techniques, Tendencies, and Topics)

Based on our organization of related work, we identified three main types (called *Data Types*) of history data that may be valuable for reflection on personal search behavior, since they have been shown to affect search performance:

- *Techniques*: The use of advanced query operators and special search engine features (e.g., [3,20,27]).
- *Tendencies*: Summative actions describing the overall characteristics of search engine use, such as the number of terms used per query (e.g., [14,27,28]).
- *Topics*: The content and subject of an individual's search engine use (e.g., [1,6,14,17,23]).

Using the formative survey and previous work, we selected 12 specific *elements* of people's search history to display in the Dashboard. These elements represented different points in the space of Data Types (i.e., Techniques, Tendencies, and Topics) and were those most often selected as being of interest by survey respondents (mean selection rate = 47%, as compared to an average of 34% for other aspects). These elements, and how they are calculated, are now described.

Techniques Data

Operator use: There are a number of advanced search operators understood by Web search engines, including quotation marks (used to group query terms together), '+' (used to mark a term to be unaltered by the search engine), '-' (exclude a term from a query), and 'site:' (used to restrict results to a particular domain). We represent a person's search operator use by counting the number queries issued by that person that contain a particular search operator.

Vertical use: In addition to general Web search, there are a number of verticals that search engines offer to enable users to search particular types of data (e.g., images, maps) or perform particular types of tasks (e.g., shopping, travel). Awareness of the different verticals available is important, so we count the number of times each vertical is used.

Direct Answer use: Direct answers are information tailored to specific queries, e.g. for weather forecasts. They are shown inline on search engine result pages to address searchers' needs quickly. We count the number of each type of direct answer a person encountered in search results.

Tendencies Data

Query length: An individual may issue long or short queries. We calculated the average number of search terms an individual uses in their queries.

Query ambiguity: We measure query ambiguity by examining the variation in the search results different people click following the same query. We use historic search log data from many users over one month before the beginning of the study, and calculate the average click entropy, as defined in [10], for each user across all queries they issue.

Clicks per query: People vary in the number of search results they select after querying. We calculate the average number of results an individual clicks following a query.

Time to click: After a person issues a query, they may click a result right away, or pause to consider the retrieved results first. To capture this, we measure, in seconds, the average time between issuing a query and clicking on a result.

Session time: Queries often appear as part of a search session, rather than in isolation. We identify sessions by looking for search activity with less than 10 minutes between each action, and compute the average duration, in minutes.

Session queries: As a general measure of activity level, we also compute the average number of queries that each user issues per session.

Topics Data

Categories: We aggregate the most popular topical categories of the search results clicked by an individual. The category of a URL is determined by selecting the highest probability category returned by a content-based Web page classifier (described in [5]), which assigns URLs to Open Directory Project (ODP, dmoz.org) categories. We used 13 top-level ODP categories for this labeling: Arts, Business, Computers, Games, Health, Home, Kids and Teens, News, Recreation, Science, Shopping, Society, and Sports.

Domains: We extract the most common domains from search results clicked by an individual. For example, if a person clicks on many Wikipedia search results, wikipedia.org is a distinctive domain for that individual.

Query terms: To show the most salient search words, we parse the individual terms from each of a person's queries, remove stop words, and count the frequency of each term.

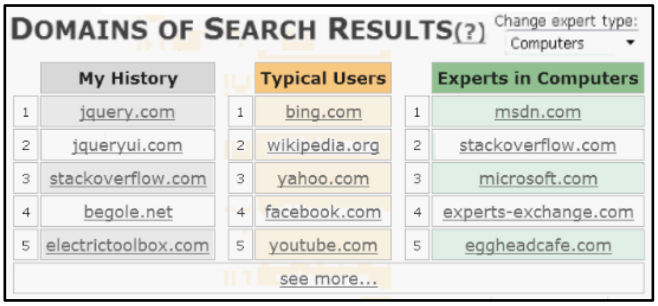
Comparison Data (User Archetypes)

Because the ability to compare one's behavior to others has been shown to be a valuable tool for learning and decision making, we gathered comparison data for several different user archetypes. While some previous work has investigated the display of the history of other individual users for comparison with personal Web browsing history [26], we decided to use the data of groups for two main reasons. First, sharing individual data raises privacy concerns. Second, about two thirds of our initial survey respondents reported wanting to able to compare their personal search information to others (65.3% agreed). They were most interested in comparing their data with "people who are experts in topics I am interested in" (68% agreed) and "people who are experts at searching the Web," (51% agreed) as compared to people they know, such as colleagues (41% agreed), or family and friends (25% agreed).

The three user archetypes we created represented *Typical Users*, *Search Experts*, and *Topic Experts*. To generate the representative values of the 12 data elements for each archetype, we algorithmically selected sets of representative users from the opt-in logs from a browser plugin widely deployed by Bing, using data from one month starting April 15, 2011. To remove variability caused by geographic and linguistic variation, we only include entries generated in the English speaking United States. Individual user history was then aggregated over all users to generate a single archetype profile. We now describe how the users were identified.

Typical Users: This archetype allows Dashboard users to compare their behavior with typical user behavior. To generate the archetype, we randomly selected 1000 users from all non-search experts during the sampling period.

Search Experts: To create this archetype, we used the approach of White et al. [27] who showed that web searchers who tend to be more successful could be identified by the use of search operators. We selected 1000 users who most frequently used search operators during the sampling period. Search experts were used for comparison with Techniques and Tendencies data. We explore potential issues with this approach in the Important Considerations section.



DOMAINS OF SEARCH RESULTS(?)

Change expert type: Computers ▾

	My History		Typical Users		Experts in Computers
1	jquery.com	1	bing.com	1	msdn.com
2	jqueryui.com	2	wikipedia.org	2	stackoverflow.com
3	stackoverflow.com	3	yahoo.com	3	microsoft.com
4	begole.net	4	facebook.com	4	experts-exchange.com
5	electrictoolbox.com	5	youtube.com	5	eggheadcafe.com

see more...

Figure 3. The Domains data in the Topics section.

Topic Experts: To create this archetype we selected topically relevant subsets of users from the 2000 users selected as typical or expert. A user was considered topically relevant if they had visited at least 10 search results in the category in a one-week period. Our approach created 13 topical expert profiles, one associated with each topic. The 13 generated topical experts were presented with Topics data only.

Displaying Data and Interacting

Information about an individual's personal search history and comparison information for the three different Data Types are displayed in the Search Dashboard system. The system is accessed using an intranet URL from a standard Web browser. When users visit the website, their personal search history is automatically loaded. Each Data Type (Tendencies, Techniques, and Topics) is displayed separately using a tabbed interface. To understand the value of comparison data, we created two versions of the Dashboard (called *Variants*): (i) the *no comparison* Dashboard showed only the user's personal data; (ii) the *comparison* Dashboard added the data of our archetypes for comparison.

Displaying Textual Data

Two types of data are displayed: textual and numeric. Textual data can be found on the Techniques and Topics sections, and are presented as lists in tables (see Figure 3). The top five values for a textual attribute are displayed, and the full list (up to 100 values) can be viewed by clicking the "see more…" link below the tables.

When comparison information is included with textual data, two additional columns are added to the table to present archetypal user data for the typical and expert user. The expert used on the Techniques tab and Tendencies tab is a search expert, and on the Topics tab is a topic expert. Although there are 13 topical expert archetypes, only one was shown at a time. A different topic expert can be selected via a drop down list positioned by each of the topic expert data tables. The topic expert that is most similar to the topic that the user is associated with is selected by default.

Because some textual data, like domains and query terms, occur more often than others, and thus are more likely to be used without an existing preference, we identified an individual's most distinctive data by normalizing the count of the textual item by the count for Typical Users. This is

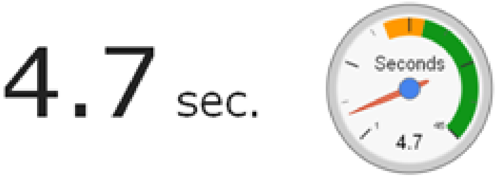


Figure 4. Alternative presentations of numeric data: without comparison data (left), and with comparison data (right).

analogous to the TF.IDF method from information retrieval [24], and resulted in data being ordered by the most distinctive items, rather than those that occurred most frequently. This approach is used for user, and expert archetype, data.

Displaying Numeric Data

Numeric data, which includes all of the Tendencies data, is displayed in two ways (both are shown in Figure 4). The first way, when no comparison data is used, only the number and unit of the measurement is displayed in large font. When comparison data is available, numeric data is shown using a gauge chart, which allows a number of data values to be concisely displayed. On each gauge the position of the needle and the number displayed indicates the user's value for the particular measure. Two regions are also displayed, an amber region representing the range of typical users (defined as the values between Typical User's value and the Search Expert's value), and a green region representing the expert user range (defined as the value from the Search Expert's value to the extent of the gauge). We opted for gauge charts after piloting alternatives (e.g., bullet graphs); they were familiar and required the least explanation.

Exploring Data and Getting More Information

The Search Dashboard also provides facilities for users to obtain explanations and more details about any of the data displayed. Each data label in the Dashboard provides a brief description of the data element and possible interpretations of values, via a tooltip. These labels are intended to be descriptive rather than instructive. For consistency, the same descriptions are used regardless of whether comparison data is shown or not. For example, the tooltip describing "clicks per query" reads: "The number of results you typically click on following a query. Some people are very selective with their search results, leading to very few clicks per search; some people average less than 1 click per result. Other people are happy to explore many search results."

Table text is also hyperlinked, and clicking performs an exemplary action, such as launching a search engine query illustrating the functionality. Hovering over text data also provides more details. For the values in the Domains tables (in Figure 3), hovering on a domain name calls a popup with the list of queries that led to that domain being visited. Clicking any query in the popup issues the query to a search engine. Figure 5 shows the popup for direct answers.

STUDY

We conducted a five-week study with 90 participants to understand how people make use of different types of information in their search history summary. Our study design aimed to address four main questions:

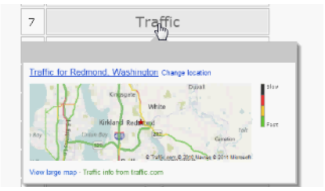

7
Traffic
Traffic for Redmond, Washington Change location
View large map · Traffic info from traffic.com

Figure 5. An example of the Traffic direct answer. When a user hovers over an answer type an example is shown.

- Do participants perceive information about their personal search history as valuable?
- How are the different types of personal search history data perceived and used?
- Does reflecting on personal search behavior lead people to change their attitudes about search engines and adopt new search behavior after seeing the Dashboard?
- Does the ability to compare one's personal search history with others provide benefit?

We chose to focus our research on changes in behaviors and attitudes, rather than on search success because success is less accurate to measure, particularly within the context of real day-to-day workplace searches. This is an important first-step, as the ability to effectively impact search behavior can be valuable regardless of search outcome.

Data Collection

To address our four main questions, we collected three forms of data: (i) We logged all participant web browsing activity from which we were able to extract queries issued to all Web search engines and what, if any, advanced search engine functionality was being used. (ii) We collected survey data during each of the three parts of the study (described below). Surveys contained a number of Likert-scale and free-text questions. (iii) Finally, we also collected log data about how participants interacted with the Dashboard.

Conditions

Our study involved a mixed 3 × 2 design (*Data Type* × *Dashboard Variant*). Because our first research questions related to how people would value and perceive different aspects of their search history, we grouped our three Data Types (Techniques, Tendencies and Topics) into three tabs (separate sections) on the Dashboard. For the study, the three Data Types were presented in random order, and participants worked with only one Data Type at time, all participants saw all three Data Types (i.e., within subject).

We hypothesized that having the data of other users available for comparison would increase how valuable participants found reflection. However, current search history tools provide personal data only. So, we were interested in identifying the differences between personal data and having personal data augmented with the data of archetypal users for comparison. For this reason, we developed the two Dashboard Variants (comparison and no comparison) that were studied between-subjects – participants were randomly assigned to use only one of the Dashboard Variants.

Participants

Ninety people participated. All were employees of Microsoft at the company's Redmond, WA headquarters. Most (73, 81%) were male, which is consistent with the company's demographics. All reported searching the Web at least daily, with over half (50, 56%) reporting that they searched more than 10 times a day. Participants were randomly selected from the company directory, and recruited via email. In exchange for participation in all five weeks of the study, they were entered into a sweepstakes for one of four prizes (one $300 gift card, three $100 gift cards).

Procedure

The study was in three parts over five weeks: (i) initial registration, (ii) introduction to the Search Dashboard after three weeks, and (iii) a final exit survey after five weeks.

Part I: Registration

At the study outset, participants were asked to enroll using their primary work computer, by completing an initial survey, and configuring a piece of software to log all of their search engine activity for the study period. The registration survey asked basic demographic information, as well as perceptions and attitudes towards Web search.

Part II: Introduction to Search Dashboard

Three weeks (on average) after participants enrolled in the study, they were sent an email requesting that they view their personal Search Dashboard. The actual time between enrolling and completing the study ranged from two to four weeks depending on participants availability.

The search engine activity that was logged during the time between Part I and II of the study was used to populate the personal data in each individual's Search Dashboard.

Participants were randomly assigned to one of the two between-subject Dashboard Variant conditions (either compare or no compare). When participants visited the Dashboard for the first time they were initially presented with a walkthrough that explained their task, guided them through the features of the search dashboard and described how their data would be presented. They were then asked to view each of the three tabs, one at time, for as long as they like. Recall that each tab represents one of the three Data Types (Techniques, Tendencies, and Topics). To ensure that participants could not simply click through the study without viewing their data, the system enforced a one minute delay on each tab before the participant could proceed. After viewing each tab, a survey was presented that asked about the Data Type they just viewed. After all three surveys were completed, a final survey was presented that asked participants about their overall experience with the Dashboard and asked them again about their attitudes and perceptions of Web search. After the final survey participants were told they could visit the Dashboard at any time.

Part III: Exit Survey

Finally, after roughly another two weeks, participants were contacted again. The time between Part II and III ranged from 6 days to 19 days (average 12 days). Participants were asked to visit the other Dashboard Variant – i.e., participants who initially saw the comparison Variant were asked to view the no comparison, and vice versa. Participants were also asked to complete a final survey that was focused on the differences between the two Dashboard Variants.

Data Analysis

Surveys contained 5-point Likert scales (1=strongly disagree, 3=neutral, 5=strongly agree) or 7-point self-rating of skill (1=min, 7=max). Analysis of within-subject factors for survey questions used Friedman's ANOVA for related samples; post-hoc pairwise comparisons used Wilcoxon Signed Ranks Tests for two related samples. For the analysis of questions for between-subject conditions, we used the Mann Whitney U test. In addition to analyzing the quantitative data, we used survey comments to help explain results found in statistical analysis. Our post-hoc tests of survey data and analysis of search log data used many dependent variables, so using Bonferroni corrections to control the experiment-wise error rate, we set α to 0.05 divided by the number of dependent variables.

Results

We now present our results, organized by main findings. We show that participants found their interactions with the Search Dashboard (and in particular with their Techniques and Tendencies) to be valuable, and that it impacted their self-assessment of their search skills. The use of comparison data, especially search expert data, led to increased engagement and insights, as well as observed behavioral changes for Tendencies and Techniques.

The Search Dashboard is Engaging

We assessed participants' level of engagement, through the Dashboard interaction logs. Although we did not provide specific time guidelines for using the Dashboard, analysis of the Dashboard log data reveals that the participants spent almost half of an hour (28.65 min.) exploring their data. During this time, participants looked at 39.96 tooltips on average. They were also likely to return to the Dashboard on their own after their initial study visit, with 72.53% doing so at least once and the logs recording an average of 2.28 visits per participant.

Techniques & Tendencies More Insightful Than Topics

To understand how insightful the Data Types were we elicited ratings on "I learned something new", "I was surprised by some aspect of my history", and "Based on what I saw, I will change how I search". All three aspects showed significant differences between Data Types. Statistics are presented in Figure 6, and are discussed below.

Overall, users did not feel the Topics provided much insight in terms of the questions posed, agreeing below neutral that Topics would lead to change in how they search. A few participants really enjoyed the Topics data. One participant reported, "I love the area of my topics and how it separates them out into categories ... That type of information I find very interesting." However, more participants stated that

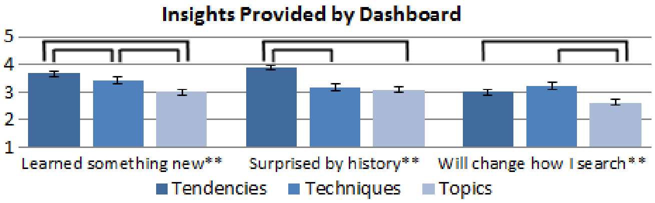


Figure 6. Mean ratings and ± SEM of insights provided by the Dashboard for each of the three Data Types. All questions were sig. diff. (*p*<.001). Sig. diff. pairs are indicated by lines (*p*<.05*)*.

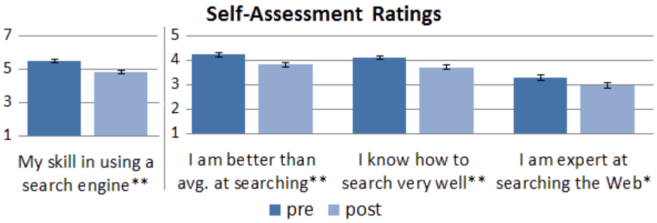


Figure 7. Mean and ± SEM for self-assessment ratings before (pre) and after (post) using the Search Dashboard. Sig. diff. were found for all ratings (*=*p*<.01, **=*p*<.001).

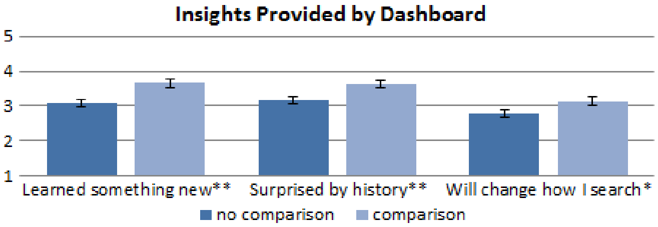


Figure 8. Mean and ± SEM for ratings of insights between Dashboard variants. Sig. diff. were found for all aspects (*=*p*<.05, **=*p*<.001).

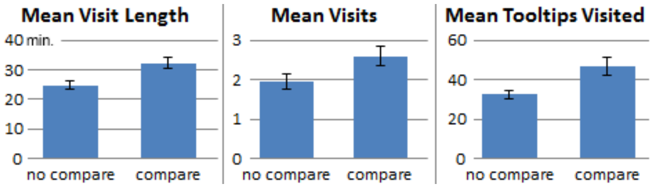


Figure 9. Mean and ± SEM usage statistics for both Dashboard variants. For all aspects, usage was significantly higher in the compare condition than the no compare condition (*p*<.01).

they were not certain how to make use of it: “Interesting information. But, not necessarily actionable.”

The Tendencies and Techniques data were viewed more positively than Topics. While insights were rated higher for both Data Types, participants were particularly likely to be surprised or learn something new when viewing data related to their Tendencies. Again, many users expressed not knowing how to take action on the data with users commenting: “probably the [Tendencies] tab [was the most interesting] although nothing was really ‘actionable’”, and “pretty cool, although not sure how I would modify my behavior based on that”. Despite Topics data receiving low ratings, some participants reported that they found the most value in seeing the data of Topic Experts (such as being able to see the domains that computer experts most often visit). For example, one participant said, “There is good information about what I would use to augment my searches in the future. It's helpful to find… sources of information aside from what I've been searching for in the past. It's like having a friend tell you ‘have you checked out this site?’”

Overall, the comments revealed a strong preference for information that participants could see an immediate application for. In these terms, they felt Techniques data would most likely change how they searched. While this was likely due to most people learning something new, being reminded was also valuable: “There are features … that I’ve forgotten about that this has been a great reminder for.”

Self-Assessment of Search Skills Changed

To see how the Dashboard influenced views about search skills, participants were asked to rate their search skills along a number of dimensions when they first agreed to participate in the study (*pre*), and then again several weeks later after viewing the Dashboard (*post).* As shown in Figure 7, through the use of the Dashboard, participants came to believe they were less skilled search engine users. This suggests that they learned they have room for improving their search behavior, whereas initially they saw less room.

Comparison Leads to Increased Engagement and Insights

To see if comparison data led to increased engagement and insights we looked at the ratings with Dashboard variant as a between-subjects factor (see Figure 8). In all cases participants rated the comparison dashboard higher than the non-comparison version. In particular, participants in the comparison condition were more likely to report that the Search Dashboard would change how they search, and as we will see in the subsequent section, this proves to be true. Participants’ comments also revealed enthusiasm for the comparison data (e.g., “Interesting. Fun to compare.”).

We hypothesized that participants would explore and spend more time with the comparison Dashboard. A series of one-tail *t*-tests to compare each of the three system-usage variables system usage revealed significant differences (with α=.0167). Users in the comparison condition spent more time visiting the Dashboard during the study, read tooltips more often, and paid more visits to the Dashboard after their initial visit (see Figure 9).

Behavior Change Observed for Tendencies & Techniques

To assess whether or not participants’ experiences with the Search Dashboard led them to change aspects of their behavior, we compared search behavior *pre* using the Dashboard to behavior *post*. For this analysis we looked at the differences between participants in the Dashboard Variant conditions for each of the data elements presented in the Dashboard. We also created a control group from the browser logs to provide a baseline for each of our metrics and to assess whether there were external factors (such as a new search engine feature) that could have caused any observable behavior change during the study period.

To build the control group, for each variable we sampled 1000 users separately using the log data described earlier, for a time period coinciding with the study. Separate samples for each variable were used to best resemble the mean and variance of that variable across all of our participants' pre-dashboard behavior. We did this because we believed that a simple random sample of users would not yield a set of users who were sufficiently representative of our participants' behaviors and search expertise across all dimensions. Only users who performed the actions of a measure were used in analysis (e.g., only users who clicked on a result were used to compute time to click).

The means and standard deviations for all measures across all groups and time periods are presented in Table 1. We used a mixed-design ANOVAs with group (no comparison, comparison, and control – independent measures), and time (pre dashboard and post dashboard – repeated measures), as the factors. The analysis revealed no significant difference between the experimental groups on any of the metrics for the *pre* dashboard period. There were significant differences between groups in some of the metrics *post* using the Dashboard (all $F(2{,}560\text{-}1100) \geq 6.24$, $p \leq .002$), where *560-1100* denotes the minimum and maximum degrees of freedom in the error term, based on the total number of users, across all groups, who performed the actions for each dependent variable and the number in the control group.

Participants in the comparison group changed Tendencies – taking longer to click on search results and issuing longer search queries – and Techniques – using operators, answers, and verticals more frequently – as compared to the other groups. There was a significant effect of time on behavior within the comparison group for the same Tendencies and Techniques (all $F(1{,}561\text{-}1101) \geq 10.94$, $p \leq .001$), and significant interactions for the same variables between group and time for the comparison group (all $F(2{,}558\text{-}1098) \geq 7.01$, $p \leq .001$). The findings suggest that participants in the comparison group significantly changed aspects of their behavior after seeing their Dashboard.

In looking for behavior change metrics for Topics data, we decided to assess whether the Dashboard led participants to explore more domains than they would typically consider, and whether they would consider exploring the topical information they were exposed to by using more unique queries. Neither of these measures showed a significant change in behavior after users viewed the Dashboard.

DISCUSSION

Our study of 90 peoples' experiences with the Search Dashboard provides the following five main results:

- Reflecting with the Search Dashboard changed peoples' attitudes about search engines and their own search skills.
- Reflection led participants to adjust their search behavior, but only when comparison data was available.
- Participants changed their behavior for 5 of the 12 data elements presented.

	measure	time	control	no compare	compare
Techniques	**Operator use**	pre	3.5% (1.1)	3.7% (1.2)	3.2% (0.9)
		post	3.4% (0.8)	3.3% (1.4)	***7.3% (1.0)***
	Vertical use	pre	2.98 (0.3)	2.97 (0.3)	3.01 (0.3)
		post	2.97 (0.3)	3.12 (0.4)	***3.74 (0.4)***
	Answer use	pre	31.2% (10.0)	30.1% (8.0)	29.9% (8.0)
		post	30.9% (8.0)	31.1% (7.0)	***33.1% (9.0)***
Tendencies	**Query length**	pre	2.91 (0.6)	2.91 (0.4)	2.93 (0.4)
		post	2.95 (0.50)	2.94 (0.4)	***3.07 (0.3)***
	Click entropy	pre	1.51 (0.7)	1.48 (0.7)	1.50 (0.7)
		post	1.50 (0.8)	1.51 (0.7)	1.49 (0.7)
	Clicks per query	pre	0.51 (0.2)	0.50 (0.2)	0.51 (0.2)
		post	0.52 (0.2)	0.51 (0.2)	0.51 (0.2)
	Time to click	pre	16.39 (2.4)	16.52 (3.3)	16.31 (3.11)
		post	16.41 (3.0)	16.43 (3.9)	***17.88 (3.1)***
	Session time	pre	13.6 (8.4)	13.4 (8.5)	13.8 (8.3)
		post	13.9 (8.4)	13.3 (8.3)	13.4 (8.3)
	Session queries	pre	1.42 (1.1)	1.39 (1.0)	1.43 (1.0)
		post	1.48 (1.1)	1.42 (1.0)	1.45 (1.1)
Topics	Domains	pre	85.40 (50.3)	85.29 (58.8)	86.44 (57.4)
		post	86.11 (49.5)	86.12 (59.2)	87.22 (59.2)
	Query terms	pre	77.67 (50.2)	77.93 (50.3)	77.11 (50.2)
		post	80.13 (51.4)	78.04 (49.6)	78.26 (50.3)

Table 1. The 11 metrics for the 12 data elements presented pre and post visiting the Dashboard. Sig. differences were found *within subject* and *between groups* for the same variables (bolded). For these variables comparison led to sig. higher usage.

- Participants preferred, and changed behavior for, data on Techniques and Tendencies but not Topics; likely due to a strong preference for data the can be easily applied.
- Comparison data also increased user engagement, changes in attitudes, and insight drawn from reflection.

Explanation for Results

Presenting personal data and descriptions of the different data elements provided participants with aspects of their behavior, which they may not have previously considered as affecting their search performance. This likely caught their interest, and led them to explore and consider the different data elements carefully. In this way, personal data alone was sufficient to allow participants to have valuable reflection on their behavior, by providing new insights and interaction possibilities, and also by showing them that there may be more to search engine use than they had previously thought.

The ease with which the data presented can be acted upon seems to be important to participant views of how valuable particular data is, and how likely they were to adjust that aspect of their behavior. Comparision data from user archetypes improved participants' perceptions of the value of the data and increased the degree of insight they were able to draw. Behavior change only occurred when comparison data was available. In other words, without the clear targets that were provided by seeing how other users behave, participants found the data less useful because the goal was not as obvious. Further, participants may not have known that there was room for change, but seeing that experts or even typical users were different from them, clarified the possibility of making positive changes.

There was likely a similar effect for the particular data elements themselves, where 5 of the 12 elements led to participants changing that aspect of their searching. The Topics data did not provide a clear use case to all participants, whereas data on Techniques and Tendencies were more straightforward. For example, a participant seeing that the weather direct answer is commonly used by experts, would have easily been able to infer that experts use this technique to save time (because the forecast is shown with the search results). For Tendencies data, comparison provided clear targets for how behavior could be changed, which would not have been available in the non-compare condition. For example, showing a user that they take eight seconds on average to select a result, does not necessarily suggest they should slow down. However, when participants saw that Search Experts take longer than Typical Users to select a result, they were able to infer that experts must study and select results more carefully than typical users (as suggested in survey responses).

Even with comparison data, some elements do not provide a clear path towards behavior change. For example, query ambiguity is not observable in search results or while formulating a query, so exactly how to influence one's own query ambiguity rating is not apparent. Also, data relating to Topics were likely difficult to apply, since these data largely relates to the queries people use. We believe that Topics data can provide useful information about topically relevant search terms, topic categories, and websites. However, because the range of potential information needs is so enormous, the data presented in the Topics section of the Dashboard may have be unrelated to a person's current need. For example, knowing that computer experts make frequent use of the stackoverflow.com domain might be interesting, but would be of little help in buying a laptop.

Finally, data that does not lead to behavior change can still have value. We believe that the Search Dashboard helped to change users' views of Web search overall because a complete picture of different search aspects was given. Further, because different people valued the data elements differently, a fuller data set could engage a wider audience.

Important Considerations

While our results clearly show that reflecting on personal search behavior *can* lead to behavior changes, we must interpret the results remembering how participants encountered their data. The surveys likely aided in inducing reflection; they created an artificial requirement for people to look at and think about their search history in a way they might not have otherwise done. Further, it is possible there was a novelty effect. While there may have been confounding influences, we believe their role was minimal compared to the interface itself because behavioral changes occurred for the compare condition, and not for the non-compare.

Another important consideration is how experts were identified. Our approach of selecting the users who most often employed search operators likely misidentifies some experts and misses others. However, the approach was correct often enough that the aggregated behavior patterns of expert and typical archetypes were consistent with previous findings [27]. Regardless of how experts are characterized, our work makes clear that search behavior can be impacted by reflection, and the comparison with the archetypes we created were essential for the observed changes.

We also do not know whether the Dashboard necessarily helped participants become better searchers or to improve their search performance. For example, users may try to use more query terms than necessary, leading a search engine to return results that are less relevant. This would mean participants are worse off because of their changed search behavior. In this first study, our primary research questions relate to understanding how people make use of different types of data, and whether or not it can impact search behavior and attitudes. However, related work has suggested that the knowledge and use of particular search engine functionality and other characteristics of search engine use can predict improved searching performance [2,27], and that people can infer the thought process of others by imitating behavior [8]. We hypothesize that people may have been able to *think* more like search experts, can apply their new knowledge appropriately, and did increase their search performance; but we leave this evaluation to future work.

Designing Search Feedback for Reflection

Reflecting on search behavior can lead to changes in terms of behavior and attitudes about search, and therefore should be used by search engines. Given recent interest in providing search feedback and in educating searchers, we believe our approach can be used with little change. A Search Dashboard-like system would fit in well with existing search engine history facilities. Because all of the data used in the Dashboard are currently collected by search engines, the development cost of such a feature would be low.

It is also interesting to consider how users would encounter and use a Search Dashboard system if it were part of search engines. We feel that current search history functionality may not be widely used, except in specific situations, such as trying to refind a previously viewed website. However, the Dashboard offers a very different, and likely more engaging, view of personal search behavior. We feel that upon learning about a new Dashboard-like system many people would be inclined to test its functionality. And, as we demonstrated in our study, even a single session is sufficient to affect people's attitudes and search behaviors. We believe that after an intial encounter people would use the Dashboard as a reference when they have difficulties or an opportunity arises. This being said, it is also interesting to consider how aspects of the Dashboard may be inserted directly into the search experience for more frequent user contact, as a reminder or for encouragement to improve their Techniques, Tendencies or Topics. Rather than a "Tip of the Day" style widget, reminders should be personally meaningful to capture user interest and suggest a target for change. For example, after a period of observed search engine use, a widget could be added directly to search results that says: “You have used fewer terms in your queries this month than last month. Search experts average more query terms than novices.”

CONCLUSION

We presented and studied the Search Dashboard, a search history feedback system that lets users reflect on their own search behavior and, in some cases, see how that behavior compares to others. The findings of our user study provide evidence that personal history data can lead users to changes in both search behavior and attitudes about search. Participants found their interactions with the Search Dashboard valuable and particularly valued viewing information about their Techniques and Tendencies. Also important to participants was being able to compare their behavior with that of others, and led to changes in aspects of their search behavior to better match search experts. The strong results of our study demonstrate the potential that reflective interfaces have to help people learn to better utilize search engines.

REFERENCES

1. Aula, A., Khan, R.M. & Guan, Z. How does search behavior change as search becomes more difficult? *Proc. CHI,* 2010, 35-44.
2. Aula, A. & Nordhausen, K. Modeling successful performance in web searching. *JASIST 57,* 12 (2006), 1678-1693.
3. Aula, A., Jhaveri, N. & Kaki, M. Information search and re-access strategies of experienced Web users. *Proc. WWW*, 583-592.
4. Bandura, A. *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ, US: Prentice-Hall, 1986.
5. Bennett, P., Svore, K. & Dumais, S. Classification-enhanced ranking. *Proc. WWW*, 2010, 111-120.
6. Brand-Gruwel, S., Wopereis, I. & Vermetten, Y. Information problem solving by experts and novices: Analysis of a complex cognitive skill. *Computers in Human Behavior*, *21*, (2005), 487-508.
7. Boud, D., Keogh, R. & Walker, D. *Reflection: Turning Experience into Learning.* Routledge, 1985.
8. Collins, A., Brown, J.S. & Newman, S.E. Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. *Knowing, Learning and Instruction 8*, *1* (1989), 453-494.
9. Consolvo, S., McDonald, D. & Landay, J. Theory-driven design strategies for technologies that support behavior change in everyday life. *Proc. CHI*, 2009, 405-414.
10. Dou, Z., Song, R. & Wen, J. A large-scale evaluation and analysis of personalized search strategies. *Proc. WWW*, 2007, 581-590.
11. Evans, B. M. & Chi, E.H. An elaborated model of social search. *IP&M, 46*, 6 (2010), 656-678.
12. Fogg, B.J. Motivating, influencing, and persuading users. In *The human-computer interaction handbook*, Julie A. Jacko and Andrew Sears (Eds.). L., (2002), 358-370.
13. Froehlich, J., Findlater, L. & Landay, J. The design of eco-feedback technology. *Proc. CHI*, 2010, 1999-2008.
14. Hölscher, C. & Strube, G. Web search behavior of Internet experts and newbies. *Computer Networks 33*, (2000), 337-346.
15. Khan, K. & Locatis, C. Searching through the cyberspace: the effects of link display and link density on information retrieval from hypertext on the World Wide Web. *JASIST*, *49*, 2 (1998), 176-182.
16. Kodagoda, N. & Wong, B.L.W. Effects of low and high literacy on user performance in information search and retrieval. *Proc. BCS-HCI*, 2008, 173-181.
17. Lazonder, A.W., Biemans, H.J.A. & Worpeis, I.G.J.H. Differences between novice and experienced users in searching information on the World Wide Web. *JASIST 51*, 6 (2000), 576-581.
18. Li, I., Forlizzi, J. & Dey, A. Know thyself: monitoring and reflecting on facets of one's life. *Proc. CHI EA*, 2010, 4489-4492.
19. Moraveji, N., Morris, M.R., Morris, D., Czerwinski, M. & Riche, N. ClassSearch: Facilitating the development of web search skills through social learning. *Proc. CHI*, 2011, 1797-1806.
20. Moraveji, N., Russell, D., Bien, J. & Mease, D. Measuring improvement in user search performance resulting from optimal search tips. *Proc. SIGIR*, 2011, 355-363.
21. Morris, D., Morris, M.R. & Venolia, G. SearchBar: a search-centric web history for task resumption and information re-finding. Proc. *CHI*, 2008, 1207-1216.
22. Nielsen, J. Incompetent research skills curb users’ problem solving. *Alertbox.* April 11, 2011. (available at: http://www.useit.com/alertbox/search-skills.html)
23. Saito, H. & Miwa, K. A cognitive study of information seeking processes in the WWW: Effects of searcher’s knowledge and experience. *Proc. WISE*, 2001, 321-333.
24. Spärck-Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28,* 1 (1972), 11-21.
25. Thaler, R.H. & Sunstein, C.R. *Nudge: improving decisions about health, wealth, and happiness*. Yale, 2008.
26. Van Kleek, M., Moore, B., Xu, C., & Karger, D.R. Eyebrowse: Real-time web activity sharing and visualization. *Proc. CHI EA*, 2010, 3643-3648.
27. White, R.W. and Morris, D. Investigating the querying and browsing behavior of advanced search engine users. *Proc. SIGIR*, 2007, 255–262.
28. White, R.W., Bilenko, M., and Cucerzan, S. Studying the use of popular destinations to enhance web search Interaction. *Proc. of SIGIR*, 2007, 159–166.
29. White, R.W., Dumais, S., and Teevan, J. Characterizing the influence of domain expertise on web search behavior. *Proc. WSDM*, 2009, 132–141.