Individual Differences in Gaze Patterns for Web Search

Susan Dumais
Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
sdumais@microsoft.com

Georg Buscher
DFKI
Knowledge Management Dept.
Kaiserslautern, 67663, Germany
georg.buscher@dfki.de

Edward Cutrell
Microsoft Research India
196/36 2nd Main, Sadashivnagar
Bangalore, 560 080, India
cutrell@microsoft.com

ABSTRACT

We investigate how people interact with Web search engine result pages using eye-tracking, to provide a detailed understanding of the patterns of user attention. Previous research has examined the visual attention devoted to the 10 organic search results, and we extend this by also examining how gaze is distributed across other components of contemporary search engines, such as ads and related searches. This provides insights about searcher's interactions with the "whole page", and not just individual components. In addition, we used clustering techniques to identify groups of individuals, with distinct gaze patterns. The groups varied in how exhaustively they examined the search results and in what regions of the search result page they paid most attention to (organic results vs. ads). These results further our understanding of how attention is distributed across increasingly complex search result pages, and how individuals exhibit distinct patterns of attention and interaction.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems – *Human information processing, Human factors.*
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Search process, Selection process.*
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *User issues.*
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Evaluation/methodology, Interaction styles.*

General Terms

Experimentation, Human Factors.

Keywords

Eye-tracking, Individual differences

1. INTRODUCTION

In developing interactive retrieval systems it is important to go beyond the analysis of off-line measures of the relevance of individual results to the query, e.g., popular measures such as precision and recall, or discounted cumulative gain (DCG). While these measures provide some indication of the quality of the results, they do little to elucidate our understanding of how searchers interact with results. In the research reported in this paper, we use gaze tracking to enable us to understand detailed patterns of user attention to and interactions with search results.

Previous studies have used eye-tracking to understand how people attend to different elements of search engine result pages (SERPs). This work has developed well-known terms to describe typical gaze distributions on SERPs, such as the "golden triangle" [12]. Figure 1 shows an example of a characteristic heat map for a SERP, with the most visual attention being devoted to the first result along with the next few results. These studies tend to be fairly high-level, with qualitative descriptions of gaze behavior aggregated across participants and tasks. Other researchers have taken a more controlled experimental approach and reported quantitative summaries of eye movements on SERPs, often explicitly controlling the search tasks that people are asked to conduct. These studies characterize how visual attention is distributed on the 10 organic results, e.g., [6][9][13][16][18]. However, all of today's major commercial search engines include

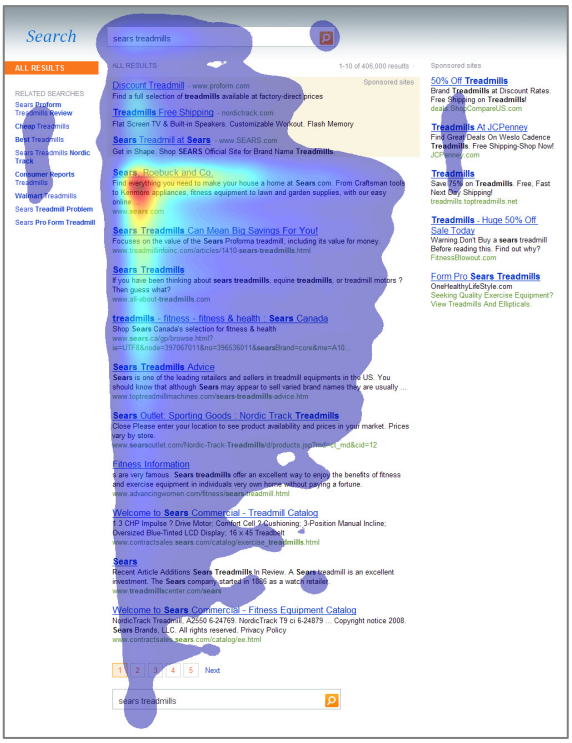

Figure 1: Gaze heat map on a search engine results page.

IIiX 2010, August 18–22, 2010, New Brunswick, NJ, USA.

additional elements on a SERP such as sponsored links or ads at the top and on the right, related searches, graphical elements such as images and maps, deep links, etc.

In the research reported in this paper we examine how the visual attention devoted to organic results is influenced by these other page elements, in particular ads and related searches. This provides an initial understanding how individual elements combine to create a "whole page" experience. In addition, most previous eye-tracking studies have reported aggregate data, but here we look at individual differences in how searchers distribute their attention to different elements on the SERP. By examining in detail how people attend to search engine results pages and indentifying distinct patterns of visual attention and interaction, we can provide a richer understanding of information seeking behavior and interaction with retrieval systems.

After presenting an overview of related research, we describe the experimental design and methods for our eye-tracking study. We then provide an analysis of individual differences in the amount of attention devoted to different regions of a search results page, with three distinct clusters of interaction behavior being identified. The relation of these clusters to search strategies, task behavior and questionnaire data is further explored. We conclude with a summary of the implications of the results and some directions for future research.

2. RELATED WORK

Two general lines of research are related to our work – studies that have used eye-tracking methods to examine how people attend to search results pages, and research that has examined individual differences in information seeking strategies and search behaviors.

There is a long history in information science of understanding individual differences in search strategies, tactics, and performance (see Saracevic [19] for an overview). Allen [1] and Ford et al. [7][8], for example, identified several differences among web searchers that influence search strategies and task performance. Important dimensions included prior experience, gender, age and cognitive styles. More recently, Gwizdka [10] and Kules et al. [15] have examined the relationship between different interfaces and search behavior. Bhavnani [3] and Thatcher [20] used a combination of qualitative and quantitative methods to identify web search strategies. Bhavnani further investigated how domain expertise influences the choice of search strategies and task success. These studies provided detailed modeling but involved only a small number of tasks. On the other end of the spectrum, large-scale log analyses involving millions of users and a wide range of tasks have examined the relationship between search expertise (White et al. [23]) and domain expertise (White et al. [22]) on web search behaviors.

In the last few years, several groups have used eye-tracking to provide detailed quantitative analyses of eye movements as people examine Web pages in general (e.g., [4]) and search engine results pages more specifically (e.g., [6][9][13][16][18]). Since eye position is highly correlated with visual attention, these studies provide a unique insight into what people are doing as they interact with search result pages or destination Web pages. Most of these studies characterized how visual attention is distributed on the 10 organic results. For example, Joachims et al. [13], Guan and Cutrell [9] and Pan et al. [18] showed that the way in which searchers examined search results was influenced by the position and relevance of results. Searchers have a strong bias towards results presented at the top of a SERP. Cutrell and Guan [6] examined how gaze duration is influenced by the length of snippets used to present search results. However, all of today's major commercial search engines include additional elements on a SERP such as sponsored links or ads at the top and on the right, related searches, graphical elements such as images and maps, deep links, etc. Recent work by Buscher et al. [5] examined how visual attention is distributed among different elements on the search results page (e.g., organic results, top ads, right ads, related searches, etc.). They find that most of the attention is devoted to the top search results (like the positional bias noted above), but that there is also substantial attention to ads at the top of the page and that the amount of attention devoted to elements is influenced by their quality. In particular, poor quality ads received less visual attention than good ads, and both types of ads received less attention when their quality was unpredictable across trials. All of these results summarize behavior aggregated over all participants.

A few studies have examined individual differences in gaze patterns. Klockner et al. [11] examined the order in which participants look at the top 25 search results in order to identify relevant results. By hand-coding video records, they found that 52-65% of participants used a depth-first strategy (in which they opened potentially relevant items as they encountered them), 11-15% used a breadth-first strategy (in which they looked at the entire list before selecting a result), and 20-37% used a mixed strategy (in which they looked somewhat ahead before selecting a result). Aula et al. [2] used a more carefully controlled experimental procedure in which the initial search results were fixed for each task. They identified two patterns that people used in examining search results – exhaustive evaluators (54% of the participants, who looked at more than half of the visible results for more than half of the tasks), and economic evaluators (46% of the participants). Finally, Lorigo et al. [17] examined in detail the sequence and patterns of gaze actions. They found that the type of search task (information vs. navigational) influenced task completion time and time on documents, but that gender did not have large effects.

The research reported in this paper builds on and extends previous work on understanding individual differences in detailed gaze behaviors on search result pages. Instead of focusing on just the organic results, we examine user interactions with the whole page including results, ads, and related searches. In addition we use clustering techniques to identify groups of individuals who exhibit similar patterns of visual attention.

3. METHODS AND DATA COLLECTION

We use eye-tracking as an instrument to provide detailed information about the searcher's visual attention. Eye-tracking data can provide valuable insights about search strategies and processes. We supplemented this very detailed gaze data with task completion measures as well as subjective measures of search engine quality and search strategies.

Participants in the experiment completed 32 search tasks using a Web search engine. Three variables were manipulated in the experiment – task type (informational or navigational), the quality of ads, and the order in which ads of different quality were shown. In a previous paper we reported aggregate results examining the effects of ad quality and ordering on gaze duration and task success (Buscher et al. [5]). In this paper we build on results from the same experiment but focus on individual differences in gaze patterns. For the analyses reported in this paper, we collapse

Table 1: Examples of task descriptions and initial queries.

Task Description	Initial Task Query	Task Type
Find the official website of Tesla Motors – a startup that builds powerful electronic cars.	tesla electric cars	Nav
Find the symptom checker webpage of WebMD.	symptom checker web md	Nav
What is the size of a modern implantable pacemaker of today?	heart pacemaker size	Info
What basic equipment do you need for kite surfing?	kite surfing equipment	Info

across the task and ad quality variables, which were counterbalanced across participants. In addition, unlike most previous eye-tracking studies which looked at gaze patterns on the ten organic search results, we examine search behavior using a more realistically composed search results page, which includes ads and related searches (see Figure 2, described in more detail below).

We now describe our experimental design and the behavioral measures that we examined in more detail. See Buscher et al. [5] for additional experimental details and previous results.

3.1 Experimental Design and Procedure

Tasks

Each participant had to solve the same set of 32 search tasks. Half of the tasks were navigational and half were informational. All of the tasks were of a commercial nature so that ads would be a realistic component of the SERPs. Ad quality was varied across tasks (and counterbalanced across participants), and all ads on a SERP were either good or bad.

Each task included a short description of what the participants should look for. In order to make the initial SERP comparable across participants, we provided them with an initial query for each task. Some examples of task descriptions and the corresponding initial task queries are given in Table 1. After the initial SERP was presented, participants were free to proceed as they wished. They could click links, view the next page of results, or re-query. The combination of an initial fixed SERP and full search functionality provides a good balance between experimental control and search realism for a laboratory study.

We cached results for each initial query. This allowed us to have a consistent initial set of results for each task. All tasks contained at least one solution to the task on the first page of organic results, although participants did not know this and sometimes viewed additional pages or generated new queries. For 24 (75%) of the tasks, the static first SERP contained a solution within the top 3 organic results, for 6 tasks (19%), a solution could be found in positions 4-6, and for 2 tasks (6%), a solution was after position 6. Depending on ad quality, solutions could sometimes also be found in the ads. The tasks were pretty simple taking 2-3 minutes to complete, on average.

Search Page Composition

The layout of the SERPs was modeled after a commercial Web search engine. As depicted in Figure 2, a SERP contained the following important elements:

- upper and lower search boxes,
- 10 organic results (not containing any special elements like maps, videos, images, or deep links),
- 3 top ads and 5 right ads, and
- related searches on the left rail for queries for which they were available (20 of 32 initial queries contained related searches).

To generate the SERP for a query, we implemented our own search interface shown in Figure 2. For the initial task query the interface showed a locally cached version of the first SERP for the query. For any other user-generated query, the interface queried a commercial Web search engine in the background, took the organic results and the related searches (if any), inserted ads, and displayed them using our modified interface layout.

We controlled the SERP generation process so that the aforementioned SERP elements were contained on every SERP that was presented to the participant (except for the related searches that were only present on 63% percent of the initial SERPs). Any other advanced interaction techniques that are sometimes available for commercial Web search engines were turned off.

Procedure

After a short introduction to the study, the eye tracker was calibrated using a 5-point calibration. Then, the participants started with one practice task to illustrate the procedure and continued in the same way for the remaining 32 tasks.

For each task, we provided the participants with a written task description and the corresponding initial query. After reading the description and the query aloud, the participants pressed a search button to begin searching using the initial query. The first SERP was always the cached version. From here on, participants were free to interact with search results. To solve the task, they had to navigate to an appropriate Web page and point out the solution on it to the experimenter. After finding a solution, they had to answer the question: "How good was the search engine for this task?" (5-point Likert scale).

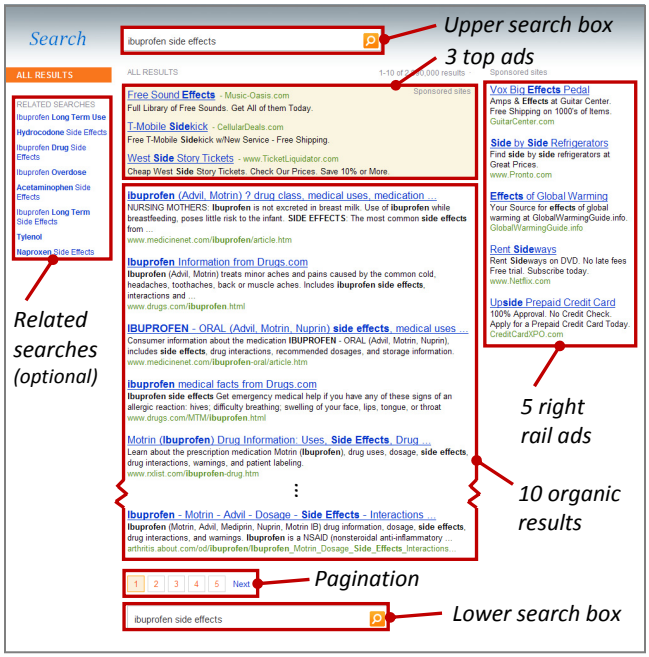

Figure 2: Layout of our search engine result page (SERP). The main areas of interest (AOIs) are identified.

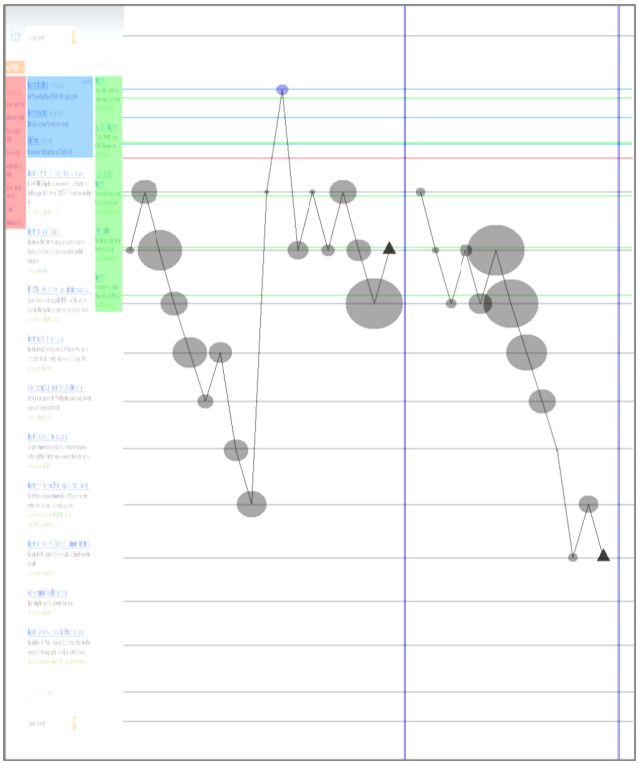

Figure 3. Scan path representation of the sequence of searcher gaze and click activities. Circles represent gaze; the diameter of the circles corresponds to fixation time. Triangles signify clicks. On the far left, a miniature version of the corresponding SERP is shown.

After completing the practice task and all 32 main tasks, the participants completed a questionnaire asking about their Web search experience and practices during the study and in general. The experiment took about one hour per participant.

3.2 Apparatus

The experiment was performed on a 17" LCD monitor (96 dpi) at a screen resolution of 1280x1024 pixels. We used the Internet Explorer 7 browser with a window size of 1040x996 pixels. With this setting, the page fold was usually between the organic results at positions 6 and 7. For gaze tracking, we used a Tobii x50 eye tracker which has a tracking frequency of 50 Hz and an accuracy of 0.5° of visual angle. Logging of click and gaze data was done by the software Tobii Studio.

3.3 Participants

Participants were recruited from a user study pool. They ranged in age between 26 and 60 years (mean = 45.5, σ = 8.2), and had a wide variety of backgrounds and professions. 21 participants were female and 17 were male.

3.4 Measures

For our analysis, we wanted to know how the different elements on the SERPs attract visual attention as well as how clicks are distributed among them. Therefore, we assigned gaze and click data to areas of interest (AOIs) on the SERPs.

3.4.1 AOIs

Since all SERPs presented during the study had the same kinds of elements, we created common areas of interest. All regions labeled in Figure 2 were defined as AOIs. For the top ads, the right rail ads, and the organic results, we introduced further AOIs matching the different result entries (i.e., separate AOIs for each of the 10 organic results, for the 3 top ads and for the 5 right rail ads).Fixation impact on each of these AOIs provides the input for our clustering analyses.

3.4.2 Fixation Impact

Fixations were detected using built-in algorithms of Tobii Studio. The algorithms generate a fixation if recorded gaze locations of at least 100ms are close to each other (within a radius of 35 pixels).

We used the measure fixation impact *fi(A)* to determine the amount of gaze an AOI *A* received. This measure was introduced by Buscher et al. [4] and is a modified version of simple fixation duration. Fixation duration assigns the entire duration to the AOI(s) that contain the center point of the fixation, but the fixation impact measure spreads the duration to all AOI(s) close to the fixation center using a Gaussian distribution. Thus, fixation impact prorates the duration of a fixation to all AOIs that are projected on the foveal area of the eyes. We also measured clicks in each AOI, *c(A)*.

3.4.3 Gaze Sequences and Scanpaths

AOIs are very useful in understanding where searchers are spending their time, but do not say anything about the sequence of activities. To better understand the temporal sequence of activities, we developed a visualization based on a technique used by Aula et al. [2]. We extend this by incorporating activities in different areas of interest, clicks and multiple page views.

Figure 3 shows a visualization of the sequence of gaze activities, for one participant for one task. The y-axis represents the position on the SERP. A small version of the SERP with different regions color coded is shown on the far left to provide a visual reference for interpreting click positions. Organic results are shown in grey, top ads in blue, right ads in green, and related searches in red. The x-axis represents the temporal order and duration of gaze and click events. Circles represent a fixation within an AOI, with the size being proportional to the duration of the fixation. Triangles represent a click. Vertical blue lines denote individual SERP visits.

Figure 3 shows data for participant 1 on task 20. The participant starts by looking briefly at results 2, then result 1, then back to result 2 (where they now spend more time), gradually goes down to result 7, changes direction and goes to the first ad (small blue circle), looks again at the first few results before clicking on the second result which ends the first view on the SERP. The result was apparently not relevant so they returned to the SERP, and scanned down to result 7, clicked on it, and found what they were looking for.

Using this representation of search scanpaths, we define several abstractions of search strategies that we believe will provide insights about how different individuals examine the SERP as a whole. For example: Do people search systematically from top to bottom? Do they look at many results before clicking? Do they focus only on results or do they also attend to other elements of the page such as ads or related searches? Are these behaviors consistent across individuals?

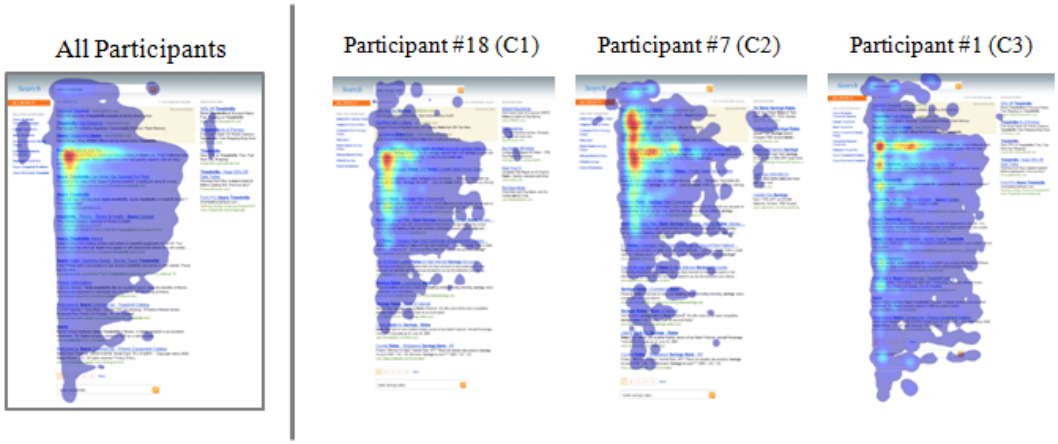

Figure 4. Comparison of an aggregate heat map for all participants (left), and individual heat maps from three participants (right). P18 is in Cluster 1 (economic examination, time on organic results 1-3); P7 is in Cluser 2 (economic examination, time on ads especially top ads); and P1 is in Cluster 3 (exhaustive examination).

3.4.3.1 Ecomomic vs. Exhaustive Evaluation

Aula et al. [2] differentiated between two groups of users: economic and exhaustive evaluators. A participant was classified as an *economic evaluator* if he or she scanned at most half of the result entries visible on the screen without the need to scroll (in their case at most 3 results) for over 50% of the tasks. Otherwise, they were classified as *exhaustive evaluators*.

Inspired by this measure, we call a single SERP evaluation an *economic evaluation*, if and only if at most half of the results on the SERP above the fold (including the 3 top ads and the organic results, but not including the right rail ads) were inspected before the first click (in our case at most 4 results). Otherwise, we call it an *exhaustive evaluation*. We do this for all tasks and compute the proportion of economic evaluations for each participant.

3.4.3.2 Completeness

Similar to Lorgio et al. [17], we call a scanpath *complete* if a participant inspected all of the result entries above the clicked entry. However, for the determination of scanpath completeness, we focus on two different sets of result entries on a SERP. We say that a scanpath is *complete-organic* if it is complete with respect to the 10 organic results (all other elements on a SERP are simply ignored). Furthermore, a scanpath is *complete-all* if it is complete with respect to the top 3 ads and the 10 organic results.

3.4.3.3 Linearity

For the computation of linearity, we adopt the same basic definitions as introduced by Lorgio et al. We obtain a numbered *scan sequence* by assigning numbers to all top ads (i.e., -2, -1, 0) and organic results (i.e., 1, 2, …, 10) and using these numbers to describe the scanpath. For example, the beginning of the scanpath shown in Figure 5 can be represented by the sequence "2, 1, 2, 3, 4, 5, 4, 6, …". The *minimal scan sequence* can be obtained by removing repeat visits to a result entry. For example, the scan sequence from above would turn into "2, 1, 3, 4, 5, 6, …". Similar to Lorgio et al., we define a scanpath to be *linear* if the minimal scan sequence is monotonically increasing in steps of 1. Likewise, a scanpath is *strictlyLinear*, if the scan sequence is monotonically increasing in steps of 1.

Again, we define a scanpath to be *linear-organic* and *strictlyLinear-organic* if the scanpath is linear / strictlyLinear when only considering organic result entries in the scan sequence. When considering top ads as well as organic result entries in the scan sequence, then a scanpath may be *linear-all* or even *strictlyLinear-all*.

3.4.3.4 Change of Scan Direction

In addition, we measure how often a participant scanned up or down a result list until the first click. The measures *ScanUp* and *ScanDown* count the number of times a participant began scanning upwards or downwards a SERP. Whether a participant scanned downwards is determined as follows:

a) Either two subsequent transitions from one result entry to the next have the same downward direction (e.g., from position 3 to 4 and then from 4 to 5).

b) Or a transition between two result entries skips at least one result entry in between (e.g., from position -2 to 0).

Scan sequences in the upward direction are determined analogously. The measures ScanDown and ScanUp count the number of times the scan sequence changed to the downward / upward direction. For example, the scan sequence in Figure 3 contains two downward and one upward scans before the first click.

3.4.3.5 Number of Gaze Actions

In order to further our understanding of the dynamics while viewing result entries, we compute additional simple gaze measures relative to an AOI A. *GazeEntries* on A counts the number of times the gaze moves into AOI A from any other region of the SERP. *GazeEventsBeforeFirstEntry* / *BeforeLastEntry* on A measures the number of gaze changes between AOIs before first / last gazing at A.

3.4.4 Task-Level Measures

Although we focus on measures that seek to understand the location and sequence of visual attention over SERPs, we also measured task-level summaries including overall task completion time, errors, number of queries, number of clicks, etc.

3.4.5 Questionnaire Measures

We also supplemented the detailed gaze tracking data with some measures of subjective impressions. After each search task participants were asked to the following question: "How good was the search engine for this task?" (5-point Likert scale).

In addition, at the end of the experiment, participants completed a short questionnaire asking about their web and search experience, their overall impression of the search engine used in the experiment, and their general search practices (e.g., "If I can't find what I am looking for in the top 1 or 2 result entries, I usually [look further down | go to the next page | click on related searches | try another query]").

4. RESULTS

We begin by showing some examples of individual differences using heat maps. We then take a more fine-grained look at the distribution of attention across different AOI regions, and use this as input to a clustering algorithm to identify consistent patterns of user interaction. Finally, we discuss the relationship between the identified clusters and other measures including fixation time, the temporal order of gaze patterns (as characterized by scanpaths), task performance, and questionnaire data.

4.1 Heat Maps

Heat maps are often used to visually summarize gaze patterns. Figure 1 and Figure 4 (left) show the overall heat map for our experiment averaged over more than 1200 search tasks, comprising 38 participants each of whom conducted 32 search tasks. Color is used to represent the overall amount of attention, ranging from red (most) to blue (least). As can be seen, most attention (red) is devoted to the left portion of the first organic result, with some attention (yellow) to the right of the first result and the second and third results, less attention (light blue) to results further down the list and to the top ads, and very little attention (dark blue) to other regions.

Figure 4 (right) also shows the heat maps for three individual participants. Even though all participants completed exactly the same 32 tasks, they show quite different aggregate gaze patterns. As we describe in the next section, these participants represent three main clusters of searchers identified in our analyses. It is these individual differences that we seek to understand in more detail.

Although heat maps generate interesting pictures that provide some intuitions about aggregate behavior, they are not that helpful in understanding user interaction with specific page elements. To examine this in greater detail we use the breakdown of a search result page into areas of interest (AOIs), as shown in Figure 2. Further we cluster participants using the way in which they distribute their attention to different AOIs.

4.2 Areas of Interest (AOIs) and Clusters

In this section we describe our clustering analysis of participants. For these analyses, we first summarize an individual's search behavior using the fixation impact on the different areas of interest. Specifically, we compute average fixation impact over all 32 tasks for an individual. We represent each participant using the average fixation impact for the main AOIs (shown in Figure 2), and further divide the organic results into three groups reflecting how far down the results list individuals looked (Results 1-3, Results 4-6, and Results 7-10). (As noted in Section 3, all participants completed the same number of navigations and information tasks and the same number of tasks with good and bad ads.)

We summarized attention to different page elements using both the average fixation impact for each AOI, and a version of these times normalized by the total fixation time for an individual. The normalized data reflects the proportion of time spent on different regions of the page independent of the overall amount of time taken. Both analyses resulted in exactly the same clusters, so in this paper we describe analyses obtained using the normalized data.

We also examined search behavior for the entire task (i.e., for all page visits and for all queries needed to find an answer), and for the first visit to the first SERP. Because the first SERP is fixed across participants, it makes the results easier to interpret. (For 74% of the queries, participants found the desired answer on their first page visit.) The overall findings are quite similar when we look at all page visits or just the first page visit, so we focus on the first SERP visit but also compare to the more global behavior.

To summarize, the data we use as input to the clustering analysis is based on the average fixation impact on each of the main AOIs shown in Figure 2 (with the results further broken down by Results 1-3, 4-6 and 7-10) for an individual across all 32 tasks. In the main results presented below, we use fixation impact on the first SERP visit (which is the same of all participants) and normalize the fixation impact for each AOI by the total fixation impact for the individual.

4.2.1 Clusters of participants

We used the Cluto clustering package [14] to identify groups of participants who shared similar distributions of attention to AOIs on the search result page (as represented by the normalized AOI data described above). Specifically, we use repeated-bisection clustering with a cosine similarity metric and the ratio of intra- to extra- cluster similarity as the objective function. In practice we find that clusters are fairly stable regardless of the specific clustering or similarity metric. By varying the number of clusters and testing within- and between-cluster similarity we find that the objective function levels off at around 3 clusters, so we use 3 clusters in the analyses below.

Table 2 (top section) shows some general characteristics of the clusters and participants in each cluster. As we describe in more

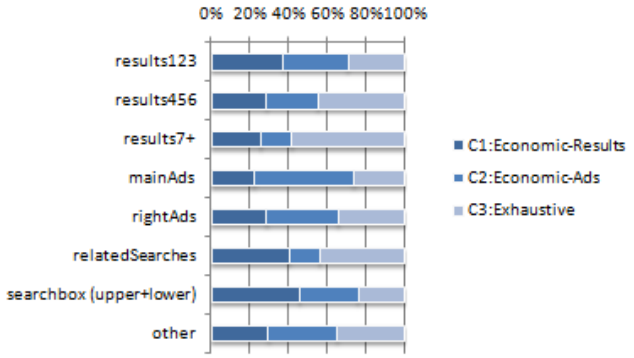

Figure 5. Proportion of time spent on each AOI, for each cluster. Data is normalized by row (AOI elements) to show which cluster of participants spends proportionally more time for individual page elements.

detail below, there are three main clusters of participants, those who explore the SERP broadly (which we call C3: *Exhaustive* searchers), and those who explore more narrowly – further broken down by those who also look at some additional results (C1: *Economic-Results*) and those who also regularly look at ads (C2: *Economic-Ads*). Participants are pretty evenly split among clusters, with 15 (39%), 11 (29%), and 12 (32%) participants respectively. Participants in the Exhaustive group are somewhat older and less experienced with computers, the internet, and search engines based on their responses to the questionnaire.

Table 3 and Figure 5 show summary measures for the three clusters, using the normalized data. Table 3 (AOI-normalized section) shows the proportion of total gaze time spent in each AOI. Not surprisingly, participants in all clusters spend the largest proportion of their time on results 1-3, although the proportion that this represents ranges from 68% for the Economic-Results group to 54% for the Exhaustive group. Participants in the Economic-Results group spend more than 10% of their time looking at results 4-6, and little time in other regions – they are very much focused on the top few results. Participants in the Economic-Ads group spend almost 20% attending to the main ads – this group spends almost 80% of their time on the top results and ads. Finally participants in the Exhaustive group spend more than 20% of their time attending to results 4-6, more than 10% of their time on results 7-10, and more than 9% on the main ads – they spread their attention broadly throughout the page. In all groups less than 3% of attention is devoted to right ads, related searches or the search boxes.

Figure 5 shows the same data, normalized by row (AOI elements). This allows us to see which cluster of participants spends proportionally more time for individual page elements. The Exhaustive group (light blue) spends the most time proportionally in results 4-6, results 7+, and related searches. On average, they examine the organic search results more thoroughly than participants in the Economic groups. The Economic-Ads group (medium blue) spends proportionally more time attending to ads, especially the main ads at the top of the page. Finally, the Economic-Results group (dark blue) spends proportionally more time on results 4-6 and results 7+ (compared to the Economic-Ads group), and more time on related searches and the search box.

It is important to note that all participants conducted exactly the same 32 search tasks, so that observed differences are the result of individual characteristics or experiences and not the result of task differences per se.

4.3 Relationship between Clusters and Other Measures

We now consider the relationship between the clusters of searchers identified using the relative amount of time on different AOIs with other measures, including total fixation impact, scanpaths, task outcomes and questionnaire data.

4.3.1 AOI – Fixation Impact Time

Table 3 (AOI – raw times section) shows the total fixation impact on the AOI regions for participants in each of the clusters. Exhaustive participants are the slowest in terms of overall time (totalFixationImact), an average 14633 msecs. They examine more information and it takes them more total time to do so. Participants in the other two clusters are faster, taking 32% and 40% less time, respectively for Economic-Results and Economic-Ads.

All three groups spend the most absolute time on results 1-3, again with the Exhaustive group being substantially slower than the other two groups by 11% and 32%, respectively. The Exhaustive and Economic-Results group spend the next most amount of time on results 4-6. Interestingly, the Economic-Ads group spends more time on the main ads than on results 4-6 (1625 vs. 1169 msecs on average). They spend more than twice as much time on the main ads as the Economic-Results group and even more time on main ads than the Exhaustive group, even though the Exhaustive group spends more time on all other measures.

4.3.2 Scanpath Analyses

We now consider the temporal dynamics of user attention, examining not just where people look but also the order in which they do so. Table 3 (ScanPath section) shows two measures of the number of gaze events before the initial click, two measures of the extent of gaze (min and max position), as well as all scanpath strategies, described earlier in Section 3.4.3. Exhaustive searchers exhibit 30% more gaze events before their initial click than searchers in the other two groups (13.2 vs. 8.6 and 9.3), and they also show the highest maximum gaze position on average (4.3 vs. 3.6 and 3.2). Economic-Ads searchers show the lowest minimum gaze position (-1.17), indicating that they look at more than one main ad, on average. (Gaze position for main ads is encoded -2, -1, 0 for the first, second and third ad, respectively.)

Using our modification of Aula et al.'s [2] definition of economic vs. exhaustive examination of search results, we see a good correspondence between the cluster assignments and the extent to which people consider fewer than 4 results before their first click. Participants in the Exhaustive cluster considered fewer than 4 results on only 36% of the trials, compared with 42% and 43% for participants in the two Economic clusters.

The measures counting gaze events for individual elements are summarized at the bottom of the ScanPath section. These results show that Economic-Ads searchers not only spend relatively more time on main ads and gaze there more often (1.33 vs. 1.13 and 1.11), but they also go to the ads before they gaze on other page elements (0.60 vs. 1.34 and 1.57). In contrast, Economic-Results and Exhaustive searchers both get to the main results earlier (0.65 and 0.79 vs. 1.01) and to the ads later.

All measures of completeness and linearity in examining the organic results (complete-organic, linear-organic and strictlyLinear-organic) are roughly comparable across clusters indicating that participants generally scan down in similar ways, although the extent of the scan varies across clusters. When considering the sequence of interaction with the ads, there are some differences – e.g., for the complete-all measure the Economic-Results searchers show less of a linear progression from top to bottom since they tend to attend less to the ads.

The ScanUp and ScanDown measures highlight the fact that Exhaustive searchers more frequently look up and down the SERP before their initial click, although the differences are not statistically significant. Conversely, the Economic searchers are more likely to click on the result at the furthest extent of their gaze, Exhaustive searchers more often look further down the list without necessarily clicking.

4.3.3 Task Measures

The analyses we have considered so far have focused on the visual attention devoted to the first search result page. We now look at the extent to which the clusters and other patterns

identified using these data correspond to overall task performance, which may include multiple queries, clicks and page visits.

Table 3 (Task click and time section) shows the results of these analyses. There are no differences in the total number of queries issued in the three groups. However, participants in the Economic-Ads and Exhaustive clusters click on more results and ads, and tend to view more non-SERP pages.

There is a large difference in the overall task completion time, with the Exhaustive searchers taking about 25% longer to complete search tasks on average (63423 msecs vs. 47017 or 48085 msecs, for the two Economic groups). They also take a longer time to make their first click. Finally, as noted earlier, there are no differences in overall accuracy, since participants are not allowed to go on to the next task unless they have found the answer (or a three minute time limit has been exceeded, which happened very infrequently).

4.3.4 Questionnaire Data

Table 2 (Questionnaire section) summarizes the results of the post-experiment questionnaire. There are some relationships between the groups identified using gaze patterns and subjective impressions of the search engine in the study and more general search behaviors, although most of the differences are not reliable statistically.

The Ecomomic-Ads group was the least satisfied overall with the system used in the experiment, for both the overall quality of the search engine and the queries used. The Economic-Results groups' self-reported strategies were consistent with their economic behavior – e.g., in response to the query about what they did when the desired item was not in the top 1 or 2 results they were the least likely to say that they look further down the list, go to next page, click related query or click ads. However, they also reported that they usually looked at top ads as much as the Economic-Ads group, and the reason for this discrepancy is not clear without further investigation.

5. CONCLUSIONS AND FUTURE WORK

In this paper we used an eye-tracking methodology to provide a detailed analysis of the patterns of user attention on realistic search engine result pages which consist of organic results, ads and related searches. We find that most attention is devoted to the top three results, but that substantial visual attention is also placed on the next three results and on the top ads, with less attention devoted to other regions of the search page (related searches, right ads and search and navigational elements) for most tasks. We further developed visualizations and summary measures to characterize the order in which and the depth to which participants scanned search result pages. This broad set of measures provides insights about how people interact with whole search engine pages as well as individual areas of interest.

In addition to these aggregate summaries of visual attention on the SERP page, we also clustered participants into three groups who showed distinct patterns of visual attention. The three main clusters we identified were searchers who explore the SERP broadly (*Exhaustive* cluster, 32%), and searchers who explore more narrowly – further broken down by those who also look at some additional results (*Economic-Results* cluster, *39%*) and those who look regularly at ads (*Economic-Ads* cluster, *29%*). These clusters are also associated with differences in total fixation impact, scanpaths, task outcomes and questionnaire data. By identifying distinct behavioral patterns in search we believe that we can begin to design search interfaces that better support these different search strategies. For example, for exhaustive searchers we might provide capabilities to mark results of potential interest during their initial exploration, or enhance snippets to aid in their decision making process.

This research represents an initial attempt to understand in detail the distribution of visual attention and the sequences of search behaviors. We would like to extend our work to examine the broader space of information needs, searchers and information environments. To do this we will consider a broader range of search tasks, especially more exploratory tasks that take longer to accomplish, and extend our work to include even richer search result presentations, especially involving the integration of different types of results (web pages, news, images, answers, etc.). Finally, we would like to see the extent to which the patterns we identified in the laboratory can be seen in large-scale search logs to provide broader coverage of individuals and tasks *in situ*.

6. REFERENCES

[1] Allen, B. Individual differences and the conundrums of user-centered design. *JASIS, 51(6),* 2000, 508-520.

[2] Aula, A., Majaranta, P. and Raiha, K-J. (2005). Eye-tracking reveals personal styles for search result evaluation. In *Proceedings INTERACT 2005*, 1058-1061.

[3] Bhavnani, S. Important cognitive components of domain-specific search knowledge. In *Proceedings TREC 2001*, 571-578.

[4] Buscher, G., Cutrell, E. and Morris, M. R. What do you see when you're surfing?: Using eye tracking to predict salient regions of web pages. In *Proceedings CHI 2009*, 21-30.

[5] Buscher, G., Dumais, S. and Cutrell, E. The good, the bad, and the random: an eye-tracking study of ad quality in web search. To appear in *Proceedings SIGIR 2010*.

[6] Cutrell, E. and Guan, Z. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings CHI 2007*, 407-416.

[7] Ford, N., Miller, D. and Moss, N. The role of individual differences in internet searching: An empirical study. *JASIST, 52 (12),* 2001, 1049-1066.

[8] Ford, N., Miller, D. and Moss, N. Web search strategies and human individual differences: Cognitive and demographic factors, internet attitudes and approaches. *JASIST, 56 (7),* 2005, 741-756.

[9] Guan, Z. and Cutrell, E. An eye tracking study of the effect of target rank on web search. In *Proceedings CHI 2007*, 417-420.

[10] Gwizdka, J. What a difference a tag cloud makes: Effects of tasks and cognitive abilities on search results interface use. *Information Research, 14 (4),* Dec. 2009.

[11] Klockner, K., Wirschum, N. and Jameson, A. Depth- and breadth-first processing of search results lists. In *Proceedings CHI 2004*, 1539.

[12] Hotchkiss, G., Alston, S. and Edwards, G. Eye tracking study, 2006. Retrieved January 18, 2010 from http://www.enquiro.com/eyetrackingreport.asp.

[13] Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings SIGIR 2005*, 154-161.

[14] Karypis, G. Cluto — a clustering toolkit. www.cs.umn.edu/~cluto, retrieved Jan 2009.

[15] Kules, B., Capra, R., Banta, M. and Sierra, T. What do exploratory searchers look at in a faceted search interface? In *Proceedings JCDL 2009*, 313-322.

[16] Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., Granka, L., Pellacini, F. and Pan, B. Eye tracking and online search: Lessons learned and challenges ahead. *JASIST, 2008, 59 (7)*, 1041-1052.

[17] Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L. and Gay, G. The influence of task and gender on search evaluation and behavior using Google. *Information Processing and Management, 42(4)*, 2006, 1123-1131.

[18] Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G. and Granka, L. In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication, 12*, 2007, 801-823.

[19] Saracevic, T. Individual differences in organizing, searching and retrieving information. In *Proceedings of ASIS 1991*, 82-86.

[20] Thatcher, A. Information seeking behaviours and cognitive strategies in different search tasks on the WWW. *International Journal of Industrial Ergonomics, 36 (12),* 2006, 1055-1068.

[21] Tseng, Y-C. and Howes, A. The adaptation of visual search strategies to expected information gain. In *Proceedings CHI 2008*, 1075-1084.

[22] White, R., Dumais, S. T. and Teevan, J. Characterizing the influence of domains expertise on Web search behavior. In *Proceedings WSDM 2009*, 132-14.

[23] White, R. and Morris, D. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings SIGIR 2007*, 255-262.

Table 2. Participant characteristic and questionnaire results broken down by clusters. Post hoc t-test are shown where the ANOVA was significant at the .05 level.

	Cluster1 Economic-Results	Cluster2 Economic-Ads	Cluster3 Exhaustive	Overall	T-tests 1 vs 2	2 vs 3	1 vs 3
Characteristics							
Number of participants	15	11	12	38			
Age	42.5	43.0	51.5	45.5	ns	<.01	<.01
Percent Female	67%	45%	50%	55%			
I use a computer [rarely (1) - several times a day (4)]	3.7	4.0	3.4	3.7			
I use the Web [rarely (1) - several times a day (4)]	3.6	4.0	3.2	3.6			
I use a Web search engine [rarely (1) - several times a day (4)]	3.5	3.7	3.1	3.4			
Questionnaire [5-point Likert scale from 0 to 4]							
Regarding the study today how good was the search engine overall?	3.67	3.18	3.42	3.45			
Regarding the study today how close were the search terms to what you would have chosen?	3.00	2.64	3.17	2.95			
In general how relevant are the search results to your queries?	3.07	2.82	2.83	2.92			
If can't find in top 1 or 2 results, I look further down on the same results page	3.33	3.45	3.58	3.45			
If can't find in top 1 or 2 results, I go to the next page of search results	1.73	2.64	2.00	2.08	<.05	0.10	ns
If can't find in top 1 or 2 results, I click on a related query	1.53	1.91	2.00	1.79			
If can't find in top 1 or 2 results, I click on an ad	0.47	0.82	1.00	0.74			
If can't find in top 1 or 2 results, I try another query	2.87	3.18	2.75	2.92			
If can't find in top 1 or 2 results, I try another search engine	1.53	1.64	1.42	1.53			
Do you usually look at the related queries from the search engine?	3.13	3.00	3.00	3.05			
Do you usually look at the ads at the top of search engine results pages?	3.27	3.27	2.58	3.05	ns	0.05	0.05
Do you usually look at the ads on the right side of search engine results pages?	2.60	2.45	2.50	2.53			

Table 3. Gaze patterns and task performance results broken down by clusters. Post hoc t-tests are shown where the ANOVA was significant at the .05 level.

	Cluster1 Economic-Results	Cluster2 Economic-Ads	Cluster3 Exhaustive	Overall	T-tests 1 vs 2	T-tests 2 vs 3	T-tests 1 vs 3
AOI-normalized [%]							
results123	0.680	0.610	0.537	0.615	0.06	<.01	<.01
results456	0.134	0.125	0.211	0.156	ns	<.01	<.01
results7+	0.046	0.029	0.105	0.059	0.06	<.01	<.01
mainAds	0.077	0.181	0.091	0.111	<.01	<.01	<.01
rightAds	0.007	0.009	0.009	0.008			
relatedSearches	0.009	0.003	0.010	0.008			
searchbox (upper+lower)	0.026	0.017	0.014	0.020			
other	0.021	0.025	0.024	0.023			
AOI-raw times [in msec]							
totalFixationDuration	9938	8719	14634	11068	ns	<.01	<.01
results123	6795	5194	7653	6602	0.05	<.01	ns
results456	1333	1169	3063	1832	ns	<.01	<.01
results7+	465	261	1689	792	<.05	<.01	<.01
mainAds	740	1625	1478	1229	<.01	ns	0.10
rightAds	76	88	100	87			
relatedSearches	96	30	110	82			
searchbox (upper+lower)	233	135	176	187			
other	200	216	364	256	ns	0.10	<.05
ScanPaths							
gazeEventsBefore1stClick	8.61	9.26	13.20	10.25	ns	ns	<.05
gazeEventsBetweenInitialGazeAndClick	4.06	4.60	7.13	5.18			
minGazePos	-0.75	-1.17	-0.71	-0.86			
maxGazePos	3.64	3.23	4.26	3.72	ns	<.01	<.05
clickPosIsHighestAttendedPos [0: no, 1: yes]	0.63	0.60	0.55	0.60	ns	ns	0.01
economicEvaluation [0: no, 1: yes]	0.42	0.43	0.36	0.41			
complete-organic [0: no, 1: yes]	0.85	0.82	0.84	0.84			
complete-all [0: no, 1: yes]	0.11	0.28	0.21	0.19	<.01	ns	0.09
linear-organic [0: no, 1: yes]	0.75	0.79	0.73	0.76			
linear-all [0: no, 1: yes]	0.27	0.21	0.27	0.25			
strictlyLinear-organic [0: no, 1: yes]	0.22	0.25	0.20	0.22			
strictlyLinear-all [0: no, 1: yes]	0.12	0.10	0.09	0.11			
scanDownCount	1.26	1.25	1.60	1.36			
scanUpCount	0.92	1.01	1.28	1.06			
mainAds: gazeEntries	1.11	1.33	1.13	1.18			
results123: gazeEntries	1.76	1.77	1.99	1.83			
results456: gazeEntries	0.68	0.52	1.04	0.75	ns	<.01	<.05
mainAds: gazeEventsBeforeFirstEntry	1.34	0.60	1.57	1.20	ns	<.01	ns
results123: gazeEventsBeforeFirstEntry	0.65	1.01	0.79	0.80	<.01	ns	ns
mainAds: gazeEventsBeforeLastEntry	3.92	3.09	5.66	4.23	ns	<.05	ns
results123: gazeEventsBeforeLastEntry	4.48	4.97	6.92	5.39	ns	ns	<.05
Task clicks and time							
numberOfQueries	1.13	1.11	1.14	1.13			
numberOfSERP1ClicksOnResultsOrAds	1.09	1.25	1.28	1.20	<.05	ns	<.05
numberOfSERP1ClicksOnResults	1.07	1.14	1.20	1.13			
numberOfSERP1ClicksOnTopAds	0.03	0.11	0.08	0.06	<.01	ns	0.09
numberOfNonSerpPageViews	2.08	2.25	2.41	2.23			
taskCompletionTime [in msec]	47018	48085	63424	52508	ns	<.05	<.05
timeToFirstClick1stSERP [in msec]	9367	8446	13523	10413	ns	<.01	0.02