What Users Do: The Eyes Have It

Paul Thomas[1], Falk Scholer[2], and Alistair Moffat[3]

1. CSIRO and the Australian National University, Australia
2. RMIT University, Australia
3. The University of Melbourne, Australia

Abstract. Search engine result pages – the ten blue links – are a staple of document retrieval services. The usual presumption is that users read these one-by-one from the top, making judgments about the usefulness of documents based on the snippets presented, accessing the underlying document when a snippet seems attractive, and then moving on to the next snippet. In this paper we re-examine this assumption, and present the results of a user experiment in which gaze-tracking is combined with click analysis. We conclude that in very general terms, users do indeed read from the top, but that at a detailed level there are complex behaviors evident, suggesting that a more sophisticated model of user interaction might be appropriate. In particular, we argue that users retain a number of snippets in an "active band" that shifts down the result page, and that reading and clicking activity tends to takes place within the band in a manner that is not strictly sequential.

Keywords: Retrieval evaluation, user behavior, user model

1 Introduction

Web and enterprise search systems process billions of queries per day, making them amongst the most highly-used computing services. A typical service provides a dialog box for query input, and in response generates a *search engine result page*, or SERP, which contains a ranked list of (often) ten query-biased summaries (or *snippets*), together with matching links to the underlying documents. Users are normally presumed to examine the SERP "from the top", reading the snippets in the order they are presented, making a decision about each in regard to the likely usefulness of the underlying document, and *clicking* to access those documents for which the snippet suggests relevance. If the bottom of the SERP is reached, users can access a second page of results for the same query; or reformulate the query to fetch a fresh page of (possibly overlapping) results; or switch to a different search system (again to get possibly overlapping results); or can exit their search entirely. Users might also undertake any of these actions even before reaching the bottom of the first SERP.

In this paper we take a fresh look at this presumed behavior, presenting the outcomes of a user experiment in which gaze tracking was coupled with an instrumented web browser in order to measure both the explicit actions users took while viewing a SERP (clicks and query reformulations), and also their implicit actions, as indicated by their gaze behavior.

2 Models and Metrics

Knowledge of user behavior allows better interfaces to be constructed, and hence allows more efficient searching. Another way we can exploit knowledge of user behavior is in understanding search effectiveness metrics. Given a query and a particular ranking of documents in a SERP in response to that query, it is natural to enquire whether or not the ranking is “good” compared to the SERP generated by some other system, or by some other configuration of the same system. To quantify search effectiveness, a range of metrics have been described, varying from simple ones such as Prec@k (the fraction of the first k documents in the SERP that are relevant) through to complex mechanisms such as normalized discounted cumulative gain (NDCG) [6].

Moffat and Zobel [11] drew a direct relationship between the user’s behavior while reading the SERP and a metric for measuring retrieval effectiveness. They argued that search quality could be measured in units of “expected relevant documents identified per snippet examined”, and that users could be modeled as starting at the top of the SERP, and proceeding from rank i to rank $i+1$ with some probability p, with p adjusted according to the nature of the query and the persistence of the searcher. That is, they suggest that at rank i the user has probability p of next accessing rank $i+1$, and probability $(1-p)$ of exiting the search without examining any document snippets at ranks $i+1$ or beyond. The resultant effectiveness metric, RBP, is formulated as a weighted sum over a vector of relevance-at-rank values,

$$\mathrm{RBP} = \sum_{i=1}^{\infty} \left((1-p)p^{i-1}\right) \cdot r_i \,,$$

where r_i is the relevance of the ith-ranked document as a fractional value between zero and one, with one meaning “completely relevant”.

That is, the RBP metric can be thought of as being a direct consequence of a simple one-state, three-transition, *user model* in which the probability of exiting the reading state remains p throughout. Other possible models are then apparent: in Prec@k, the user model is that users read exactly k of the presented snippets; and in reciprocal rank, or RR, the user is modeled as reading until the first relevant document is encountered, and then exiting the reading state. Similarly, average precision, AP, defined as the average of the precision values at each depth at which a relevant document occurs, can also be regarded as being a *weighted-precision* metric:

$$\mathrm{AP} = \sum_{i=1}^{\infty} \left(\frac{\mathrm{Prec}@i}{R}\right) \cdot r_i$$

where $R = \sum_{i=1}^{\infty} r_i$ is the total relevance in the collection (for this query). That is, AP can also be regarded as being of the form $\sum_i w_i \cdot r_i$, where w_i is the *weight* assigned to the ith-ranked document. In the case of RR and AP the user model is *adaptive*, since the weights w_i are a function not only of i, but of r_i. Robertson [12] proposes a corresponding user model in which (for binary relevance values $r_i \in \{0,1\}$) the user is assumed to proceed through the ranking examining snippets/documents until an identified relevant document is encountered, picked at random from amongst all of the R relevant documents for this query. Other *adaptive* metrics – ones in which the exit probability p is a function of relevance, $p_i = f(r_1..r_i)$ for some relationship $f()$, have also been described [17]. A "to depth k" truncated and scaled version of discounted cumulative gain (DCG) [6] is another example of a static weighted precision metric, in which w_i depends only on i. Note that DCG itself cannot be fitted to this weighted-precision structure, since the set of weights $1/(\log_2(i+1))$ used to discount the relevance scores r_i is a non-convergent series, and would give rise to a model in which the user is expected to read an arbitrarily large number of documents for each query that they pose.

Another metric has also been proposed recently – the *time-biased gain* (TBG) of Smucker and Clarke [13]. Rather than assessing user effort by counting the number of snippets viewed, Smucker and Clarke measure effort in terms of time spent on task. In this model of user behavior, inspection of snippets takes a certain length of time, and viewing of the underlying document adds a further variable time, depending on the length of the document, and whether it has been viewed previously. User willingness to continue reading down the ranking is assumed to erode as a function of time, rather than of snippets viewed, a further point of difference. Smucker and Clarke analyze the behavior of a set of 48 users, each spending up to ten minutes undertaking each of four search tasks. From this data they infer values for a number of critical parameters that drive their model, including probabilities that a document will be viewed, given that it is relevant (according to external relevance judgments), and the probability that it will be saved (regarded as being relevant by the subject), given that it has been viewed.

Common to many of these metrics is that the effectiveness value can be interpreted as being the rate at which relevance is accrued in terms of documents inspected (or, in the case of TBG, time spent) by a probabilistic user; and hence the only difference between these metrics is the estimate of how the user behaves. All of these models also share another feature – they are plausible only if users do indeed read "from the top"; or, at least, read SERPs in such a way that it can be logically equated in some way to a "from the top" reading order. A range of user studies have suggested that this is indeed the case, albeit with some variation [1, 4, 5, 7]; and a key purpose of our study was to further explore that assumption.

3 User Experiment

To investigate user behavior, an instrumented web browser was used to access a commercial search service via its API. A total of $n = 34$ subjects were each asked to carry out six search tasks, after exploring the system via a training task. The six search tasks were of three different difficulty levels, following the categorization given by Wu et al. [16]. Table 1 lists half of the tasks, one of type *remember*, one of type *understand*, and one from the hardest category, *analyze*.

The browsing interface used in the experiments allowed users to click on documents and read them, but not open tabs or further windows. A pre-determined "starter query" was the first one run for each topic; thereafter, subjects were free to run other queries while they explored that topic, and to move to further pages in the results listing for any query. If a document was selected and opened for reading, it could only be closed by the user selecting one of two buttons, indicating whether the viewed document was "useful" or "not useful" for answering the information need. Only when that assessment had been

Information specification	Starter query
(*remember*) You recently attended an outdoor music festival and heard a band called Wolf Parade. You really enjoyed the band and want to purchase their latest album. What is the name of their latest (full-length) album?	*wolf parade*
(*understand*) You recently became acquainted with one of the farmers at the local farmers' market. One day, over lunch, they were on a rant about how people are ruining the soil. They were clearly upset, so you're interested in finding out more. What are some human activities that degrade soil fertility?	*damage soil fertility*
(*analyze*) Your sister is turning 25 next month and wants to do something exciting for her birthday. She is considering some type of extreme sport. What are some different types of extreme sports in which amateurs can participate? What are the risks involved with each sport?	*extreme sport*

Table 1. Three of the test queries employed in the user study, together with the first query evaluated for each subject. The second and subsequent queries were at the discretion of the subjects.

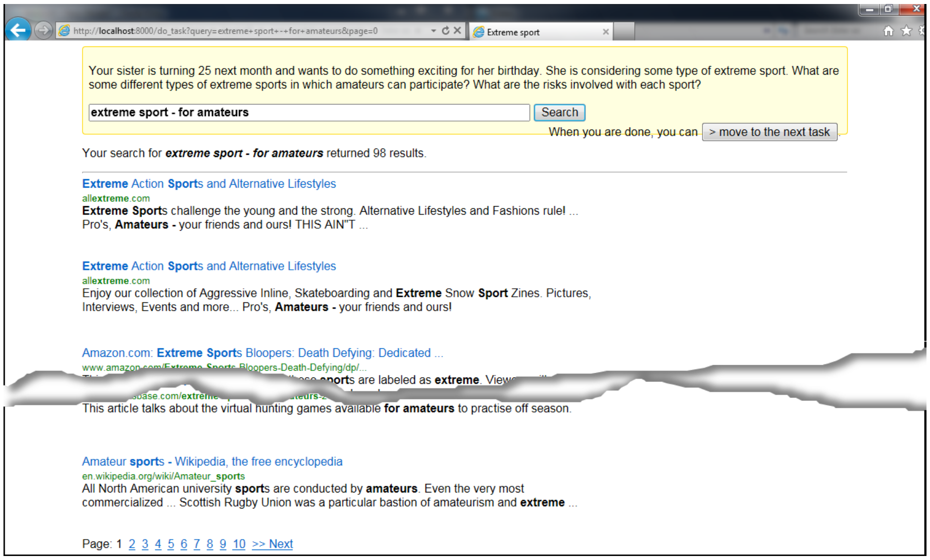
lodged did the original SERP become accessible again. We took these user-supplied judgments as being definitive of relevance for that user, and did not carry out any further judgments, working on the principle that the user's behavior is based on what they think at the time, rather than what an expert says via a post-hoc assessment. Users were asked to "collect a set of answer pages that in your opinion allow that information need to be appropriately met"; they were free to elect when they had reached that point, and move on to the next topic.

Search results were presented in SERPs that each contained ten links to documents, with seven of the links visible "above the fold" at the time the page was opened, and three more visible on scrolling. A set of "next page" links was provided at the bottom of each SERP, while a query box at the top of the SERP allowed fresh queries to be issued. The information need statement for the topic was displayed at the top of each answer page, as a reminder to the participant of what they were looking for. Figure 1 shows (part of) a typical SERP screen for one of the *analyze* tasks.

"Facelab" gaze-tracking equipment[1] monitored the user's gaze on the screen throughout each session. The stream of observations from the tracker was then reduced to *fixations* of at least 75 milliseconds duration within a 5-pixel radius; and sequences of fixations within the area of each displayed snippet were further amalgamated. That sequence was then integrated with the browser log data that noted user actions, including: queries and query reformulations; clicks and document opens; and document judgments made. The resulting processed data can be thought of as sequences of snippet numbers that follow the user's gaze, interspersed with notes about explicit actions undertaken, such as clicks, judgments, and query reformulations.

Topics were presented to participants in a structured manner so as to minimize any ordering effects. As an independent dimension of the exploration, we also systematically degraded the SERPs in half of the subject-topic combinations, using a technique described by Jones et al. [8]. In these *diluted* results, all of the odd-rank positions were replaced by documents that contained words matching the query, but were off-topic. In this paper we focus exclusively on the undiluted results; an analysis of the differences in user behavior arising from the dilution is part of a separate study [14]. Each of the 34

[1] http://www.seeingmachines.com/product/facelab/



http://localhost:8000/do_task?query=extreme+sport+-+for+amateurs&page=0
Extreme sport

Your sister is turning 25 next month and wants to do something exciting for her birthday. She is considering some type of extreme sport. What are some different types of extreme sports in which amateurs can participate? What are the risks involved with each sport?

extreme sport - for amateurs

When you are done, you can

Your search for ***extreme sport - for amateurs*** returned 98 results.

Extreme Action **Sport**s and Alternative Lifestyles
all**extreme**.com
Extreme Sports challenge the young and the strong. Alternative Lifestyles and Fashions rule! ...
Pro's, **Amateurs -** your friends and ours! THIS AIN"T ...

Extreme Action **Sport**s and Alternative Lifestyles
all**extreme**.com
Enjoy our collection of Aggressive Inline, Skateboarding and **Extreme** Snow **Sport** Zines. Pictures, Interviews, Events and more... Pro's, **Amateurs -** your friends and ours!

Amazon.com: **Extreme Sport**s Bloopers: Death Defying: Dedicated ...
www.amazon.com/Extreme-Sports-Bloopers-Death-Defying/dp/...
sports are labeled as **extreme**. View

base.com/**extreme**
This article talks about the virtual hunting games available **for amateurs** to practise off season.

Amateur **sport**s **-** Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Amateur_**sports**
All North American university **sport**s are conducted by **amateurs**. Even the very most commercialized ... Scottish Rugby Union was a particular bastion of amateurism and **extreme** ...

Page: 1 2 3 4 5 6 7 8 9 10 >> Next

Fig. 1. A typical SERP as viewed by the study participants. The first seven links are visible when the page is opened, the other three, and the "next pages" buttons, are visible after scrolling.

users can thus be thought of as contributing three topics to the pool of data, with one topic drawn from each of the three categories shown in Figure 1.

The majority of the study participants were undergraduate or graduate students in STEM topic areas, all fluent in English (though many as a second language), majority male, and for the most part aged in their twenties. A total of 37 participants were identified initially, but technical issues meant that the data from three people was not used in the analysis. Ethics approval for the experimental design was granted by RMIT University's Ethics Advisory Board.

4 Results

We now examine some of the data we collected during our experimentation.

Click-Throughs. The relative frequency of click-throughs is shown in Figure 2, plotted as a function of snippet rank in the results page. The expected downward trend is present [7], and serves as a useful confirmation of two effects: the search service is more likely to place promising items near the top of the ranking; and users are more likely to view items near the top of the SERP. In combination, these two factors mean that click-throughs are more top-biased than document viewings.

Click information, and in particular whether clicked-on documents were subsequently marked as being useful, can be used to investigate whether participants were taking their search tasks seriously. For each pair of participants, the mean percentage agreement (calculated as the number of documents that were given the same relevance rating by

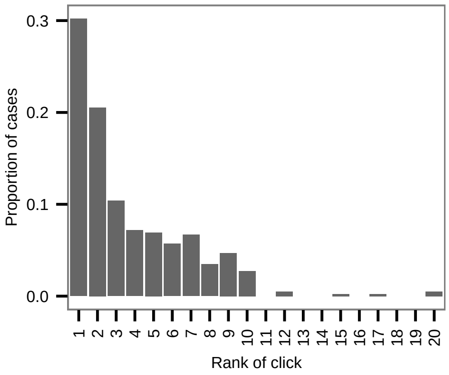


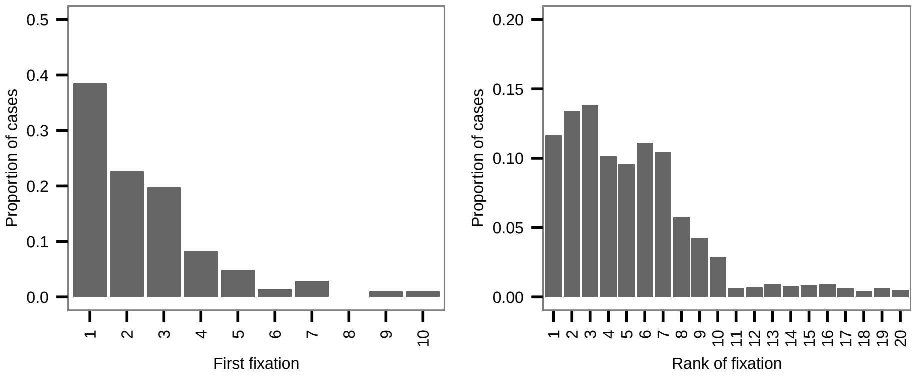
Fig. 2. Click-throughs averaged over topics and subjects, plotted by rank.

both participants, divided by the total number of documents that were viewed by both participants) was 79%. For the relevant-only class, the proportion of specific agreement is 85%. This is very high; for example, Voorhees [15] reports pairwise positive agreement of 42% to 49% between primary and secondary TREC assessors. However, it should be noted that the latter evaluation was carried out over a full set of TREC relevance pools, while our comparison is over the subset of documents that were clicked on by users, and hence likely to include a higher rate of relevant documents. In any case, the high level of agreement suggests that our user study participants were in fact attentive to their task.

Fixations. Figure 3 shows how the set of fixations was distributed over the rank positions in the SERP. In Figure 3(a), the distribution of first fixation points for the subject-topic-query combinations is plotted. The top snippet in the SERP dominates, and is the first one read 38% of the time. But ranks two and three also attract a significant fraction of the first fixations, and the participants were more likely to start with either the second or third snippet than they were to start with the first. Snippets that fall below the bottom of the screen (ranks eight and above) are also sometimes the first one viewed; indicating that in some rare cases the user's first action is to scroll the results window.

Across all fixations recorded during the experiments, shown in Figure 3(b), snippets that are closer to the top of the results page are more likely to be viewed than snippets lower down. But the relationship is not monotonic, and the first rank position is not the one that is most frequently viewed – positions 3 and then 2 enjoy that status. The role of the number of snippets in each SERP, and of the location of the "fold" – the point below which users needed to scroll down the page in order to reveal more snippets – is apparent in this second graph. (Where users requested a second page of results for the same query, the snippets were labeled as being at ranks 11–20, and so on.) The three snippets below the fold are rather less likely to be viewed; and snippets on the second results page are even less likely to be looked at.

Fixation Progressions. We define a *jump* as the difference between consecutive fixations for the same subject and topic. For example, if the t th snippet viewed in the SERP



(a) (b)

Fig. 3. Fixation distributions: (a) rank of first fixation for each submitted query; and (b) total fixations for each query. There were seven snippets presented "above the fold" on each SERP, and ten snippets in total on each SERP.

$<$	-4	-3	-2	-1	$+1$	$+2$	$+3$	$+4$	$>$
0.047	0.033	0.049	0.069	0.230	0.347	0.104	0.046	0.032	0.043
		0.427					0.573		

Table 2. Observed jump probabilities, expressed as fractions of a total of 2,633 overlapping two-fixation observations. A further 234 fixations occurred as singleton events, and were not included here. The median of this distribution is 1.0, the mean is 0.15.

is at rank d_t, then the t th jump is given by $j_t = d_{t+1} - d_t$. An ideal "from-the-top" reader would have $d_t = t$ and hence generate a jump sequence of $j_t = +1$ values; while at the other extreme, a genuinely random reader would make a selection from $j_t \in \{-d_t + 1 \cdots + \infty\} \setminus \{0\}$ (assuming that the SERP is, in effect, infinite in length).

Table 2 gives an overview of the jump distribution observed in our experiments. Negative movements are nearly as likely as positive ones, with 57% of the jumps positive and 43% of the jumps negative. The median jump value is $+1$, as expected; however the mean jump value is only $+0.1$. At face value, these observations suggest that the document reading order is neither "from the top" nor random.

In fact, there is a certain amount of embedded structure in the jump sequence, but at a higher level than is revealed by Table 2. Table 3 lists observed occurrences of forwards and backwards jumps, first as overall totals matching the second row of Table 2, and then broken down by three conditioning categories: those jumps that are the difference between the first two fixations (that is, the first jump in each SERP displayed); those that took place immediately following a prior backwards jump; and those that took place immediately following a forwards jump. This is, each column of the table represents an estimate of the relative probability of backwards and forwards jumps (as shown by the parenthesized values), given one unit of knowledge of the gaze sequence. As has already been noted, the overall count of backwards jumps is only modestly smaller than

	overall	first	after $j_{t-1} < 0$	after $j_{t-1} > 0$
jump $j_t < 0$	1125 (0.427)	85 (0.392)	296 (0.295)	744 (0.527)
jump $j_t > 0$	1508 (0.573)	132 (0.608)	709 (0.705)	667 (0.473)

Table 3. Conditional probabilities of forward and backward jumps. Values in parentheses are observed proportional split between positive and negative jumps, in four different contexts. A positive jump is much more likely after a negative jump than it is in the other three contexts.

$<$	-4	-3	-2	-1	0	$+1$	$+2$	$+3$	$+4$	$>$
0.055	0.045	0.043	0.089	0.070	0.242	0.094	0.161	0.092	0.046	0.060
		0.303						0.454		

Table 4. Observed two-jump probabilities, expressed as fractions of a total of 2,416 overlapping three-fixation observations. The median of this distribution is 0.0, the mean is 0.33.

the overall count of forwards jumps. But when the estimates are conditioned by one prior event, a different picture emerges – after a backwards jump, it is very likely that the next transition will be forwards again; and after a forwards jump, there is heightened likelihood of a backwards jump.

Table 4 sums adjacent pairs of jumps to get a net change over sequences of three consecutive fixations. For example, the gaze sequence “$1,3,2,3,4,3,5$” would reduce to the 1-jump sequence “$+2,-1,+1,+1,-1,+2$” and then be further reduced to the 2-jump sequence “$+1,0,+2,0,+1$”. As the table shows, when adjacent pairs of jumps are combined, the dominant outcome is “0” – around a quarter of the time the user will be looking at the same document again two steps from now. Table 5 provides further details. The most common 2-jump is “$+1,+1$”; with the “$-1,+1$” and “$+1,-1$” combinations also relatively common. The only double-negative combination in the top 12 is “$-1,-1$”; after that, the next double negative combinations are “$-2,-1$” at rank 16 (0.012), and then “$-1,-2$” and “$-1,-3$” at equal rank 22, with probabilities below 1%.

More importantly, the direction and magnitude of the first jump in each pair influences the second. Figure 4 explores this connection. Here, the horizontal axis gives the first jump, the vertical gives the second, and the level of shading gives the probability of the second jump conditioned on the first (so each column “adds up to 1”). Regardless of what jump has just happened, a jump of $+1$ (that is, reading down the results list) is very common, although this effect is weaker following a large positive (downward) jump since there are fewer results left. Other patterns are also evident. A jump in one direction ($+$ or $-$, down or up) is commonly followed by a jump in the other. In particular, jumps are commonly in the opposite direction and are of about the same magnitude, an effect that gives rise to the shaded band around the diagonal. This explains the tendency to an overall outcome of “0”, in Table 4.

The relative abundance of these effects – jumps of $+1$, and jumps in one direction being followed by equal jumps in the other – might describe a user who is consciously or unconsciously looking at a result, often going back to some sort of “best so far” to compare it, then going forward a little and repeating the sequence.

Comb.	prop.	Comb.	prop.	Comb.	prop.
$+1,+1$	0.128	$+1,+2$	0.037	$+1,-2$	0.024
$-1,+1$	0.098	$+2,-1$	0.033	$-2,+1$	0.023
$+1,-1$	0.096	$-1,+2$	0.029	$+1,-3$	0.017
$-1,-1$	0.044	$+2,+1$	0.027	$+3,+1$	0.014

Table 5. The most frequent two-jump combinations, expressed as a proportion of the 2,416 overlapping three-fixation observations.

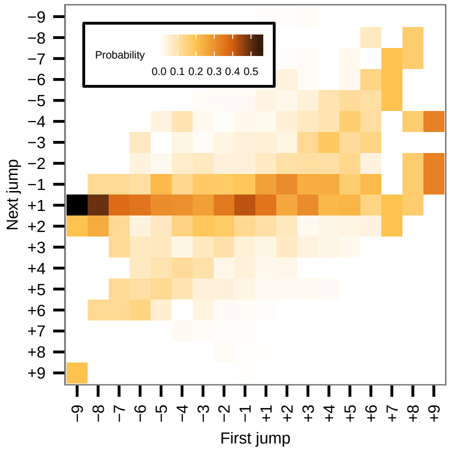


Fig. 4. Two-jump combinations. The horizontal axis gives the first jump in each pair; the vertical the second; and the shading gives the probability of this second jump conditioned on the first.

Taken together, these statistics suggest that there are a wide variety of reading behaviors, and the assumption that the average user reads a search results list from the top until they stop is somewhat simplistic, even if that is what click-through patterns might suggest. Instead, it appears that a modified sequential reading process takes place: searchers maintain a "zone of interest" that is a small number of snippets (two or three) wide, and read backwards and forwards freely within that zone, maintaining a localized set of potentially interesting snippets that are evaluated against each other before a click-through takes place. It is the zone that is likely to start near the top of the page, and then steadily progress downwards, rather than the fixations themselves.

It is also worth noting that some of the effect that has been observed may be due to the inherent imprecision of the gaze-tracking hardware and software (our tracker is generally accurate to within 10 pixels), and it might be that a sequence "$-1,+1$" reflects the user's eyes drifting slightly offline while reading a single snippet, and that only a single fixation was involved.

Exit From a SERP. Figure 5 shows the distribution of the lowest-ranked snippets that were viewed, for each query. Distinct peaks can be observed at ranks 7 and 10. These peaks are a consequence of the screen layout within the browser: 7 snippets showed above

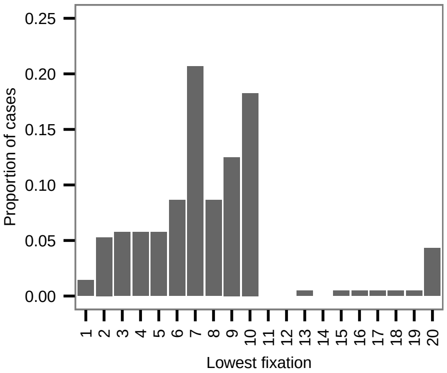


Fig. 5. Distribution of the ranks of the lowest snippet viewed for each query.

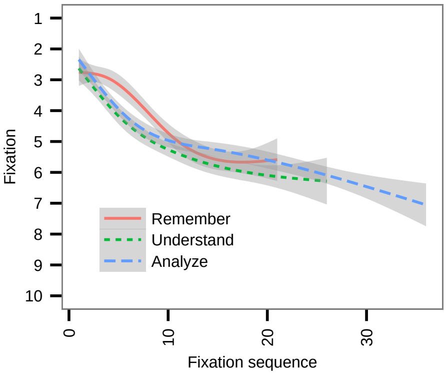


Fig. 6. Fixation rank as a function of fixation sequence. The three task types, in increasing order of complexity, were: *remember*, *understand* and *analyze*.

the fold, and 10 were presented on each query page. The smaller peak at 20 similarly marks the end of the second page of results. Putting all modeling aside, there is clearly a strong influence on user behavior caused by presentation geometry. It seems questionable whether the default “ten results per page” is an optimal setting, since some users won’t scroll at all. On the other hand, those that make the additional cognitive commitment to do so are rewarded with only three additional answers, before encountering an even more challenging hurdle in the form of a “next page” link. Investigating the influence of screen geometry, and the relative impact of the two barriers (needing to scroll, and needing to click on a link), is an interesting area for future work.

Impact of Task Type. Recall that participants in our user study carried out search tasks of three complexity levels: *remember*, *understand* and *analyze*. Figure 6 plots the mean fixation rank as a function of fixation sequence (fitted with a polynomial). For the simplest task category, *remember*, views were unlikely to move below the fold. On the other hand, for the more complex *understand* and *analyze* tasks, views were likely to continue to the bottom of the first results page. The different slopes also suggest that reading speed tended to be higher for more complex tasks. When users need to assemble a larger number of answer documents, they may be more inclined to scan the entire results list first, to get a feel for the range of answer documents that are available to them. For simpler tasks, a more common strategy seems to be to inspect the top rank positions more carefully, until one or two satisfactory items are found and the task is completed.

5 Conclusions and Future Work

Information retrieval systems provide searchers with multiple results in response to a query, typically formatted as a ranked list of document summaries. This paper investigated the long-standing assumption that users read through such a list from top to bottom, one item at a time.

Analysis of eye-tracking and click data from a study of 34 searchers showed that there are large variations in viewing behavior. While rank one was the most common single place to start looking, in over 60% of cases participants began their exploration of a results page from a different position; overall, the most frequently viewed positions were at ranks two and three. Examination of sequences of gaze movements showed that most users in fact shifted their attention freely within a zone of interest, typically consisting of two to three snippets. On average, this zone tended to start near the top of a results page, and shift slowly downwards. This detail is not apparent from click behavior alone, which suggests that click behavior is not a good proxy for viewing behavior and may not be a good proxy for users' decision processes or effort.

The majority of information retrieval evaluation metrics are based on either positional (static) or cascade (adaptive) models of user behavior, both of which assume linear, top-to-bottom reading patterns [3]. Given the findings above, it appears that these models are not capturing complexities that are present in searcher behavior. Investigating how such patterns can be incorporated into refined models, and how this might impact on evaluation metrics, is an interesting avenue for future work.

Our analysis also showed differences in gaze behavior for tasks of different complexity levels: for simple *remember* tasks, searchers tended to constrain their attention to the top half of the results list, with their zone of interest flattening out at around rank position five, presumably after they have found a sufficient number of relevant documents to satisfy their information need. For more complex tasks, users worked their way down the results list more quickly, and also to a greater depth, on average. This effect of task type on search behavior needs to be better understood. For example, different tasks might be approached with different expectations about the number of documents that need to be found; this might partially explain why users tended to read faster for the more complex tasks. We have investigated some of these relationships in other work based on the same user study [10]; and also explored the effect that answer quality has on user behavior [14].

A related issue is the potential impact that the instructions given to user study participants might have on their search behavior. Such effects have been observed when administering questionnaires, for example [9]. While prior work has demonstrated that framing artificial information needs in a task-based scenario can increase the fidelity of searcher behavior [2], the impact on gaze behavior when reading a search results screen is an open research question.

Acknowledgments We thank Dingyun Zhu for his help running the experiments. This work was supported by the Australian Research Council.

References

1. Aula, A., Majaranta, P., Raiha, K.J.: Eye-tracking reveals the personal styles for search result evaluation. In: Proc. INTERACT. pp. 1058–1061. Rome, Italy (2005), LNCS 3585
2. Borlund, P.: Experimental components for the evaluation of interactive information retrieval systems. J. Documentation 56(1), 71–90 (2000)
3. Chapelle, O., Zhang, Y.: A dynamic Bayesian network click model for web search ranking. In: Proc. WWW. pp. 1–10. Madrid, Spain (2009)
4. Dumais, S., Buscher, G., Cutrell, E.: Individual differences in gaze patterns for web search. In: Proc. IIiX. pp. 185–194. London, England (2010)
5. Granka, L.A., Joachims, T., Gay, G.: Eye-tracking analysis of user behavior in WWW search. In: Proc. SIGIR. pp. 478–479. Sheffield, England (2004)
6. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Information Systems 20(4), 422–446 (2002)
7. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Proc. SIGIR. pp. 154–161. Salvador, Brazil (2005)
8. Jones, T., Hawking, D., Thomas, P., Sankaranarayana, R.: Relative effect of spam and irrelevant documents on user interaction with search engines. In: Proc. CIKM. pp. 2113–2116. Glasgow, Scotland (2011)
9. Kelly, D., Harper, D.J., Landau, B.: Questionnaire mode effects in interactive information retrieval experiments. Information Processing & Management 44(1), 122–141 (2008)
10. Moffat, A., Thomas, P., Scholer, F.: Users versus models: What observation tells us about effectiveness metrics. In: Proc. CIKM. San Francisco, California (2013), to appear
11. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Information Systems 27(1), 2:1–2:27 (2008)
12. Robertson, S.: A new interpretation of average precision. In: Proc. SIGIR. pp. 689–690. Singapore (2008)
13. Smucker, M.D., Clarke, C.L.A.: Time-based calibration of effectiveness measures. In: Proc. SIGIR. pp. 95–104. Portland, Oregon (2012)
14. Thomas, P., Scholer, F., Moffat, A.: Fading away: Dilution and user behaviour. In: Proc. 3rd Europ. Wrkshp. HCI and IR. pp. 3–6. Dublin, Ireland (2013)
15. Voorhees, E.M.: Variations in relevance judgements and the measurement of retrieval effectiveness. Information Processing & Management 36(5), 697–716 (2000)
16. Wu, W.C., Kelly, D., Edwards, A., Arguello, J.: Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In: Proc. IIiX. pp. 254–257. Nijmegen, The Netherlands (2012)
17. Yilmaz, E., Shokouhi, M., Craswell, N., Robertson, S.: Expected browsing utility for web search evaluation. In: Proc. CIKM. pp. 1561–1564. Toronto, Canada (2010)