ViewSer: Enabling Large-Scale Remote User Studies of Web Search Examination and Interaction

Dmitry Lagun
Mathematics & Computer Science Dept.
Emory University
dlagun@emory.edu

Eugene Agichtein
Mathematics & Computer Science Dept.
Emory University
eugene@mathcs.emory.edu

ABSTRACT

Web search behaviour studies, including eye-tracking studies of search result examination, have resulted in numerous insights to improve search result quality and presentation. Yet, eye tracking studies have been restricted in scale, due to the expense and the effort required. Furthermore, as the reach of the Web expands, it becomes increasingly important to understand how searchers around the world see and interact with the search results. To address both challenges, we introduce ViewSer, a novel methodology for performing web search examination studies remotely, at scale, and without requiring eye-tracking equipment. ViewSer operates by automatically modifying the appearance of a search engine result page, to clearly show one search result at a time as if through a "viewport", while partially blurring the rest and allowing the participant to move the viewport naturally with a computer mouse or trackpad. Remarkably, the resulting result viewing and clickthrough patterns agree closely with unrestricted viewing of results, as measured by eye-tracking equipment, validated by a study with over 100 participants. We also explore applications of ViewSer to practical search tasks, such as analyzing the search result summary (snippet) attractiveness, result re-ranking, and evaluating snippet quality. These experiments could have only be done previously by tracking the eye movements for a small number of subjects in the lab. In contrast, our study was performed with over 100 participants, allowing us to reproduce and extend previous findings, establishing ViewSer as a valuable tool for large-scale search behavior experiments.

Categories and Subject Descriptors

H.4 [**Informational storage and retrieval**]: evaluation, search process.

General Terms

Design, Experimentation, Human Factors.

Keywords

Web search behavior; web search evaluation, remote user studies.

SIGIR'11, July 24–28, 2011, Beijing, China.

1. INTRODUCTION

Web search engines serve billions of searches a day, providing information for a diverse range of information needs. Understanding and analyzing how users interact with search has emerged as an important area of research. In particular, eye tracking has proven to be an invaluable technology for studying search behavior, providing important insights into search interface design. Yet, despite these advantages, eye tracking studies remain relatively small scale, as they require in-lab participation and supervision, and thus are "too expensive" for day-to-day search evaluation.

This paper proposes a new methodology for *performing large-scale behavioral studies of web search*, while maintaining many of the benefits of the controlled in-lab eye tracking studies of search. For this, we present a specially designed search engine result interface, which we call ViewSer (for Viewport Examination of Web Search Results). ViewSer aims to induce result examination behavior similar to unrestricted viewing, yet allowing us to track precisely the viewed portion of the search result page. For this, ViewSer blurs most of the search result page, except for the search result currently examined (pointed to) by the cursor, which creates a clear "viewport", illustrated in Figure 1. This viewport follows the cursor position, allowing a subject to examine the search results, while the viewport position is tracked.

The kinds of web search evaluation for which ViewSer is designed, focus on evaluating search results individually. Examples of such tasks are: measuring the rates of result examination and estimating snippet attractiveness – valuable for accurate clickthrough interpretation [24] and for learning to rank from click data; and evaluating snippets (result abstracts), e.g., by using the proportion of views to clicks on a result [25], as we demonstrate in this paper. Indeed, in this paper we explore multiple practical applications of ViewSer. As a first task, we show that ViewSer can serve as an effective method to measure and estimate snippet attractiveness - indicating that a snippet tends to "attract" clicks. This in turn can help better interpret clickthrough data for tasks such as learning to rank. Another crucial task is generating and evaluating search result snippets. In this paper we explore an application of ViewSer to detect bad (misleading) snippets which can serve as a valuable feedback to snippet generation algorithms. The contributions of this paper include:

- A novel web search evaluation methodology that could enable large-scale studies of search examination behavior (Section 3).

- A validation of our ViewSer prototype, showing that the search result viewing exhibited by remote participants using our ViewSer interface closely approximates unconstrained search examination (Section 4).

- A demonstration of effectively applying ViewSer to crucial web search tasks of learning to estimate snippet attractiveness, which in turn can be used for result ranking, and for automatically detecting bad (misleading) snippets (Section 5).

This work is just a first step in developing and applying the ideas introduced in this paper. As we continue to refine the ViewSer technology, other applications and evaluation methods will be explored. In summary, we believe that the ViewSer technology will form a crucial component in the future of web search evaluation.

2. RELATED WORK

Our goal is to enable large-scale, remote behavioral studies of web search, focusing on search result examination, normally done with eye-tracking equipment. Thus, our work spans the areas of user studies in IR, search behavior modelling, crowd-sourced IR evaluation, and eye tracking-based evaluation of search.

Crowdsourcing methodology has emerged as a viable way to cheaply obtain human input for a wide range of tasks, including document relevance assessments. One of the most popular web sites providing a marketplace for hiring internet workers is Amazon Mechanical Turk (AMT). Previous efforts studied various aspects of document relevance rating crowd-sourced via AMT, including task completion time, worker's responsiveness, locality and ratings quality in terms of accuracy and inter-rater agreement [2, 21, 6, 20]. Kely et. al. reports an important study of searcher behavior for in-lab and remote participants [19]. In contrast to these studies, our focus is searcher behavior - specifically, search result examination. Somewhat related to our approach, [11] describes crowdsourcing user studies of graphical perception conducted via AMT. However, we are not aware of any published user study of web search examination behavior conducted for hundreds of users in crowdsourcing framework.

Our work is inspired by the emergence of the large-scale, passive logging and analysis of search behavior as an alternative to in-lab studies: the log data has been used for search evaluation [2], for improving search engine ranking [1, 15] among other tasks. However, such log-based studies are a blunt instrument - they are more appropriate for overall search performance evaluation, whereas our proposed methodology enables precise tracking and characterization of searcher behavior, at the level of detail previously only possible with eye tracking studies of search.

To enable this vision, our implementation of ViewSer builds upon the previous work on restricted focus viewing (RFV) described in references [4, 13, 3], where the authors explored the effect of restricted viewing in usability studies of user interfaces. However, our work substantially differs from prior applications of this idea, as our work is, as far as we know, the first to apply this idea to web search. Our approach is also more general, scalable, and efficient compared to previous work: our implementation is based on the Scalable Vector Graphics (SVG) technology natively supported by the Firefox browser, which in turn enables ViewSer to render and blur rich XHTML content such as text formatted with cascade style-sheets, images and videos, while [4] describes an application to image examination.

Figure 1: An example of the ViewSer interface displaying a blurred search engine result page (SERP) for the query "toilet", with the viewport revealing the first result.

As we will show, ViewSer has many potential applications to web search. In particular, estimating the "attractiveness" (with respect to clickthrough) of search result summaries, or snippets can improve click interpretation, which is in turn helpful for more accurate ranking models. Previously, Clarke et al. [7] found statistically significant changes in clickthrough patterns due to caption features. The more recent work of [25] confirmed some of the findings of [7] in a different experiment using the concept of "fair pairs". Both of these approaches used methodology based on changes in clickthrough, whereas our work directly measures the searcher examination of the captions, and subsequent behavior. Previous work on snippet evaluation and generation [17] followed a more classical approach, based on explicit labels for snippet quality or language models.

As another application, snippet attractiveness could be useful to improve learning to rank (LTR) from click data, by adjusting the clickthrough counts according to snippet attractiveness. LTR has demonstrated promising results in ranking large scale document collections based on both relevance judgements as well as implicit user feedback, relevant algorithms include [9, 16]. In fact, result attractiveness was already incorporated in click models, though was modelled as an unobservable (hidden) variable [5]. In references [26], snippet presentations features were used as part of user features to better model result examination. Our work is also related to [1], where the authors used features such as query terms, and the browsing and clickthrough statistics, to improve result ranking. In summary, previous models have studied result attractiveness based on changes in click-

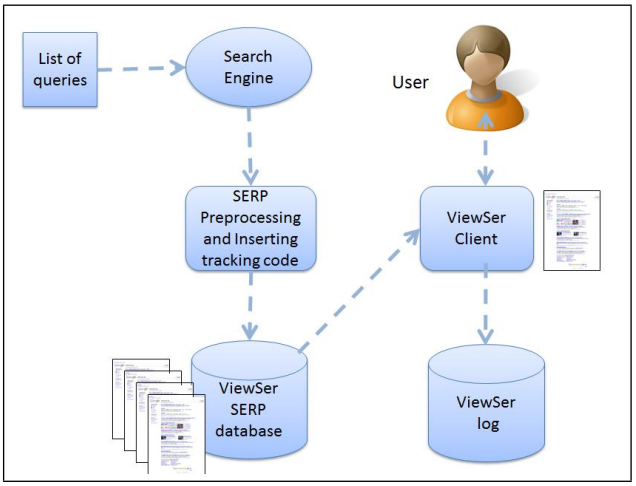


Figure 2: The ViewSer architecture for large-scale search result examination studies.

through patterns or statistical inference based on user clicks. The work described in this paper differs from prior work in that we present and validate a novel methodology for directly measuring result examination. In turn, this enables training models to reproduce and extend previous findings quickly, effectively, and at scale.

3. VIEWSER SYSTEM IMPLEMENTATION

The ViewSer architecture is outlined in Figure 2. First, ViewSer retrieves and pre-processes the search engine result pages (SERPs), by inserting code into each SERP to modify the appearance and to enable tracking of the user interaction events. These SERP pages, and the landing pages of the results (the actual documents) are cached in a database for the subsequent studies, enabling fully reproducible and repeatable experiments. A participant opens one of the pre-processed SERPs, which causes her browser to blur all but one results on the SERP. As the participant moves the viewport around to view the rest of the results, the precise position of the viewport and other searcher interactions are sent to the server and logged for future analysis. The rest of this section describes these steps in more detail.

SERP pre-processing: to emulate the "viewport" of the ViewSer interface, we automatically modified the SERPs by inserting the JavaScript/Scalable Vector Graphics (SVG) code, which is directly supported by the Firefox browser, without requiring any additional plugins or other downloads. The code leaves clear one result region at a time (identified using the HTML DOM tags), and blurs the rest of the SERP. Specifically, the SVG specification is used to describe the blurring effect and incorporate it into regular cascade style sheet class, which can be added to any HTML DOM element in a web page.

More concretely, the fragment of the code in Figure 3, *make-blur*, defines a Gaussian filter for blurring a search result, and can be referenced in a style specification of any HTML element to blur the element's content accordingly. This specific filter is a Gaussian filter with the σ parameter set to 2.5. The second operation performed by make-blur is gray-scaling the element appearance. Conveniently, each search result on a SERP is described within list element

```
<svg:svg>
<svg:filter id="make-blur">
<svg:feGaussianBlur stdDeviation="2.5"/>
<svg:feColorMatrix values="
0.3333  0.3333  0.3333  0  0
0.3333  0.3333  0.3333  0  0
0.3333  0.3333  0.3333  0  0
0       0       0       1  0"/>
```

Figure 3: A fragment of the Support Vector Graphics (SVG) code used by ViewSer to blur HTML elements such as search results.

<LI>. Therefore we modified the style of <LI> elements on the SERP, thus blurring all of the results.

ViewSer front-end : Initially, all out-of-focus elements are blurred and discolored to grayscale in order to imitate peripheral vision. Then, when a viewport moves over an element (e.g., a search result), the element's style can be changed back to the original appearance by detecting the onMouseOut event. The SVG-based implementation makes ViewSer scalable for crowdsourcing, as it does not require any additional installation, and responsive by exploiting the optimized native browser support, while allowing precise tracking of the viewing of any HTML element (such as the result position on SERP). These are significant advantages over previously proposed implementations using browser plugins or Java applications [4, 13]. A limitation of ViewSer is that only the complete HTML DOM elements can be revealed, not allowing for partial or gradual occlusion.

Logging the Searcher behavior: To track the viewport movement and other interface events, we injected additional JavaScript code into the SERP shown on the client's machine. This code logged window events such as clicking, scrolling, mouse moves, and finally the hovering events over a search result lasting 200ms or longer (corresponding to the typical duration of a eye movement fixations [23]). These events were buffered and periodically sent to the server via asynchronous HTTP requests for subsequent processing. The ViewSer code and tracking infrastructure is available on the project web-page [1].

4. VALIDATING VIEWSER

To validate the ViewSer technology we performed two main user studies: first, to collect "ground truth" eye tracking data; and second, to collect examination data using our ViewSer interface to compare to the eye-tracking behavior.

4.1 Search Tasks and Study Procedure

We used 25 benchmark search tasks from the WEB Track of the TREC 2009 competition. The goal for each task (the task description) was provided to the participants. For example, the goal of the query "toilet" was stated as: "Find information on buying, installing, and repairing toilets". For each task, the query keywords were submitted to the Google search engine, and the Search Engine Result Pages (SERPs), as well as all the result documents linked from each SERP, were cached. The original SERP layout was not modified (as shown on Figure 1), recreating a realistic search experience for the participants. The participants started with a

[1] http://ir.mathcs.emory.edu/intent/

Query	*Description*
`toilet`	Find information on buying, installing, and repairing toilets.
`mitchell college`	Find information about Mitchell College in New London, CT, such as a prospective student might find useful.
`cheap internet`	I'm looking for cheap (i.e. low-cost) internet service.
`espn sports`	I'm looking for various sports scores and information from the ESPN Sports site.
`euclid`	Find information on the Greek mathematician Euclid.

Table 1: Example queries and descriptions provided to the subjects.

provided SERP for each query, and were instructed to find the needed information with least effort that is, to click only on results that appear relevant. After a subject clicked on a result to examine the document and went back to the SERP, she was asked to rate the document relevance. To be considered a valid response, we required that participants attempt all search tasks, and click and rate at least one result for each task.

4.2 Participants

Eye-tracking group: for this "ground truth" group, ten participants (6 female, 4 male, ages 23.0±1.5, all graduate and undergraduate students and fluent English speakers) were recruited. The eye tracking was performed using a Tobii x60 eye tracker paired with a 17" LCD monitor set to 1280x1024 resolution. The subject's gaze position was sampled at 60 Hz with accuracy of 0.5 degrees. For the two remote studies, participants were recruited through the Amazon Mechanical Turk website, using the standard mechanism of listing our study as an available Human Intelligence Task (HIT).

ViewSer group: The workers were required to use the popular Firefox web browser. They were instructed to view the search results using the ViewSer interface as described above. 203 MTurk workers attempted the remote study. As a first step, the data obtained from MTurk subjects were automatically filtered to discard careless or automated (robot) workers. While the instructions required providing relevance judgements, some workers did not provide relevance judgements, and/or spent less than 1 minute on the whole HIT of 25 queries (presumably, to obtain the payment and move on). After these cases were automatically filtered out, we had valid data from 106 workers (48%).

Unconstrained viewing group: to serve as "control" subjects, we recruited additional 25 MTurk workers. The task was identical to the ViewSer group, except that we removed blurring, allowing for unconstrained viewing of the SERP.

4.3 Results

Our goal was to investigate whether ViewSer indeed induces similar viewing and clickthrough behavior remotely in MTurk subjects, as in the unconstrained viewing setting for the in-lab eye tracking subjects and remote participants. Before presenting quantitative results and analysis, Figure 4 shows an example heatmap of the relative time spent viewing the SERP for the query "toilet", aggregated for all subjects in Eye-tracking group. The first vertical colorbar (a) projects the relative viewing time onto the vertical axis of the SERP for the Eye-tracking group, and the second colorbar (b) for the ViewSer group, showing a noticeable similarity between the most intensely scrutinized search results.

Overall, the ViewSer group required 1 minute and 37 seconds on average (SD=70 seconds) for each search task, compared to 55 seconds on average (SD=20 seconds) for the Eye-tracking group. While the subjects in the ViewSer group took more time for each task, this is to be expected due to more time required to move a mouse pointer. Interestingly, the resulting search behavior patterns of the two groups are remarkably similar otherwise.

ViewSer SERP Examination and Clickthrough: Figure 5 reports the viewing and clickthrough rates for each result rank for the ViewSer group. Each data point indicates the fraction of the result views at each rank position, and the corresponding fraction of the clicks landing on that position, for all searches and participants. The first 3 results were viewed for 93%, 87% , and 78% of the tune, respectively, dropping to 27% for the 10th result. The clickthrough values are correlated with the viewing, with the exception of the results in the last (10th) position, which is slightly more likely to be clicked than the 9th result. These viewing and clickthrough patterns correspond well to the previous studies of unconstrained search result examination behavior [22, 12].

Comparative Analysis of ViewSer vs. Eye-Tracking: Figures 6(a) and 6(b) report the relative viewing times and clickthrough rates for the Eye-Tracking and ViewSer groups, respectively. The values in Figure 6(a) were computed for each subject and query (that is, the viewing time for a particular abstract by a subject was divided by the total viewing time of the corresponding SERP) for an individual query, and then averaged across all queries. The relative viewing time is important, as the longer a searcher's gaze stays on a particular area, more information is processed, and therefore this area receives more attention. Thus, comparing the relative amount of time (attention) spent on examining results through the ViewSer interface, vs. that of the Eye-Tracking group, quantifies the similarity of the viewing behavior of the two groups. Interestingly, ViewSer participants spent more time viewing first results is probably because of higher speed of eye movements compared to hand (mouse) movement, which may increase the likelihood of skipping over the results when viewing the page unconstrained (eye tracking group) compared to requiring to move the mouse to reveal the next result (ViewSer group).

Figure 6(b) reports the relative clickthrough rates for eye tracking and ViewSer participant groups. Each data point corresponds to the percent of all clicks for a query, landing on the corresponding result rank; these values are then averaged across all queries. In other words, the reported clickthrough rates are normalized for each query separately, and then averaged across all queries. We found that ViewSer group exhibits lower clickthrough rates on top results. We hypothesize that this is likely due to ViewSer interface encouraging more careful examination of the results in the top

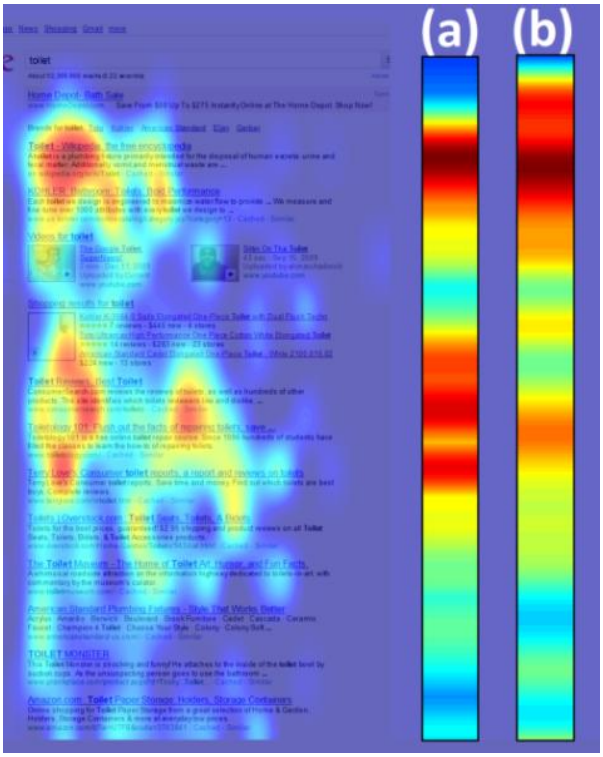


Figure 4: An example attention heatmap showing the relative viewing time over a SERP for the query „toilet" (Eye-tracking group), and the corresponding colorbar, showing the heatmap density projected onto the vertical axis (a). Overlaid as (b) is the colorbar for the viewing time for the same SERP but for the ViewSer group. This figure illustrates the similar distribution of attention between eye-tracking (a) and ViewSer (b).

positions, resulting in lower rates of "indiscriminate" clicking frequently observed for top-ranked results [1].

Comparative Analysis of Viewing and Clickthrough for Individual Queries: More detailed analysis of the Spearman correlation of viewing and clickthrough behavior for the ViewSer and Eye-Tracking groups for *individual queries* is reported in Figures 7(a) and 7(b), respectively. For the vast majority of queries (over 80%), the correlation of the viewing and clickthrough behavior of the ViewSer and Eye-Tracking groups is over 0.8 and is never below 0.6, indicating that ViewSer provides a close approximation of eye tracking for over 80% of queries, and a moderate approximation for the remainder. To gain additional intuition of the relationship between Eye-tracking and ViewSer behavior for individual queries, we plot the relative viewing time measured using Eye tracking (Y axis) and ViewSer (X axis) for each result for all queries (Figure 8). The color shading indicates the results rank position, where the red color corresponds to rank 1, and the blue color to rank 10. The result viewing times, as measured by the two methods, correlate strongly (r = 0.74). Intuitively, results with higher ranks cluster in the top right quarter as both groups spend more time viewing higher-ranked results, as expected.

Further Analysis: ViewSer vs. Unconstrained Browsing: to validate ViewSer methodology further, namely, to determine whether ViewSer participants examine the SERP differently due to restricting of their peripheral vision, we performed a follow-up study with additional 25 MTurk workers. This final group enjoyed Unconstrained viewing of the SERP, without blurring of out-of-focus results. The clickthrough rates of this group are reported in Figure 9, together with the corresponding ViewSer clickthrough rates. Remarkably, the clickthrough behavior of these groups is similar, with Spearman correlation r = 0.81.

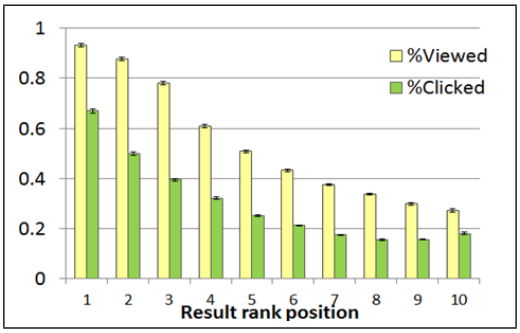


Figure 5: Viewing and clickthrough rates for each rank, aggregated for all queries and participants (ViewSer group).

5. APPLICATIONS TO SEARCH TASKS

This section describes three practical applications of the ViewSer technology to web search. First, Section 5.1 describes collection of relevance rating used in our experiments for this section. Then, we describe how ViewSer could be used to analyze snippet attractiveness (Section 5.2), to improve result ranking (Section 5.3), and to detect misleading snippets (Section 5.4).

5.1 Relevance ratings collection

To validate our findings on a bigger dataset, we collected SERP examination data for an additional 50 queries taken from the HARD track of TREC 2005, resulting in a dataset of 75 queries. Separately from the ViewSer study, we collected comprehensive relevance judgments for all of the results on the first page of results, for all queries in the WEB Track and the HARD Track. The Mechanical Turk workers (MTurk) were recruited to perform the relevance labeling as described above.

To control worker accuracy in ViewSer group we obtained relevance ratings for documents of each query in our collection. Each MTurk HIT was to assess organic (non-sponsored) results for one query. Following the recommendation of [20], the authors labeled 10% of documents as a validation set in order to estimate the worker accuracy and verify the quality of their work. On average, results for each query were rated by 6 workers. Inter-rater agreement, computed with Fleiss Kappa was 0.39, which correspond to fair/moderate agreement. We conjecture that this level of agreement is caused by the difficulty of the tasks and the informational nature of the queries. These ratings were used to compute the worker's accuracy on validation set and to filter workers with low accuracy as unreliable. For the WEB track, 17 of

Table 2: A sample of the snippet features used for estimating snippet attractiveness.

Feature	*Description*
Title (21 features)	
titleQueryTerms	Number of bolded (matched) query terms appeared in title
titleStartsWithQuery	Title starts with phrase match to query
titleQueryMatch	Number of query terms matched by title
titleFirstCapitalLetter	Number of terms with first letter capitalized
titleCapitalizedTerms	Number of capitalized terms in title
titleLengthToQueryLength	Ratio of title length in to query length words
titleHasURL	Equals 1 if title contains URL
titleNonPunctuation	Number of punctuation marks in title
Summary (17 features)	
sumaryPunctuation	Number of punctuation marks in summary
homeInSummary	Feature equals to one if summary contains word home
mapInSummary	Feature equals to one if summary contains map
summarySentanceFragments	Number of sentence fragments delimited by triple dot in the text summary
summaryLinks	Number of links in summary
URL (10 features)	
urlQuery	URL has form www.query.com where the query matches exactly with spaces removed
urlQueryTerms	Number of times query terms matched URL
urlCharacterLength	Length of URL in characters
urlSlashes	Number of slashes (i.e. subdomains) in URL

106 workers were filtered out, and for the HARD track 101 out of 263.

5.2 Application 1: Snippet Attractiveness

In this section we present one possible usage of data collected with ViewSer to analyze snippet attractiveness. The importance of snippet attractiveness as a contributing factor of clickthrough patterns has been explored by a number of researchers [7, 8, 25]. Our work has the advantage that we can directly measure the ratio of the times a snippet was examined to the number of times it was clicked, which call the *COV* ratio. In other words, *COV* is defined as probability of clicking on result given that result was examined. We hypothesize that the *COV* ratio is, in fact, a measure of snippet attractiveness that is independent of the rank position.

Experimental Setup: As a first step, we validate our hypothesis that *COV* is not dependent on the rank position, and in fact can be used as an un-biased estimate of snippet attractiveness. To this end, we calculate Pearson correlation coefficient between the result rank position and number of times the result was examined, clicked, and ratio of these counts. Here we report our comparative analysis of the (*COV*) metric based on data from Eye-tracking and ViewSer groups. We also report Pearson correlation between the *COV* ratio and textual features of the snippets.

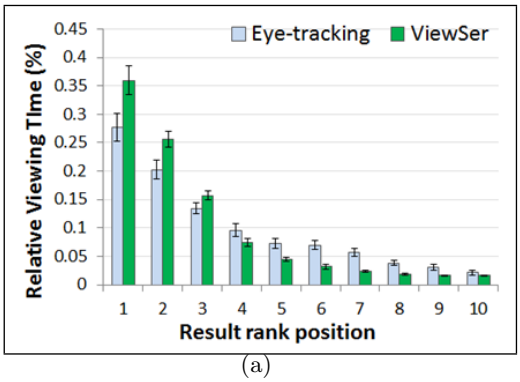
As [7, 25] indicate, there exists a strong position bias in the way results are viewed on the SERP - searchers browse the list of results from the top of the page to the bottom, which confirms previous findings [22], and more recently [12]. Such bias puts major obstacles of using click or result viewing time [12] feedback directly in search engine optimization. Different ways of eliminating of the presentation bias in clicks have been studied in [7, 25] as we highlight earlier in the text. In other words, application of additional techniques is required in order to extract useful signals from click data. It is reasonable to expect similar problems with viewing time measured using eye tracking or approximated with mouse hovering [10, 12]. However, our *COV* metric does not correlate with the result rank: the correlation between result rank and COV is 0.05 for the Eye-tracking group and 0.11 for the ViewSer group. This is a remarkable result, indicating that the *COV* ratio does not appear to be affected by result position bias.

Table 3: Snippet feature importance ranked by correlation to attractiveness.

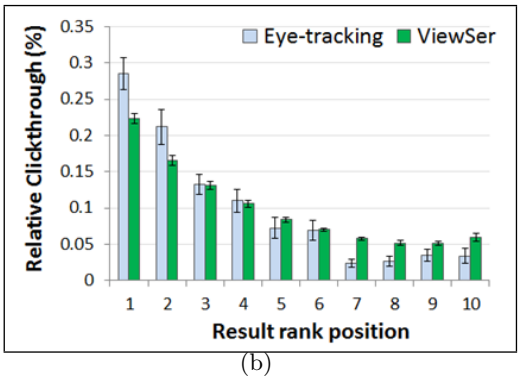
Correlation	*Feature*
0.2009	*titleStartsWithQuery*
0.1195	*summarySentanceFragments*
0.1169	*urlQuery*
0.1118	*titleQueryMatch*
0.1080	*sumaryPunctuation*
0.1060	*homeInSummary*
0.1030	*mapInSummary*

We validated the *COV* ratio measured with ViewSer on our eye tracking data. Figure 10 shows the *COV* ratio broken down by result position. On average we have observed slightly higher *COV* values in ViewSer data in comparison to Eye-tracking. Overall, Pearson correlation coefficient between Eye-tracking and ViewSer groups computed for each individual result was 0.64, which indicates substantial correlation.

In order to understand how *COV* relates to the previous work on estimation of result attractiveness [7, 25], we analyzed the correlation between *COV* and the textual features of the snippets. Table 2 shows example features that we considered. While many of the features used have already been investigated in prior work, we have extended the feature list with features capturing the rich structure of the snippets. For example, the feature *summaryLinks* indicates whether a snippet has additional embedded hyperlinks to

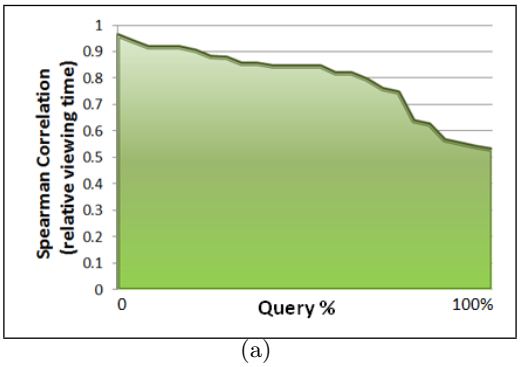


(a)

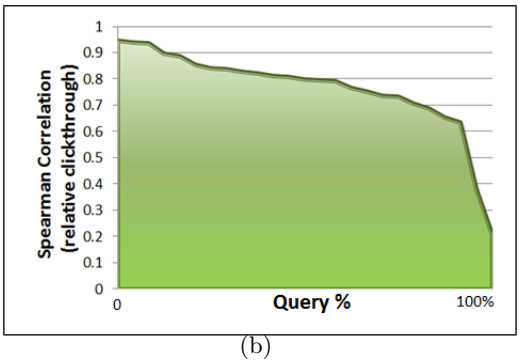


(b)

Figure 6: Viewing time (a) and clickthrough rate (b) comparison for the ViewSer and the Eye-Tracking groups, aggregated across all queries and subjects.



(a)



(b)

Figure 7: Spearman correlation of the relative viewing time (a) and clickthrough rates (b) for individual queries, for the Eye-tracking and ViewSer groups. The queries are sorted by the correlation coefficient. (mean viewing correlation: 0.79, mean clickthrough correlation: 0.76).

within-site navigation. Another example feature capturing complex snippets is *mapInSummary*, indicating whether a result contains a map with local search results. These features might be useful in predicting result attractiveness, as they can change searcher's SERP examination behavior.

The Table 3 reports the correlation between the *COV* ratio and text features of the snippet. The feature *titleStartWithQuery* has the highest correlation of 0.2. Features *titleStartWithQuery, urlQuery, titleQueryMatch,* and *homeInSummary* have higher correlation with *COV* than other features, which confirms previous findings in references [7, 25].

As we show in the next subsection, using attractiveness as an additional feature can be helpful for important and practical web search tasks.

5.3 Application 2: Result re-ranking

Learning to rank has become a very popular approach to achieve better search results ranking. In this section we investigate whether attractiveness can be used as a feature in learning to rank framework to improve original ranking. Unfortunately, the *COV* statistic as an additional feature measured directly, based on user study, would not be practical for large scale LTR experiments, since it would require collecting viewing data for each individual result. Therefore, we build a regression model to predict result attractiveness based on the textual snippet features. For this purpose we built a regression model predicting *COV* ratio from textual features described in Section 5.2. Thus, we used two additional features for re-ranking: ***COV*** (Click over Views ratio) and ***A***(estimated attractiveness) i.e., the estimated value of *COV* based on textual features.

Experimental Setup: We used the same query and document dataset as described in Section 5.1, providing labeled relevance data for 75 queries and 650 documents. We used *SVM-rank* [16] as the LTR method of choice. To estimate result attractiveness, we used the Gaussian Process regression model with radial basis function kernel. The correlation of the estimated and true COV values was 0.6 (3 fold cross-validation). Improving the estimation of snippet attractiveness will be part of our future work.

Results: Table 4 reports the NDCG [14] averaged across 3 folds for the baseline ranking system (Google) as well as for the re-ranking method, using directly measured *COV* and the the estimated attractiveness (A). The average NDCG of the original ranking was 0.8408. Our first experiment was to train a ranker based on the document position feature and the *COV* ratio, which yielded a significant improvement of 6% over the baseline. This substantial ranking improvement was surprising, given that the search engine was already highly optimized. Following the recommendation of [18], we performed significance test between the original ranking and our system, showing significance at $p < 0.05$. Once attractiveness (A) model was trained on the training data (2 folds) we use it to compute the attractiveness feature for the test fold. As reported in Table 4, the re-ranking performance based on the estimated value of snippet attractive-

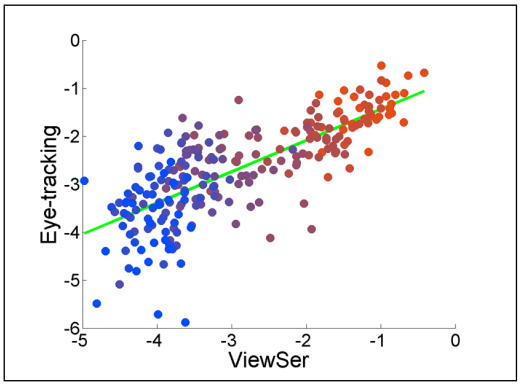


Figure 8: Relative snippet viewing time for top 10 organic results, for Eye-tracking vs. ViewSer (both on logarithmic scale). The color indicates the rank position of the results i.e., the red color corresponds to rank 1, and the blue color corresponds to rank 10. (Pearson $r = 0.74$.)

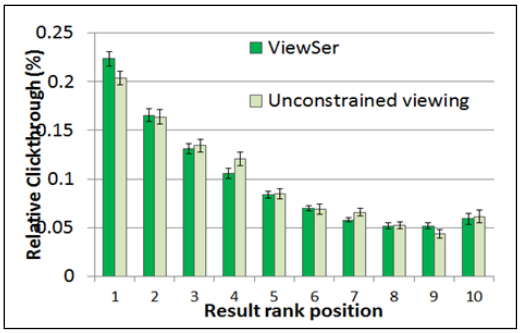


Figure 9: Clickthrough rates for ViewSer and Unconstrained viewing groups (Spearman rank correlation $r = 0.81$).

ness, outperformed the Google baseline ranking by 5.25%, reaching NDCG of 0.8848 and being statistically significant with p<0.05. The slight gap between P + *COV* and P + *A* results is due to the expected regression error, that can be further reduced with more training data or richer features. We also tried to add number of clicks received by document as an additional feature to the ranker, but the performance was slightly lower. Nevertheless, the demonstrated improvements are remarkable, considering the relatively small amount of training data that was required to estimate the snippet attractiveness and in turn improve the ranking over a state-of-the-art Google ranking.

5.4 Application 3: Detecting Bad Snippets

In this section we describe our experiments on detecting bad (i.e., misleading) search snippets. Intuitively, good snippets should clearly summarize the result document so that searcher would be able to understand whether it is worth clicking or not. Specifically, we consider good snippets to be those, that attract clicks on relevant documents, or discourage clicks on non-relevant documents. Conversely, bad snippets would discourage clicks on relevant documents, while attracting clicks on non-relevant documents. More formally, we define snippets to be *Bad* or *Good* based on the snippet

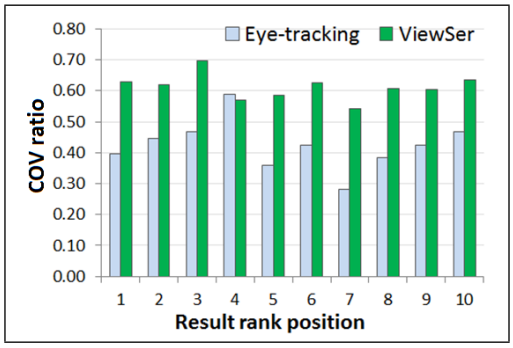


Figure 10: Clicks over Views (*COV*) for Eye-tracking and ViewSer groups, by rank position.

Table 4: Ranking system comparison. P - rank position, *COV* - Clicks over Views, C - clicks, *A* - estimated attractiveness. * indicates significance at $p < 0.05$, ** indicates significance at $p < 0.01$.

Features	NDCG
Baseline (P)	0.8408(−)
P + *COV* (ceiling)	0.8920(+6.09%)**
P +*A*	0.8848(+5.24%)$*$
P + *A* + Clicks	0.8840(+5.14%)*

COV ratio (defined in Section 5.2), and the result relevance (manually labeled as described in Section 5.1):

$$Label(COV, REL) = \begin{cases} Bad & \text{if } (REL \geq 0 \text{ and } COV < \theta_2) \\ & \text{OR } (REL = 0 \text{ and } COV > \theta_1) \\ Good & \text{otherwise} \end{cases}$$

Where the parameters θ_1 and θ_2 are set empirically by manually examining a sample of the snippets and the documents. Thus, *Good* snippets for relevant documents would have higher (i.e., greater than θ_1) *COV* ratio, since a searcher would be more willing to visit the document after examining the snippet. Similarly, a *Good* snippet allows a searcher to identify a non-relevant document, resulting in lower *COV* ratio (i.e., less than θ_2). In contrast, a snippet is considered to be *Bad* if it fails to inform the searcher about the document relevance. Hence, a snippet for a relevant document that exhibits a low *COV* ratio (i.e., less than θ_2) is considered to be *Bad*. We experimented with different values of the θ_1 and θ_2 parameters and determined that the setting θ_1=0.85 and θ_2 =0.35 provides the closest match to our definition, on a subset of the data. With this parameter setting, our dataset contained 589 *Good* snippets and 61 *Bad* snippets.

Figure 12 shows an example of a snippet that appeared in the results to the query "`wildlife extinction`" where the information need was described as *"The spotted owl episode in America highlighted U.S. efforts to prevent the extinction of wildlife species. What is not well known is the effort of other countries to prevent the demise of species native to their countries. What other countries have begun efforts to prevent such declines"*. The snippet summarizes a news article talking about recreating aurochs from reconstructed DNA, which is an attempt to save the specie, so the document was judged as a relevant to the query. However, the text summary of the snippet talks about a seemingly un-

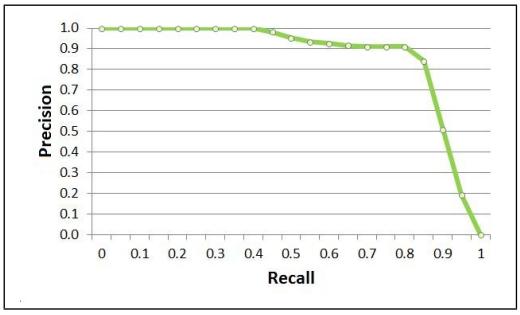


Figure 11: Precision vs. Recall for detecting bad snippets using ViewSer data.

Giant cattle to be bred back from **extinction** - Telegraph
Aurochs were immortalised in prehistoric cave paintings and admired for ...
www.telegraph.co.uk/.../**wildlife**/.../Giant-cattle-to-be-bred-back-from-**extinction**.html - Similar - Add to iGoogle

Figure 12: An example of a bad snippet: document is relevant while the COV ratio is low (0.18). This snippet was returned for the query "`wildlife extinction`" at the 4th position. The snippet summarizes a news article talking about recreating aurochs from reconstructed DNA, which is an attempt to save the specie, so the document is relevant to the query. However, text summary of the snippet is not representative, causing searchers to skip the document.

related fact that *"aurochs were immortalized in prehistoric cave paintings"*, which caused many of the searchers to skip this document. Another example of a bad snippet is shown in Figure 13, where a snippet appears relevant, but the actual document is not.

Cargo **Ship Losses** and Fatalities
Dec 11, 2003 ... safety regulations tankers, bulk carriers and passenger ships Royal Institution of Naval Architects Safety Committee general cargo ships Non ...
www.marinetalk.com/.../Cargo-**Ship-Losses**-and-Fatalities-ROY00691454IN.html - Cached - Similar

Figure 13: An example of a bad snippet: the document is not relevant while the COV ratio is high (0.85). This snippet was returned for the query "`ship losses`" at the 3rd position. The stated information need was: *"Identify instances in which weather was a main or contributing factor in the loss of a ship at sea"*. The snippet's summary lists relevant keywords, but the actual document does not discuss factors contributing to ship losses.

Experimental Setup: We treat this as a classification problem, where we attempt to predict the snippet label based on the textual features of the snippet, and our estimate of the snippet attractiveness, A, defined in Section 5.2. Specifically, we use the features listed in Table 2, as well as the A feature, as input to classification. We experimented with different classifiers such as Naive Bayes, Logistic Regression, SVM, and others, using 5-fold cross validation.

Results and Discussion: Interestingly, the highest accuracy was achieved by the LogitBoost [9] classifier, resulting in 97.7% accuracy (P = 97.6%, R = 97.7%, F1 = 97.6%, AUC = 93.6%). This improvement is significant over the majority baseline classifier, which had Accuracy of 90.6% (P = 82.1%, R = 90.6%, F1 = 86.2%, AUC = 43.9%). The Precision-Recall curve computed for the LogitBoost classifier is reported in Figure 11, showing that more than 35% of *Bad* snippets can be detected with 100% precision.

As a potential confounding factor, we did not consider whether a snippet contains an answer to the query directly on the SERP, removing the need to click on a document even when it is relevant. While this scenario could potentially violate our definition of a *Good* snippet, for the experiments in this paper, this case is extremely unlikely: the search tasks, especially those from the HARD set, are relatively complex, and are not likely to be answered directly in the snippet. As another research direction, we believe that using only the shallow text features for snippet quality classification leaves significant room for improvement, for example, by incorporating the readability and language model features proposed in [17]. We plan to explore these questions further in our future work.

6. CONCLUSIONS

We presented ViewSer, a novel methodology for crowdsourcing web search behavior evaluation studies. By restricting result viewing to a viewport, ViewSer enables tracking precisely which results are examined on a search engine result page, while preserving the natural result examination and interaction patterns. We validated the ViewSer prototype in a study with over 100 remote participants, recruited through the Amazon MTurk services, showing ViewSer induces similar search behavior compared to the in-lab users monitored using eye tracking. These results indicate that ViewSer could enable large-scale web search behavioral evaluation studies, for the fraction of the cost and effort required to perform these studies in the lab. We have presented three practical applications of ViewSer, namely a new way to learn to estimate snippet attractiveness, which could in turn improve ranking, and an exploratory application to automatically detecting bad (misleading) snippets.

As part of our future work, we plan on improving user click models using examination information collected with ViewSer, as it allows direct observation of result viewing previously only inferred by click models such as [5, 26]. In summary, we believe that ViewSer provides a crucial component that would enable realistic, large-scale, and reproducible behavioral evaluation of web search, which would in turn improve search result presentation (e.g., result summary generation), result ranking, and ultimately the overall web search experience.

ACKNOWLEDGEMENTS: This work has been supported by the NSF grant IIS-1018321 and by the Yahoo! Faculty Research Engagement Program.

7. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving Web Search Ranking by Incorporating User Behavior Information. In *Proceedings of the annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–26. SIGIR, 2006.

[2] O. Alonso, D. Rose, and B. Stewart. Crowdsourcing for Relevance Evaluation. In *ACM SIGIR Forum*, volume 42, pages 9–15. SIGIR, 2008.

[3] R. Bednarik and M. Tukiainen. Validating the Restricted Focus Viewer: A Study Using Eye-movement Tracking. *Behavior Research Methods*, 39(2):274–282, 2007.

[4] A. Blackwell, A. Jansen, and K. Marriott. Restricted Focus Viewer: A Tool for Tracking Visual Attention. *Theory and Application of Diagrams*, pages 575–588, 2000.

[5] O. Chapelle and Y. Zhang. A Dynamic Bayesian Network Click Model for Web Search Ranking. In *Proceedings of the International Conference on World Wide Web*, pages 1–10. WWW, 2009.

[6] E. Chi, P. Pirolli, and S. Lam. Aspects of Augmented Social Cognition: Social Information Foraging and Social Search. In *Proceedings of the International Conference on Online Communities and Social Computing*, pages 60–69. Springer-Verlag, 2007.

[7] C. Clarke, E. Agichtein, S. Dumais, and R. White. The Influence of Caption Features on Clickthrough Patterns in Web Search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 135–142. SIGIR, 2007.

[8] E. Cutrell and Z. Guan. What are you looking for? An Eye-tracking Study of Information Usage in Web Search. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 407–416. SIGCHI, 2007.

[9] J. Friedman, T. Hastie, and R. Tibshirani. Special invited paper. Additive Logistic Regression: A Statistical View of Boosting. *The Annals of Statistics*, 28(2):337–374, 2000.

[10] Q. Guo and E. Agichtein. Towards Predicting Web Searcher Gaze Position from Mouse Movements. In *Proceeding of the annual ACM SIGCHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3601–3606. SIGCHI, 2010.

[11] J. Heer and M. Bostock. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings of the ACM SIGCHI International Conference on Human Factors in Computing Systems*, pages 203–212. SIGCHI, 2010.

[12] J. Huang, R. W. White, and S. Dumais. No clicks, No problem: Using Cursor Movements to Understand and Improve Search. In *Proceeding of the annual ACM SIGCHI Conference on Human Factors in Computing Systems*. SIGCHI, 2011.

[13] A. Jansen, A. Blackwell, and K. Marriott. A Tool for Tracking Visual Attention: The Restricted Focus Viewer. *Behavior Research Methods*, 35(1):57–69, 2003.

[14] K. Jarvelin and J. Kekalainen. Cumulated Gain-based Evaluation of Information Retrieval Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[15] T. Joachims. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142. SIGKDD, 2002.

[16] T. Joachims. Training Linear SVMs in Linear Time. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226. SIGKDD, 2006.

[17] T. Kanungo and D. Orr. Predicting the Readability of Short Web Summaries. In *Proceedings of the Second ACM WSDM International Conference on Web Search and Data Mining*, pages 202–211. WSDM, 2009.

[18] G. Kazai, N. Milic-Frayling, and J. Costello. Towards Methods for the Collective Gathering and Quality Control of Relevance Assessments. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 452–459. SIGIR, 2009.

[19] D. Kelly and K. Gyllstrom. An Examination of Two Delivery Modes for Interactive Search System Experiments: Remote and Laboratory. In *Proceeding of the annual ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1531–1540. SIGCHI, 2011.

[20] A. Kittur, E. Chi, and B. Suh. Crowdsourcing User Studies with Mechanical Turk. In *Proceeding of the annual ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456. CHI, 2008.

[21] J. Le, A. Edmonds, V. Hester, and L. Biewald. Ensuring Quality in Crowdsourced Search Relevance Evaluation. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 17–20. SIGIR, 2010.

[22] L. Lorigo, M. Haridasan, H. Brynjarsdóttir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini, and B. Pan. Eye Tracking and Online Search: Lessons Learned and Challenges Ahead. *Journal of the American Society for Information Science and Technology*, 59(7):1041–1052, 2008.

[23] K. Rayner. Eye Movements in Reading and Information Processing: 20 years of research. *Psychological Bulletin*, 124(3):372, 1998.

[24] K. Wang, N. Gloy, and X. Li. Inferring Search Behaviors Using Partially Observable Markov (POM) Model. In *Proceedings of the ACM WSDM International Conference on Web Search and Data Mining*, pages 211–220. WSDM, 2010.

[25] Y. Yue, R. Patel, and H. Roehrig. Beyond Position Bias: Examining Result Attractiveness as a Source of Presentation Bias in Clickthrough Data. In *Proceedings of the International Conference on World Wide Web*, pages 1011–1018. WWW, 2010.

[26] Z. Zhu, W. Chen, T. Minka, C. Zhu, and Z. Chen. A Novel Click Model and its Applications to Online Advertising. In *Proceedings of the ACM WSDM International Conference on Web Search and Data Mining*, pages 321–330. WSDM, 2010.