Evaluating Search Systems Using Result Page Context

Peter Bailey, Nick Craswell, Ryen W. White
Liwei Chen, Ashwin Satyanarayana, and S. M. M. Tahaghoghi
{pbailey, nickcr, ryenw, liweich, assatya, stahagh}@microsoft.com
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052 USA

ABSTRACT

We introduce a method for evaluating the relevance of all visible components of a Web search results page, in the context of that results page. Contrary to Cranfield-style evaluation methods, our approach recognizes that a user's initial search interaction is with the result page produced by a search system, not the landing pages linked from it. Our key contribution is that the method allows us to investigate aspects of component relevance that are difficult or impossible to judge in isolation. Such contextual aspects include component-level information redundancy and cross-component coherence. We report on how the method complements traditional document relevance measurement and its support for comparative relevance assessment across multiple search engines. We also study possible issues with applying the method, including brand presentation effects, inter-judge agreement, and comparisons with document-based relevance judgments. Our findings show this is a useful method for evaluating the dominant user experience in interacting with search systems.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Measurement, Design, Experimentation

Keywords

Web search relevance, measurement, evaluation.

1. INTRODUCTION

Information retrieval (IR) systems are most frequently evaluated using Cranfield-style methodologies [12], involving standardized test collections, query sets, relevance judgments, and established measures of retrieval performance such as mean average precision (MAP) and normalized discounted cumulative gain (NDCG) [27]. Such approaches judge the relevance of the individual documents from a ranked list returned for a query, and use these judgments to compute a single score for the search system that is then averaged across many queries. This method of assessing a search system's result relevance manifests a high level of abstraction over the search task by eliminating several sources of variability [42], supporting many important experimentation, measurement and evaluation efforts. However, given the pivotal role of context in the search process [24], there is a need to study complementary evaluation methods that incorporate aspects of the search context into relevance judging.

When searching, users interact first with a search engine results page (SERP) and then with the retrieved documents. Each document has a summary, which may overstate or understate its relevance, and may include multimedia or links to additional documents. Around and interspersed in the ranked list there are other page elements designed to support the user's search task such as suggested spelling corrections, suggestions of follow-on queries, results from alternate queries, advertising links, and mini-rankings from other sources such as news, image and video search. These components comprise the *result page context*, of which search results are one aspect. Since this context may affect user perceptions of the search system it is important to consider it during search system evaluation. *Whole-page relevance* [4] defines how well the surface-level representation of all elements on a SERP and the corresponding holistic attributes of the presentation respond to users' information needs. Traditional IR evaluation methodologies (e.g., [12][43]) ignore page elements other than the main document ranking, the surface-level representation of these, and holistic results-page characteristics such as coherence, diversity and redundancy. Other techniques such as user studies (e.g., [25]) and A/B testing (e.g., [29]) are valuable but limited in terms of factors such as their scalability (user studies are costly and time-consuming) or their ability to capture qualitative feedback (A/B tests capture behaviors not rationales).

We present a method that considers all result page components when evaluating search system performance. Search engines have somewhat different SERP layouts; these are query dependent and evolve over time subject to the outcomes of experimentation on new components. An individual searcher may consider only some parts of the SERP. However, for a query that is repeated by many searchers, most, if not all, components will be viewed by at least one searcher. An approach is therefore needed to assess the collective relevance of all SERP components and consider changes to them over time. To do this we issue a query to multiple engines, and regard the SERP responses as though they were 'school assignments' from multiple students. While each student may have different styles and layout, overall they can be graded with respect to how effectively they address and satisfy the information needs represented in the assignment. Assignments can be graded both on component elements of their response (e.g., did they mention an important fact?) and on holistic aspects such as coherence, comprehension, and use of authoritative sources. This framework of analysis affords a direct analogy for for evaluating search systems using the whole-page relevance of SERPs, and gives rise to our method's name – the *School Assignment Satisfaction Index* (SASI).

IIiX 2010, August 18-21, 2010, New Brunswick, NJ, USA.

The remainder of this paper is structured as follows. We describe related work on relevance measurement in Section 2. Section 3 provides an overview of the SASI evaluation method. Contributions made possible by SASI are stated in Section 4. Section 5 describes experiments with the method and their findings. We discuss these findings and the SASI method more generally in Section 6 and conclude in Section 7.

2. RELATED WORK

Relevance measurement in Web search has an expansive history dating back over a decade. As such, this section can only briefly cover some of the significant related contributions.

The Text Retrieval Conference (TREC) series contained extensive investigations of Web search evaluation in the context of the Very Large Collection (1997-1998) and Web tracks (1999-2003) [20]. The tracks explored various search tasks, query sets and document-metrics-based evaluation methods over a set of Web document collections of different sizes. Hawking et al. characterized the challenges in applying classic ad hoc information retrieval tasks (i.e., informational search where the user is interested in potentially all relevant documents), query topics, and metrics to the Web environment [21]. They made numerous recommendations on how best to apply or adapt Cranfield-style methods to Web search evaluation. The TREC Interactive Track (1996-2002) investigated users' search processes in conjunction with experimental IR systems [17]. A significant challenge in the track was in understanding the comparability between different experiments. In 2002 [23], the track included user interaction over Web document collections using results provided by the Panoptic search engine developed by Hawking and colleagues. Various experiments were conducted, but small participant and topic samples made it difficult to draw strong conclusions.

Evaluation for ad hoc information retrieval developed through TREC had focused on mean average precision (MAP) as the dominant effectiveness measure. MAP contains both recall- and precision-oriented aspects and is sensitive to the entire search result ranking. For Web search environments, this was regarded as inappropriate since users appeared primarily interested in tasks involving high early precision, and were less concerned with recall [22]. Alternative document retrieval-based metrics have also been introduced, such as mean reciprocal rank, appropriate for evaluation in home page finding tasks [21] and (normalized) discounted cumulative gain (NDCG) [27], catering to a user model of scanning a ranked list of results to some depth.

The approach we describe in this paper evaluates search systems holistically based on complete SERPs and their components. Two important methods already exist for examining result page features. The first approach involves detailed Web user studies [25], where the experimenter monitors the search interactions of individual participants using a variety of techniques such as surveys, interviews, think-aloud protocols, gaze-tracking, and analysis of interaction behavior at the results page element level. Due to the significant human involvement in user studies, it is prohibitively expensive to obtain thousands of search interactions via these methods. However, such studies afford the collection of detailed qualitative feedback on the features presented to participants. The second approach involves online A/B testing experiments [29]. These involve setting up two parallel systems, one of which is the treatment system containing a new feature or feature variation, and one of which is the control system without the feature or containing the existing baseline experience for the feature. Large user populations are exposed at random to one of the systems and their interactions are recorded. Randomization will usually occur at the user level to ensure that each user has the same experience, either treatment or control, on each visit to the search engine. The differences between the interactions on systems are analyzed to determine the utility of the new treatment for users. Proxy signals, such as click-through rates or revenue per search, are used to determine success. This method does not provide qualitative feedback on the feature, but does provide quantitative feedback about user engagement at scale for use in understanding feature value.

A comparative variant of A/B experiments exists in the form of online website survey platforms, as exemplified by those from Keynote Systems (www.keynote.com). Keynote's survey platform provides automated analysis of multiple aspects of Website performance, including user experience. They also offer access to a large survey population which can be used to provide qualitative evaluation of task-based aspects of Website behavior, both individually and relative to other providers of similar services.

A number of researchers have assessed whether even quite substantial changes in document-based relevance metrics are truly predictive of user-perceived changes in satisfaction or effectiveness with search tasks [1][2][35]. One critique of a single document-based relevance metric (raised by Hawking et al. [21]) is that it does not characterize the multitude of user tasks and contexts in Web search. Thomas and Hawking developed a method for judging pairs of systems in situ with real users [39]. Two result lists (from two retrieval systems or ranking algorithms or caption presentation algorithms) are presented to users side-by-side (randomized in a left-right pattern by query), who interact with the lists in the usual manner. Participants may be asked to provide explicit preference ratings, or equivalent signals can be derived from interaction patterns. Carterette et al. investigated document-based preference judging systems and the efficient construction of metrics for such kinds of preference judging [8].

Preference-judging systems have typically unified and scrubbed the presentation aspects from two different systems. Presentation and other brand effects may change user perceptions of relevance. Contrary results have emerged in such studies. Bailey et al. found no preference for results purported to be from one search brand over another using the Thomas and Hawking side-by-side comparison technique with presentation aspects removed [5]. Jansen et al. conducted a user study where branding elements were removed from Google search results and Google, Yahoo!, MSN, and an in-house search engine branding was shown at the top and bottom of the results page [26]. They found various user behavior effects (e.g., the number of result list and sponsored links examined and clicked) according to the purported search brand, and determined qualitative user (and social) preference for dominant brands. Both studies show that results involving brand are highly sensitive to the experimental design.

Users' implicit ratings and preferences can be learned in more naturalistic settings by interpreting their click signals from interactions with SERPs. Joachims and Radlinski et al. have detailed various aspects of this approach in a series of papers. One line of investigation examined how to interpret search result clicks as relevance judgments [34]. Using such data for the problem of learning to rank was studied to correct for possible position bias in clickthrough rates using the FairPairs algorithm [36]. Another idea was to interleave results from two rankings for the same query into a single result set presented to the user [35]. Clicked

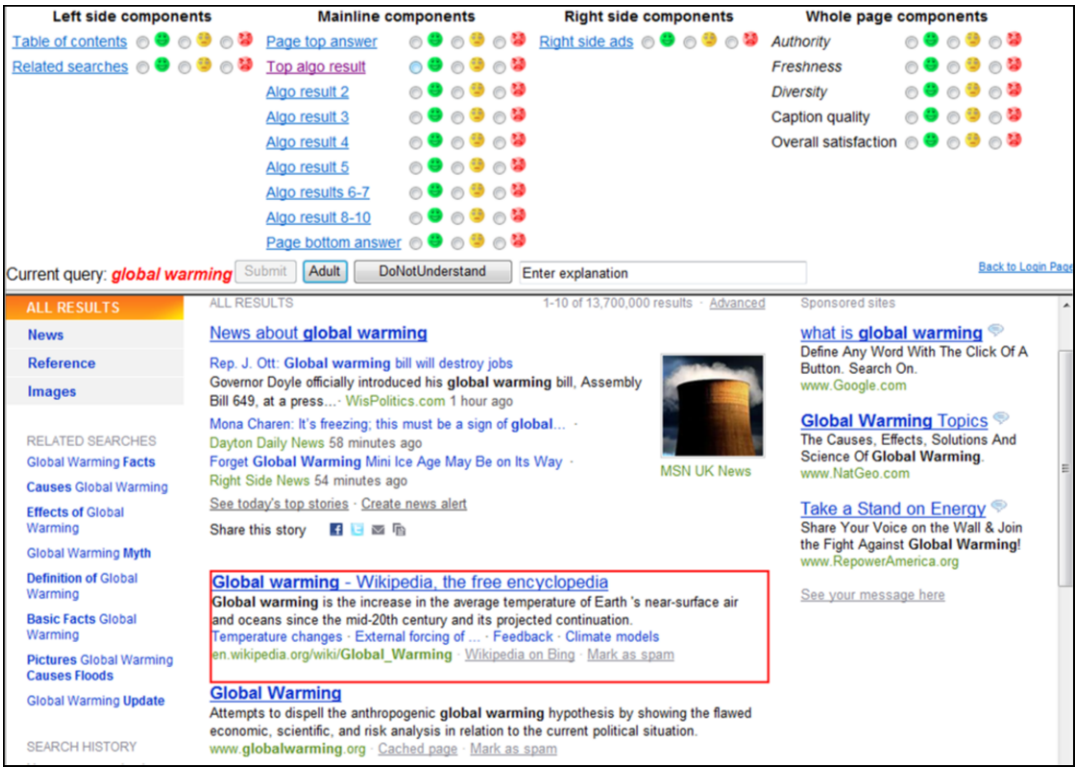

Figure 1. SASI judging interface example. Component currently being judged (*top algo result*) is highlighted.

search results can indicate preferences for one ranker over another, and clicks are sensitive to interdependence of documents in a ranked list. The overall effectiveness of interleaving has been characterized in respect to other evaluation methods [37].

In-depth analysis of user interaction with SERPs is also possible through eye-tracking [33]. Specialized equipment is used to monitor participants' eye saccades and fixations within the user interface and then analyze the areas which gather most (or least) attention and the flow patterns on the page. For example, Cutrell and Guan report on a study of user interaction with result captions [15]. They explored the effect of adding information to the caption, and discovered changes in performance for navigational tasks versus informational tasks. The same authors also investigated the effect of relevant document position for the same two task types, and found that users will prefer top-ranked results almost regardless of their relevance [19]. Buscher et al. explored the effect on overall relevance assessments by purposefully altering result page quality in a user study investigating good, bad and random advertisements for search queries [7].

The effect of presentation order in results (sometimes referred to as position bias) has also been investigated by Craswell et al. [14] and Wang et al. [44]. In both cases, they develop models to explain user behavior observed in large volumes of search engine result clickthrough logs and characterize likely document relevance to a query. Such methods are directly applicable to conducting online A/B tests, where user behavior signals must be analyzed to decide which of a pair of systems (the treatment and the control) is preferred by users. Such methods must be sensitive to many factors in real Web search. For example, not all clicks are indicative of user satisfaction; indeed, in some cases, the lack of clicks, so called "good abandonment," may be a positive signal [31]. Log-based investigations can also lead to insights regarding other SERP components beyond document relevance. For example, Clarke et al. examined features of result captions that led to changes in user click patterns [10].

The work closest in spirit to ours investigates user interactions with components of SERPs. Turpin et al. examined the impact of result caption relevance and how this affects perceptions on the relevance of (and click-through to) the underlying document for the query, using result captions generated by Google [40]. They also demonstrated that including caption evaluations as part of the relevance assessment process can change system rankings using MAP in the context of TREC Web collections. Kelly et al. investigated different methods of providing query and term suggestions to users in a qualitative user study [28]. They found that query suggestions are preferred to term suggestions, and that they can provide value to users even if they are not clicked. White et al. investigated the relative values of follow-on query suggestions and "popular destinations" (direct links to common URLs visited subsequently by others who had issued the same query and based on their trail behavior) [45].

The research of Kelly et al. and White et al. is among the few to investigate SERP components other than the results list and the corresponding result captions. The SASI evaluation method explicitly aims to support relevance examination of all search-relevant components on a SERP. We wished to support this type of analysis at a scale beyond that of a user study, but with explicit relevance labels (rather than implicit log signals). We also wished to enable judges to make assessments that considered positional effects and the relationship between components.

3. SASI METHODOLOGY

SERPs for English-language markets typically consist of a number of components, laid out in two or three columns. The main column is referred to as the *mainline* and usually contains the top ten document result summaries (titles, query-biased captions, and URLs). Other columns on SERPs are referred to as the *left rail* and *right rail* – relative to the mainline. Within each rail there are various blocks, for example the mainline top-10 block or the right-rail advertisement block. Within each block there may be multiple components. Each block supports search activities differently. For example, there may be advertisement blocks in the mainline and/or right rail; follow-on query suggestions at the top or bottom of the page, or in the left or right rail.

The SASI method is a framework for judging aspects of SERP component and holistic relevance. It differs from traditional Cranfield-style evaluations in three significant ways: (i) judging surface-level presentation on the SERP rather than documents, (ii) judging all components of the SERP, and (iii) judging the SERP components in context. Since SASI focuses on assessing the user's experience when interacting with the search engine, SERPs are captured and presented to judges in their entirety. However, there is no interaction with the interface components or any hyperlink clicking which may take the judge away from the SERP and introduce bias from landing pages encountered.

3.1 Experimental Procedure

The procedure for executing a SASI experiment is as follows (with differences from Cranfield-style Web evaluation in bold):

1. Specify the experiment:
   a. Identify a set of queries for evaluation
   b. Specify which types of SERP blocks and components should be judged
   c. Specify holistic aspects to be judged
2. **Capture SERPs from search engines and identify layout**
3. Allocate queries to judges
4. **Judging interface interactions:**
   a. Judge views query and SERP
   b. Judge labels blocks and components
   c. Judge scores holistic attributes
5. Analyze and aggregate judgments

3.2 Example SASI Judge Interface

Figure 1 shows our implementation of a SASI judging interface, with a judging pane and a pane showing a Microsoft Bing SERP for the query [global warming]. During judging, the interface can highlight the part of the SERP that is currently being judged, which in Figure 1 is *top algo result* (the first of the top-10 results selected by the search engine's ranking algorithm). Our approach of highlighting individual components and rating them is distinct from a Cranfield-style judgment system where documents are judged in isolation from their result page context. Note that this is just one example of a SASI judging interface, and others could be used. For example, although we adopted the emoticon-based judging scale proposed by Chandrasekar et al. [9], numeric or label scales could be used instead. Also, the choice of a three-point scale is arbitrary; a two-, four-, or five-point scale might be equally or more effective. There is scope for research on the effect of these and similar design decisions. We discuss this when reviewing the outcomes of some experiments performed with SASI that are described later in the paper.

3.3 Variation in SERPs and Highlighting

Each search engine has its own SERP layout, with optional blocks and components that vary on a per-query basis. Our implementation of SASI can identify what components are present on each SERP by parsing its HTML source. In the SERP pane of the judging interface this allows us to highlight the component under review. In the judging pane we list only the elements present in the current SERP that the experiment configuration specifies require judging. These facilities help us to conduct relevance investigations which focus on different aspects and sets of SERP components, as described below. Handling the significant SERP differences between engines and between queries is an engineering challenge rather than a research challenge.

Once SASI can identify the various components of a SERP, it may be configured to give more or less detail on each component. If detailed analysis of the mainline results were not required, SASI could be configured to replace Figure 1's *algo* (short for algorithmic top-10) judgments with a single *algo block* judgment. In a different experiment, SASI could be configured to target a particular user interface feature, such as the additional links placed in the top-ranked mainline result in Figure 1. SASI judgments could identify how often the links seem useful, and how often there is a problem. We added an *explanation* field to allow judges to comment on issues they may have with judging or the systems. Even without such a field we can get an overall false-positive rate on a particular interface feature by, say, measuring how well surface-level judgments based on SERP elements correspond to judgments on the full documents.

4. CONTRIBUTIONS

The SASI method is complementary to several existing relevance assessment methods used for Web search. In Table 1, we present a short summary of some of these methods, and their respective properties. Properties include: *Component* – which part of the user experience is assessed; *Users* – whether the method has real users creating the signal or label; *Contextual* – whether the method is executed within a whole-page context; *Comparative* – whether the users or judges make decisions between two or more alternatives; *Reusable* – whether the relevance labels can be re-used independently; *Holistic* – whether the technique supports identification of poor user experiences, such as missing an important result, redundant results, and result incoherence (see Section 4.1), and; *Intent/Task* – whether user intent/task goals are preserved during the evaluation process. If it is unclear if the method supports a property, it is indicated '?'.

Table 1. Comparison of relevance assessment methods.

Method	Reference	Component	Users	Contextual	Comparative	Reusable	Holistic	Intent/task
Online A/B	[29]	any	y	y				y
Interleaving	[34]	ranker	y	y	y			y
FairPairs	[36]	caption	y	y	y			y
Keynote	n/a	any	?	y	?		?	y
Side-by-side	[39]	ranker	?	y	y		y/n	
SASI	[4]	all		y			y	
Preference	[8]	any			y	y		
Captions	[40]	caption				y	y	
Documents	[12]	result				y		

The analysis in Table 1 illustrates that SASI is similar to many of the online-based measurement techniques such as interleaving or A/B testing in its contextually-sensitive properties. It also shares properties with offline techniques in its use of judges and the explicit relevance labels it creates. However, judging SERP components is substantially faster than judging documents. In our studies, we have observed judges on average complete ratings using SASI for *all* components on a SERP in the same time as they are able to judge just *two* documents. The significantly higher speed may be attributable to judging in context and may be valuable in obtaining large numbers of judgments in a short time period (improving scalability). Differences may also be attributable to judging different representations (i.e., summaries versus documents). In Section 5.2, we compare judgments assigned to SERP components with judgments of landing pages.

4.1 Judging in Context

Judges can see the whole-page of search results when judging using the SASI method. This makes it possible to identify SERP issues such as missing information, duplicated information and inter-component incoherence. The first two are related to diversity evaluation [11], but judged directly on the SERP rather than across the documents in the result set. An example of a coherence problem is showing the biography of the Greek epic poet Homer next to images of cartoon character Homer Simpson. Both may be relevant to the user's query [homer], but the user interface that groups these two is incoherent. Another problematic query is [cricket], which is a popular international game (Web documents), but also an insect (images), and a mobile phone provider in the United States (advertisements).

Showing the complete search results page also enables judges to quickly understand a query's general intent (from components such as document summaries and query suggestions). For example as shown in Figure 1, the query suggestions can indicate to the judge that users may be interested in causes, effects and facts about global warming; the presence of a news search result also suggests that this query has current topicality. The presence of this context may help the judge decide which component is useful and what was missing, redundant or incoherent.

4.2 Block-level Analysis

When judging blocks in aggregate (e.g., first obtaining judgments for the *algo block* across all queries), and then segmenting by certain query properties (e.g., query frequency, as per [16]), it becomes fairly straightforward to identify certain characteristics of current system-wide performance. For example, it may be that query suggestions are uniformly of high quality for high-frequency queries, but of low quality (or non-existent) for low frequency queries. In our experiments with the SASI method we have found that it provides useful qualitative insights which help to identify systematic issues with SERP components. See Section 5.1 for an example of a related experiment.

4.3 Position Effects

By providing appropriate judge guidelines, it is also possible to accommodate or penalize poor ranking decisions. The key is to provide instructions about which parts of the SERP are more important. For example, if an advertisement's content appears only loosely related to the dominant intent (e.g., Cricket wireless phone plans for a cricket sports query), the same advertisement content may be judged poor if shown in the mainline before the top-10 results, but fair if shown in the right rail. Anecdotally, in our studies with SASI, we have found that mingling relevance and SERP component placement can lead to challenges in communicating the guidelines for rating components to judges.

5. EXPERIMENTS

We carried out a number of experiments using SASI to explore different aspects of whole-page relevance and judging in context. The experiments demonstrate the value in considering the result page context in search system evaluation. Most experiments scraped the search results from the Google, Yahoo! and Microsoft Bing (and its predecessor Live) Web search engines. Scraping involves downloading the page contents corresponding to a website's URL; for search engines this means issuing a search query to the engine via the URL. In reporting findings in this section, the three engines are referred to (with a random permutation) as A, B and C. Non-parametric testing is used to test for statistical significance in observed differences where appropriate.

5.1 Comparative Component-level Quality

A starting point in our studies with SASI was an investigation of the quality of individual SERP components in aggregate. For the purposes of reporting here, we carried out SASI judging across engines A, B, and C using the 50 queries from the TREC 2009 Web Track [11] (hereafter referred to as *Experiment 1*). Each SERP was judged separately by many judges and at most once per judge. We used a total of 30 judges in this experiment, with an overlap factor of 5 judges per query, meaning each judge rated a randomly different subset of the 50 queries.. The average scores for some components and holistic qualities appear in Table 2 (range 0-2). Bolded text is used to depict the best performing system for each component/quality studied. The values in the table are listed in descending order in terms of the range between the top and bottom performing engine for each component. Statistical significance testing was performed using Kruskal-Wallis tests for each of the comparisons. Significant differences (both engine and degree) are highlighted in Table 2.

Table 2. Avg. scores for components and holistic qualities. Statistical significance levels are noted in sub/superscript († $p < 0.05$; ‡ $p < 0.01$). The preceding letter denotes the comparator engine with which the difference is significant.

Comp. or *Quality*	A	B	C	Range
Query suggestions	1.88	$^{A\dagger}_{C\ddagger}\mathbf{1.92}$	1.71	0.21
Right-side ads	1.76	1.61	$_{B\ddagger}\mathbf{1.77}$	0.16
Diversity	$_{C\ddagger}\mathbf{1.76}$	1.75	1.64	0.12
Caption quality	$_{C\ddagger}\mathbf{1.67}$	1.65	1.59	0.08
Top result	$_{C\ddagger}\mathbf{1.93}$	1.91	1.87	0.06
Result 5	1.80	$^{A\ddagger}_{C\ddagger}\mathbf{1.86}$	1.80	0.06
Result 4	1.81	$_{C\ddagger}\mathbf{1.86}$	1.80	0.06
Result 2	1.86	$_{C\ddagger}\mathbf{1.89}$	1.83	0.06
Overall satisfaction	$_{C\dagger}\mathbf{1.85}$	$_{C\dagger}\mathbf{1.85}$	1.80	0.06
Result 3	$_{C\dagger}\mathbf{1.87}$	1.85	1.83	0.04

The findings in Table 2 demonstrate that no single engine performs best on every count. For example, engine B was found to have superior query suggestions and poor advertisements, while engine A had the best top mainline results. Such findings support our intuition that SASI informs us about aspects of whole-page relevance beyond simple document ranking relevance.

5.2 Comparison with Document Judgments

To compare and contrast SASI with traditional IR evaluation methods, we can compare SASI judgments of mainline results against judged query sets, where relevance judgments were obtained per document, independent of the result-page context. To this end, we performed one experiment (hereafter referred to as *Experiment 2*) comparing SASI judgments and document judgments for a set of 74 judged queries from a proprietary dataset sampled with frequency bias from a large Web search engine query log. Document judgments were obtained using a separate judgment process involving pooling the top-ranked results of search engines from Google, Yahoo!, and Microsoft for the query, and judging the relevance of each of the documents. Documents were judged with respect to the query by trained human judges using a five-point relevance scale with the following options: *bad*, *fair*, *good*, *excellent*, and *perfect*. The 74 queries used in Experiment 2 reflected queries with documents judged in this way for which we had SASI judgments and a document judgment for the top-ranked mainline algorithmic result shown in the SASI interface (i.e., *top algo result*). We also performed a separate analysis comparing the SASI results from Experiment 1 (using the 50 TREC Web Track queries described earlier) against the original TREC document judgments. The purpose of these investigations was to examine areas of disagreement between rating methods and to investigate whether there was any systematic explanation for such disagreements that could be ascribed to the judgment method used.

5.2.1 Comparison using Log-sourced Queries

The graded relevance judgments available for both SASI judgments (three-point) and document judgments (five-point) of the top-ranked mainline algorithmic result meant we could compare the methods. Document judging scores were computed using normalized discounted cumulative gain [27] (NDCG). NDCG is an established measure of effectiveness used in evaluating result lists based on estimates of gain depending on where the document is in the ranked list. The reason to use just NDCG@1 is that there was a single SASI judgment to be compared with a single document label, rather than trying to create an aggregate metric over SASI component ratings. Different judge pools were used to obtain the document judgments and the SASI judgments. The judge pool for SASI ratings consisted of 10 judges and did not perform overlap judging (i.e., only one judge per query-{*top algo result*} pair). The judge pool was different from the SASI judge pool used for Experiment 1, and it was also disjoint from the judge pool used to obtain document ratings. Results for the 74 log-sourced queries in Experiment 2 are shown in the bubble plot in Figure 2. The x-axis plots NDCG@1, the y-axis the SASI judgment for the top-ranked mainline algorithmic document, and bubble size reflects the relative number of queries for that (x,y) coordinate. SASI ratings can be 0, 1 or 2; NDCG@1 was normalized to range from 0 to 100. Areas of substantial method agreement should reside on the diagonal from (0,0) to (100,2). Substantial disagreement would appear at (0,2) and (100,0).

As noted earlier, since SASI judges see only captions and document relevance judges see the full-text of documents, we expect to observe some disagreement. We note that SASI judges seemed generally more positive, with much of the judgment mass being on rating two in the y-axis. We hypothesize that it was difficult for the SASI judges to determine that all irrelevant documents were actually irrelevant from inspecting only their titles, snippets, and URLs on the SERPs. An alternative explanation is that since the captions often show a query-biased representation of each document, basing judgments on captions may cause judges to overestimate relevance. The investigations by Turpin et al. [40] corroborate that judging captions is not perfectly predictive of underlying document relevance; SASI is assessing complementary aspects of relevance.

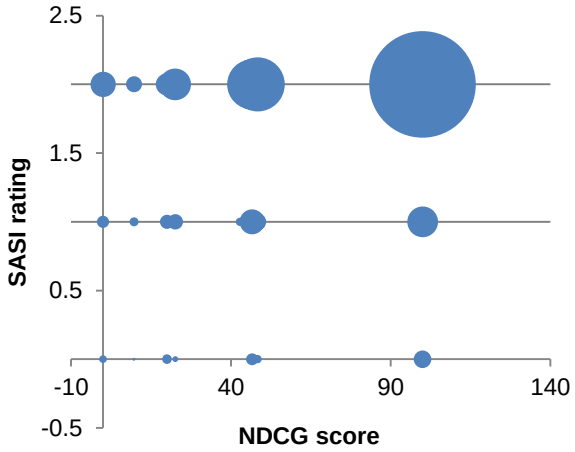

Figure 2. Bubble plot of SASI and NDCG agreement.

5.2.2 Comparison using TREC Queries

We performed a similar analysis for the 50 TREC queries with associated judgments from the TREC Web Track 2009 diversity sub-task. The judge pool for the TREC judgments came from the Web Track, for which there was no overlap with members of our SASI judge pool from Experiment 1. Of the top-ranked URLs retrieved by the Web search engines A, B, and C in response to these queries, 39% of these were un-judged in the TREC-judged document pools for the 50 TREC queries. This may reflect differences in the underlying Web corpus (since the Web has evolved since the ClueWeb09 corpus was crawled[1]) as well as ranking variations in Web search engines. In Figure 3, we plot the percentage of TREC judgments (relevant and irrelevant) against average SASI ratings for each query. We do not report on documents without TREC judgments (the remaining 39%) since these were immaterial for our comparative analysis.

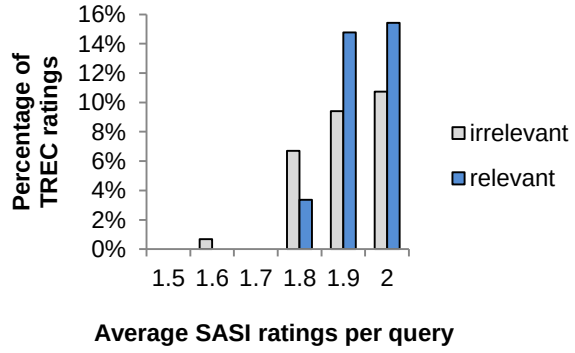

Figure 3. TREC judging ratio to SASI average rating.

[1] The ClueWeb09 is a new dataset comprising one billion Web pages crawled in early 2009, and used for new test collections to evaluate search systems at TREC.

We can observe from Figure 3 a useful degree of correspondence in that queries with highly-rated SASI judgments (either 1.9 or 2) are relevant more frequently than irrelevant. Lower average SASI ratings start to be considered more irrelevant than relevant in the TREC relevance judgment process.

To highlight differences between SASI judging and document judging, we investigated cases with high disagreement. We found that Web pages with a low SASI score and a high document relevance score tended to have a low-quality snippet. For example, a wedding budget calculator page was judged relevant by TREC judges, but its summary did not contain the word 'wedding' and hence was penalized by SASI judges. Such caption bias is expected and desired in SASI as we measure other aspects of the search system beyond document relevance.

Web pages with a high SASI score and low document relevance score tended to be associated with judgment or TREC topic development problems. For example, for the query [gogle] (spelling error intentional), when judging in context, SASI judges found the Google homepage to be a good answer, but document judges marked all retrieved Web pages as irrelevant. In TREC data, the topic [rick warren] has the diversity sub-topic "find the home page of Rick Warren," but the home page at www.rickwarren.com was judged irrelevant. SASI judges correctly judged it to be a good result. Other queries such as [volvo] and [atari] did not have a navigational sub-topic, so the TREC judges marked the homepages as irrelevant, disagreeing with SASI judges. Collectively, these comparisons illustrate the complementary aspects of the SASI method with existing Cranfield-style evaluation methods. They also suggest that clarity is required in the judging guidelines on each method so that judges fully understand the bases of their judgments.

5.3 Brand Bias Investigation

The way in which search results are presented can be an important determinant of search success. Unlike the Cranfield-style evaluation methods, issues of search-result presentation are testable using SASI. Of concern in doing so are potential variations arising from presentation effects. In particular, we were interested in whether there is an inherent brand bias that makes judges prefer one engine's results (as reported by Jansen et al. [26]).

To investigate the brand bias issue using SASI, we constructed a tool that could place the content from both top-10 search results and sponsored advertisements inside the branding elements of engines A, B, and C. This enabled us to conduct an experiment (referred to as *Experiment 3*) that grouped queries into four sets: (i) *AlgoSwap*: engine A-branded SERP with engine B's top-10 results; B-branded SERP with A's top-10 results, (ii) *AdSwap*: A-branded SERP with B's advertisements and vice versa, (iii) *AdAlgoSwap*: A-branded SERP with B's top-10 results and B's advertisements and vice versa, and (iv) *Control*: A-branded SERP with A's top-10 results and advertisements; B-branded SERP with B's top-10 results and advertisements.

In each of these four query sets, we also used the SERP results from engine C as an additional experimental signal regarding overall query difficulty in the set, and observed that the number of queries per set was small enough to see noticeable changes in average ratings per set. We selected queries from a large random sample that did not trigger any of the three engines' unique SERP features such as including news or video results within the mainline algorithmic search results, which may reveal the source engine identity and potentially bias the experiment. Ten judges carried out SASI ratings for all of the queries. There was complete overlap in the judging, meaning that the SERPs for each query were judged by all 10 SASI judges. This pool was the same pool that provided SASI ratings reported in Experiment 2.

In Table 4(a) we report on the mean average scores per component (range 0-2) and indicate (in bold) the results where a statistically significant preference is shown, treating all judge-query-component-SERP tuples as independent. We performed two-way Kruskal-Wallis tests with search engine and algo/ad swapping as factors. Post-hoc testing was used to decide if the results for a component/quality from one engine were statistically significantly different to those from another engine within each algo/ad swapping group. If brand bias were an issue in our judge pool, we would expect to see one engine always being preferred regardless of underlying relevance. As we can see in Table 4(a), this was not the case. For example, the most consistent component results are those for the mainline advertisements. For this, engine A was perceived to be better in the *Control* and *Algo* groups, and when engine A's advertisements were shown in engine B's branding they were again preferred at a statistically significant level. The results for the other components (right-rail advertisements and top-ranked search result) and the overall satisfaction rating were less consistent, but none showed consistent preference for engine A over engine B or vice versa.

We repeated the analysis but addressed the possibility of individual judge bias (for or against a particular brand) by first averaging each judge's rating per component for all SERPs from a particular set. Once again we performed two-way Kruskal-Wallis tests with search engine and algo/ad swapping as factors and associated post-hoc testing. As can be seen in Table 4(b), the variability increases in this analysis, leading to fewer differences being statistically significant. However, we again fail to see consistent judge preference for one search engine over the other in any component scores. We conclude that despite the presence of brand elements, our judges did not appear to be noticeably biased in favor or against a particular engine when using the SASI judging interface and guidelines.

5.4 Judge Agreement: Overall & By Feature

A common concern in relevance judging processes is the level of agreement among judges. Bailey et al. provided a useful overview of the history of studies in relevance assessment agreement [3]. According to their summary, agreement levels are relatively low (typically 30-50% using the Jaccard similarity coefficient) and even lower after correcting for agreement by chance. Cohen's κ is an established measure of inter-rater agreement [13]. However, since SASI experiments exceed two judges per component, Cohen's κ is inappropriate. Instead, we use Fleiss's κ [14], which measures multi-rater agreement corrected for chance. Complete agreement between raters leads to a value of one, while agreement no better than chance is below zero. The Landis and Koch scale for interpreting κ values [30] registers agreement as *poor* when < 0.0; *slight* as 0.0-0.20; *fair* as 0.21-0.40; *moderate* as 0.41-0.60; *substantial* as 0.61-0.80; and *strong* as 0.81-1.0.

In Table 5, we report on inter-rater agreement for the component and qualitative scoring of SERPs for the same 74 log-sourced queries as Experiment 2. Each cell reflects agreement among the 10 SASI judges. Statistically significant levels of agreement are shown in Table 5 in bold; fair agreement is shaded in light gray; poor agreement in dark gray. We include p-values (κ p) for cases

Table 4. Brand-bias investigation by SERP component and *qualitative* values († $p < 0.05$; ‡ $p < 0.01$).

		(a) Mean scores				(b) Means of judge means			
Component or *Quality*	**S.E.**	**Control**	**Algo**	**Ads**	**AlgoAds**	**Control**	**Algo**	**Ads**	**AlgoAds**
Mainline ads	A	**‡1.68**	**‡1.80**	1.55	1.70	1.85	**†1.80**	1.64	1.58
	B	1.56	1.52	**‡1.67**	**‡1.83**	1.61	1.39	1.73	**‡1.80**
Right-side ads	A	1.43	1.29	1.46	1.58	1.37	1.38	1.41	1.48
	B	1.43	**‡1.43**	**‡1.53**	**‡1.67**	1.39	1.29	1.50	1.58
Top result	A	1.65	**‡1.66**	**‡1.78**	1.67	1.63	1.77	1.77	1.71
	B	1.66	1.60	1.74	1.64	**†1.76**	1.70	1.77	1.65
Overall satisfaction	A	1.49	**‡1.53**	1.55	**‡1.59**	1.50	**†1.60**	1.60	1.65
	B	1.48	1.42	**‡1.58**	1.52	1.55	1.50	1.63	1.60

Table 5. Fleiss's κ inter-rater agreement for SERP components and *qualitative* values; 10 judges († $p < 0.05$; ‡ $p < 0.01$).

Search engine	A			B			C		
Component or *Quality*	**count**	**κ**	**κ *p***	**count**	**κ**	**κ *p***	**count**	**κ**	**κ *p***
Top result	73	**‡0.336**		74	**‡0.148**		74	**‡0.113**	
Right-side ads	43	**‡0.197**		46	**‡0.260**		43	**‡0.368**	
Page-top ads	28	**‡0.181**		33	**‡0.389**		27	**‡0.243**	
Top answer	5	**†0.149**	0.01	13	**‡0.179**		13	**‡0.313**	
Middle answer	25	**‡0.119**		4	-0.043	0.50	5	**‡0.173**	
Overall satisfaction	73	**‡0.085**		74	0.019	0.21	74	**‡0.096**	
Diversity	52	**‡0.074**		55	**†0.037**	0.02	51	**‡0.082**	
All top-10 results	73	**‡0.073**		74	**†0.043**	0.01	74	**‡0.101**	
Caption quality	73	**†0.042**	0.01	74	**‡0.068**		74	**‡0.063**	
Query suggestions	45	0.003	0.86	73	**‡0.062**		48	**†0.044**	0.02

where significance equals or exceeds 0.01. As shown, relatively few components or qualitative ratings reach fair agreement, and none achieve moderate agreement. These levels of agreement are slightly lower than those reported by Bailey et al. [3]. We note that qualitative values of full SERPs and multi-element components (e.g., all top-10 results, query suggestions) appear to be more challenging to reach agreement on than components with only a few elements (e.g., right-side ads).

To test whether the observed low inter-rater agreement was related to the SASI methodology, we ran a comparable experiment to that of Experiment 2 (where we compared SASI judging with document judging), this time focusing on judge agreement levels rather than SASI-document judgment correspondence. As part of a separate (non-SASI) relevance judgment process (hereafter referred to as *Experiment 4*), a pool of 11 judges rated approximately 38,000 query-URL pairs for document relevance using the five-point relevance scale described earlier (i.e., *bad*, *fair*, *good* etc.). None of these judges were part of the SASI judge pool. We computed the Fleiss's κ of the document judgments and found it to be 0.33 (fair agreement). This is similar to the agreement levels of 0.11-0.34 reported in Table 5 for the SASI judgments of the top result (the reported judged SERP component most similar to traditional query-URL judging). While SASI and document judging appear to be fairly similar in terms of agreement levels, agreement levels in SASI may be more susceptible to interaction effects between SERP elements.

In the final stage of our judge agreement analysis, we compute Fleiss's κ on ratings for a single component (first top-10 result) on a per-query basis for all three search engines with 30 SASI judges. Some of these judges were among those in the SASI judge pools in experiments 1, 2 and 3 reported earlier. The distribution of agreement levels is shown in Figure 4, with the x-axis as the κ value and the y-axis as the percentage of queries on engines A, B, or C with that agreement level. Figure 4 illustrates that there is considerable variation in agreement levels by query, and that different search engines also have different agreement distributions. The highest level of agreement is seen on the more noticeable components: the top result, query suggestions, and page-top ads. Judges show most agreement when rating results as good, and least when bad.

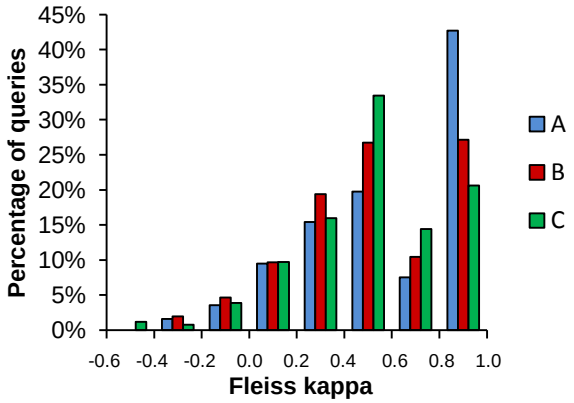

Figure 4. Distribution of per-query Fleiss's κ agreement.

Fleiss's κ may underestimate the degree to which raters conform in their ratings. If we consider our raters to be exchangeable, with no systematic difference between raters (i.e., the rating environment and displayed content are equivalent, and so the only difference would be due to the raters themselves), we would look for consistency, rather than agreement. Consistency can be measured using intra-class correlation [38]. The con-

sistency among the 30 SASI judges for the same first top-10 result is approximately 0.8 (substantial), across all three engines.

5.5 Summary of Findings

In this section we have described the findings of analysis performed using the SASI evaluation methodology. We have shown that the method can highlight significant differences in the component level quality of search systems, demonstrating that SASI can help inform us about aspects of whole-page relevance beyond the ranked list of search results. A comparison with document judgments for both log-sourced and TREC queries revealed agreement, accompanied by some positive bias in the SASI ratings. This bias could be ascribed to SERP presentation effects (judges were unable to identify irrelevant documents from captions and/or query-based snippets made many results appear relevant). Differences in SASI judgments and document judgments were attributable to either low-quality captions (when SASI ratings were low and document ratings were high) or issues with judgment guidelines or TREC topic development (SASI ratings high and document ratings low). An investigation of brand bias with SASI revealed that no apparent bias existed across all judges between engines or within each judge in favor or against a particular engine. Analysis of judge agreement on SASI revealed that it was fair and potentially dependent on the number of items in the block. In addition, SASI judge agreement was similar to document judge agreement and we found considerable variation in the agreement by query and by search engine. Follow-up analysis revealed substantial judgment consistency, signifying low variance within judges and indicating that much of the observed variation was attributable to component quality.

The experiments in this section have demonstrated some of the capabilities of SASI, including some complementary to the traditional Cranfield-style IR evaluation methods. From this analysis, it appears that there is benefit from performing judgments within a result page's context.

6. DISCUSSION

SASI is a method for performing in-context judgments relating to the holistic relevance of the SERP and about individual SERP components. While sharing some similarities with Cranfield-style evaluation, the SASI method does not generate a reusable test collection. Instead it gives visibility into different search system characteristics, which are normally evaluated using user studies or log-based A/B tests. It is also an efficient form of experimentation, with a whole SERP being judged in the time it takes to judge two documents in Cranfield-style experiments. In cases where disagreement between SASI and document judging was observed in our experiments, the issue was typically related to SERP presentation (leading judges to over-estimate result relevance) or issues related to the clarity of judgment guidelines (leading judges to employ too strict a criteria in the case of TREC judgments). More work is needed on refining the SASI methodology to reduce the effect of such factors, perhaps drawing on related research on expert judgment, e.g., [32].

As we have shown, SASI is a potentially valuable method for allowing individual SERP components to be evaluated in their result page context. However, the method has limitations and one tradeoff it makes is to evaluate the SERP component with respect to the full SERP, where issues such as diversity are in tension with others such as result coherence. One judge might make a judgment against diversity (and in favor of coherence), when actually diversity caters for more users' needs. It is therefore important to correctly balance the individual SERP preferences of any judge against those of a large disparate user population, especially for ambiguous queries. For such queries it may be prudent to employ multiple judges and/or leverage log data to develop a better understanding of the range of possible search intents associated with the particular query. A deeper analysis of inter-rater agreement, especially with and without intent hints would be valuable to conduct. Heuristic usability analysis techniques could also be incorporated to emphasize different contextual aspects through SASI's judging interface. Another abstraction in the method is to assume a relatively standardized SERP layout and non-interactive components across search engines. It does not support relevance or holistic judging for those aspects of a SERP which are highly interactive.

Our SASI judgment system has primarily been used by presenting judges with solely the individual queries to be judged. This approach is limited, in that it ignores variable query intents or significant user attributes that might affect the assessment of relevance. An additional area of investigation is to provide contextual information regarding query intent as part of the per-query instructions (where the judge first views the query to be assessed). We have carried out preliminary trials for augmenting seasonally-specific queries with descriptive information, akin to TREC topic descriptions. The same technique could be used to convey certain semantic interpretations of the query or the user's orientation in place and time (e.g., location, time of day), perhaps framed within simulated work tasks [6].

7. CONCLUSIONS AND FUTURE WORK

We have introduced SASI, a novel evaluation method that deliberately considers only surface-level information on a search results page, allowing us to evaluate search systems via the result page context. We have found that SASI provides a useful complement to Cranfield-style document relevance approaches for assessing search engine relevance. In particular, it can help to identify issues of concern, such as poor component quality, both in isolation for an engine of interest or comparatively. It also allows us to rate whole-page issues such as diversity and inter-component incoherence. Outcomes of SASI experiments involving brand bias, comparisons with document judgments, and inter-judge agreement lead us to conclude that the method is not seriously flawed and shows promise in moving beyond document judging as the primary IR evaluation technique. Judgment efficiency in SASI suggests that it may scale well to large query sets or large numbers of experimental variants.

Future development of SASI could involve tests of real user interface and relevance algorithm alternatives, rather than only comparisons between competing search systems as reported in this paper. The flexibility of the approach should allow judges to focus on the components in question, even if their effect is on holistic SERP attributes. The method might also be broadly applicable to other search systems, such as those used for enterprise search or libraries. Another extension would be to bridge the gap with Cranfield methods by developing a result page context-blind variant of SASI that could create reusable judgments. However, this would sacrifice some of the strength of the SASI method, derived from judging *in* context.

REFERENCES

[1] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. (2008). The good and the bad system: does the test collection predict users' effectiveness? *Proc. SIGIR.*

[2] J. Allan, B. Carterette, and J. Lewis. (2005). When will information retrieval be "good enough"? *Proc. SIGIR*.

[3] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A.P. de Vries, and E. Yilmaz. (2008). Relevance assessment: are judges exchangeable and does it matter. *Proc. SIGIR*.

[4] P. Bailey, N. Craswell, R.W. White, L. Chen, A. Satyanarayana, and S.M.M. Tahaghoghi. (2010). Evaluating whole-page relevance. *Proc. SIGIR*, (poster paper).

[5] P. Bailey, P. Thomas, and D. Hawking. (2007). Does brandname influence perceived search result quality? Yahoo!, Google, and WebKumara. *Proc. ADCS*.

[6] P. Borlund. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3): 71-90.

[7] G. Buscher, S. Dumais, and E. Cutrell. (2010). The good, the bad, and the random: An eye-tracking study of ad quality in web search. *Proc. SIGIR*.

[8] B. Carterette, P.N. Bennett, D.M. Chickering, and S. Dumais. (2008). Here or there: preference judgments for relevance. *Proc. ECIR*, 16-27.

[9] R. Chandrasekar, M.R. Scott, D. Slawson, A.R.D. Rajan, and D. Makoski. (2008). Measuring search experience satisfaction using explicit context-aware feedback. *Proc. Workshop on Human-Computer Interaction and Information Retrieval*.

[10] C.L.A. Clarke, E. Agichtein, S. Dumais, and R.W. White. (2007). The influence of caption features on clickthrough patterns in web search. *Proc. SIGIR*.

[11] C.L.A. Clarke, N. Craswell, and I. Soboroff. (2009). Overview of the TREC 2009 web track. *Proc. TREC*.

[12] C.W. Cleverdon. (1960). ASLIB Cranfield research project on the comparative efficiency of indexing systems. *ASLIB Proceedings*, XII, 421-431.

[13] J. Cohen. (1960). A coefficient for agreement for nominal scales. *Education and Psych. Measurement*, 20: 37-46.

[14] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. (2008). An experimental comparison of click position-bias models. *Proc. WSDM*, 87-94.

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

[16] D. Downey, S. Dumais, D. Liebling, and E. Horvitz. (2008). Understanding the relationship between searchers' queries and information goals. *Proc. CIKM*, 449-458.

[17] S. Dumais and N. Belkin. (2005). The TREC interactive tracks: putting the user into search. *TREC: Experimentation and Evaluation in Information Retrieval*. MIT Press.

[18] J. Fleiss. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378-382.

[19] Z. Guan and E. Cutrell. (2007). An eye tracking study of the effect of target rank on web search. *Proc. CHI*, 417-420.

[20] D. Hawking and N. Craswell. (2005). The very large collection and web tracks. *TREC: Experimentation and Evaluation in Information Retrieval*, 199-232. MIT Press.

[21] D. Hawking, N. Craswell, P. Bailey, and K. Griffiths. (2001). Measuring search engine quality. *Information Retrieval*, 4(1): 33-59.

[22] D. Hawking, N. Craswell, P. Thistlewaite, and D. Harman. (1999). Results and challenges in Web search evaluation. *Proc. WWW*, 1321-1330.

[23] W. Hersh. (2002). TREC 2002 interactive track report. *Proc. TREC*.

[24] P. Ingwersen and K. Järvelin. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer.

[25] B.J. Jansen and U. Pooch. (2001). Web user studies: a review and framework for future work. *JASIST*, 52(3): 235-246.

[26] B.J. Jansen, M. Zhang, and C.D. Schultz. (2009). Brand and its effect on user perception of search engine performance. *JASIST*, 60(8): 1572-1595.

[27] K. Järvelin and J. Kekäläinen. (2000). IR evaluation methods for retrieving highly relevant documents. *Proc. SIGIR*.

[28] D. Kelly, K. Gyllstrom, and E.W. Bailey. (2009). A comparison of query and term suggestion features for interactive searching. *Proc SIGIR*.

[29] R. Kohavi, T. Crook, and R. Longbotham. (2009). Online experimentation at Microsoft. *Proc. Workshop on Data Mining Case Studies and Practice Prize*.

[30] J. R. Landis and G. G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33: 159-174.

[31] J. Li, S. Huffman, and A. Tokuda. (2009). Good abandonment in mobile and PC internet search. *Proc. SIGIR*.

[32] M.B. Meyer and J.M. Booker. (2001). *Eliciting and Analyzing Expert Judgment: A Practical Guide.* ASA-SIAM.

[33] J. Nielsen and K. Pernice. (2009). *Eyetracking Web Usability*. New Riders Press.

[34] F. Radlinski, M. Kurup, and T. Joachims. (2008). Learning diverse rankings with multi-armed bandits? *Proc. ICML*.

[35] F. Radlinski, R. Kleinberg, and T. Joachims. (2008). How does clickthrough data reflect retrieval quality? *Proc. CIKM*, 43-52.

[36] F. Radlinski and T. Joachims. (2006). Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. *Proc. AAAI*, 1406-1412.

[37] F. Radlinski and N. Craswell. (2010). Comparing the sensitivity of information retrieval metrics. *Proc. SIGIR*.

[38] P.E. Shrout and J.L. Fliess (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2): 420-428.

[39] P. Thomas and D. Hawking. (2006). Evaluation by comparing result sets in context. *Proc. CIKM*, 94-101.

[40] A. Turpin, F. Scholer, K. Järvelin, M. Wu, and J.S. Culpepper. (2009). Including summaries in system evaluation. *Proc. SIGIR*, 508-515.

[41] A. Turpin and F. Scholer. (2006). User performance versus precision measures for simple search tasks. *Proc. SIGIR*.

[42] E. M. Voorhees. (2008). On test collections for adaptive information retrieval. *Information Processing and Management*, 44(6): 1879-1885.

[43] E. M. Voorhees and D. Harman. (2005). *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.

[44] K. Wang, T. Walker, and Z. Zheng. (2009). PSkip: estimating relevance ranking quality from web search clickthrough data. *Proc. KDD*, 1355-1364.

[45] R.W. White, M. Bilenko, and S. Cucerzan. (2007). Studying the use of popular destinations to enhance web search interaction. *Proc. SIGIR*.