Comparing the Sensitivity of Information Retrieval Metrics

Filip Radlinski
Microsoft
Cambridge, UK
filiprad@microsoft.com

Nick Craswell
Microsoft
Redmond, WA, USA
nickcr@microsoft.com

ABSTRACT

Information retrieval effectiveness is usually evaluated using measures such as Normalized Discounted Cumulative Gain (NDCG), Mean Average Precision (MAP) and Precision at some cutoff (Precision@k) on a set of judged queries. Recent research has suggested an alternative, evaluating information retrieval systems based on user behavior. Particularly promising are experiments that interleave two rankings and track user clicks. According to a recent study, interleaving experiments can identify large differences in retrieval effectiveness with much better reliability than other click-based methods.

We study interleaving in more detail, comparing it with traditional measures in terms of reliability, sensitivity and agreement. To detect very small differences in retrieval effectiveness, a reliable outcome with standard metrics requires about 5,000 judged queries, and this is about as reliable as interleaving with 50,000 user impressions. Amongst the traditional measures, NDCG has the strongest correlation with interleaving. Finally, we present some new forms of analysis, including an approach to enhance interleaving sensitivity.

Categories and Subject Descriptors

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

General Terms: Algorithms, Measurement

Keywords: Interleaving, Evaluation, Search

1. INTRODUCTION

A tremendous amount of research has improved information retrieval systems over the last few decades. As effective approaches mature and relative improvements become smaller, the sensitivity of evaluation metrics and their fidelity to actual user experience becomes increasingly critical. Without sensitive measurement we might reject a small but significant improvement. This becomes a problem if we reject a large number of independent small improvements, because we have forgone an overall large improvement. Without fidelity in measurement, a small change in a retrieval model might be taking into account some bias of relevance judges, rather than the preferences of real users.

The predominant form of evaluation in information retrieval is based on test collections (e.g. [19]) comprising query topics, a document corpus and human relevance judgments of topic-document pairs. This allows the application of standard metrics such as NDCG, MAP and Precision@k. Sensitivity depends on the number of topics and judgments. Fidelity depends on whether the test collection reflects real-world search behavior. For example, the TREC Web Track found that changing from informational to navigational [4] assumptions when judging can change the outcome of an evaluation [19, chapter 9]. Experiment outcomes can also be affected by an assessor's level of background knowledge [14].

An alternate evaluation approach is based on user behavior, estimating user success by measuring click, re-querying and general browsing patterns on search results. This can be motivated on grounds of fidelity and cost. On fidelity, judges are usually far removed from the search process, so may generate unrealistic query topics from observed queries, and have a hard time assessing documents in a way that reflects a user's actual information need. Additionally, traditional measures combine document judgments to obtain a score per query, for example based on discount and gain, but these may not match real user experience. Finally, judgments are slow and expensive to collect. For a system with real users, usage-based evaluation is far cheaper, despite the fact that the click data collected may not be reusable in the way that most test collections are.

This paper considers the reliability, sensitivity and agreement of these competing evaluation approaches. On the Cranfield/TREC side, we consider relevance judgments for up to 10,000 queries. On the user metric side, we perform click-based tests involving the interleaving of two retrieval functions over 200,000 user *impressions*, which we define as events where a user runs a query and clicks a result. Using a large commercial dataset, we establish results that we believe would also hold true in an academic setting.

We test sensitivity by measuring outcomes with varying numbers of queries/impressions. This is done on pairs of retrieval functions with varying degrees of difference, including one pair with a very small difference in effectiveness. Our results show that both approaches can be very sensitive, but judged evaluation may require thousands of judged queries to obtain the required sensitivity.

We test agreement in overall outcomes between traditional measures and interleaving. We tend to find agreement, which is an indication of the fidelity of the judgment-

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

based metric, since it is agreeing with an experiment involving real users. We then study various new ways of aggregating and analyzing the interleaving data, showing how to improve agreement with traditional metrics and also attain reliability with fewer impressions. We also show that, in contrast to judgment-based metrics, interleaving can measure the fraction of users for whom a ranking change was meaningful. This allows assessment to move beyond an assumption that relevance for all users is identical, and that relevance of individual documents should be aggregated identically for all queries.

2. RELATED WORK

A small number of previous studies have evaluated the sensitivity of MAP and Precision@10 in the TREC setting [18, 19]. Voorhees and Buckley [8] concluded that an absolute difference in MAP of five to six percent was needed between two retrieval functions before the direction of the difference between them as measured on fifty TREC topics is reliable. Sanderson and Zobel [13] found that an even larger difference is necessary. This paper compares traditional measures (with larger query sets) against interleaving, which was found by Joachims and collaborators to be particularly sensitive to ranking changes [17, 10].

A few previous papers studied the agreement between TREC-style evaluation and user studies. Hersh, Turpin and their collaborators found that MAP does not correlate with the time it takes users to find relevant documents [11, 2]. Allan et al. found that bpref correlated with user search effectiveness only for some quality differences [12].

On the other hand, Al-Maskari et al. [1] found that various metrics including search time, number of relevant documents found and users' perceived satisfaction differ significantly when comparing behavior between the best and worst of performing of three common information retrieval systems for TREC topic queries (as measured by MAP). Additionally, large differences in precision have been found to correlate with how long it takes users to find relevant documents [9], and user perception of result relevance [7]. However none of these evaluations detected small changes in ranking quality, which are of interest when developing retrieval algorithms.

Simply observing user clicking behavior on a real search system, Carterette and Jones [3] found a correlation between clicks and DCG on advertisements. In addition, Huffman and Hochster show that satisfaction (as estimate by judges) correlates with a DCG performance metric based on judgments of the top three retrieved documents [15]. Finally, Radlinski et al. found that a number of commonly measured usage-based ranking metrics, such as time to first click, rank of click and fraction of abandoned queries, do not reliably correlate with ranking quality on an academic article collection given large differences in ranking quality [10] . However, they found that an interleaved evaluation did allow clicks to identify the better of two rankings quickly and reliably. Despite this, the lack of relevance judgments on their collection left open the question as to whether metrics such as MAP, NDCG and precision correlate (or even agree) with interleaving.

Finally, a less common evaluation approach asks users or judges to select the better of two rankings shown side-by-side [16]. When done by judges, the same challenges exist as with judging topic-document pairs. If done by users, this requires a different search interface, meaning the evaluation cannot be done with users in a natural setting.

3. EXPERIMENT DESIGN

In this section we detail the retrieval systems evaluated, and the metrics we use.

3.1 Retrieval Systems

We evaluate the differences between five pairs of *rankers* (retrieval functions) produced during normal development by a large commercial search engine. We treat the rankers as black boxes: for a query, each ranker produces an ordered set of results. We split our experiments by the magnitude of changes they measure.

Major experiments: The first three experiments we present involve major revisions of the web search ranker, which we refer to as *rankerA*, *rankerB* and *rankerC*. Experiment *majorAB* compares rankers A and B, with experiments *majorBC* and *majorAC* named equivalently. The differences between these rankers involve changes of over half a percentage point of MAP and NDCG. These were chosen because the changes in retrieval quality are of similar magnitude to those commonly seen in recent research publications.

Minor experiments: The remaining two experiments involve minor modifications to the ranking system – we term these *minorD* and *minorE*. The overall differences involve changes in retrieval performance of under 0.2 points (out of 100) of MAP and NDCG, chosen as they are typical of incremental changes made during algorithm development. Experiment *minorD* involves a change in the processing of rare queries, with a large effect on the performance of a small fraction of queries. Experiment *minorE* involves a small change in search engine parameters, with a small effect on the performance of many queries.

3.2 Evaluation with Judgement-Based Metrics

Each ranker was evaluated using both standard information retrieval metrics and based on user traffic. The standard metrics were evaluated using approximately 12,000 queries uniformly sampled from a real workload as part of previous work (allowing frequent queries to appear multiple times, and omitting queries classified as adult by human annotators). The relevance of the top ten results returned by each ranker were assessed by trained judges on a five-point scale ranging from "bad" to "perfect". As precision and MAP both require binary relevance judgments, we binarized the ratings by taking the top two levels as relevant, and bottom three as non-relevant. This is consistent with the recent observation by Scholer and Turpin that precision and user metrics are better correlated when slightly relevant documents are grouped with non-relevant documents rather than with highly relevant documents [9].

Given Q queries, we compute precision at cutoff 5[1] for retrieval algorithm R as follows

$$Precision@5(R) = \frac{1}{Q}\sum_{i=1}^{Q}\left[\frac{1}{5}\sum_{i=j}^{5} rel^b(d_j^i)\right]$$

where d_j^i is the j^{th}-ranked document returned by R in response to query q_i, and $rel^b(d_j^i)$ is the binarized relevance assessment of this document. We compute Mean Average Precision (MAP) similarly, except that instead of measuring MAP down to a deep rank (such as 1,000 in TREC), we

[1]Chosen because few users look at results below the top 5.

Algorithm 1 Team-Draft Interleaving

1: **Input**: Rankings $A = (a_1, a_2, \dots)$ and $B = (b_1, b_2, \dots)$
2: **Init**: $I \leftarrow (); TeamA \leftarrow \emptyset; TeamB \leftarrow \emptyset;$
3: **while** $(\exists i : A[i] \notin I) \wedge (\exists j : B[j] \notin I)$ **do**
4: **if** $(|TeamA| < |TeamB|) \vee$
$((|TeamA| = |TeamB|) \wedge (RandBit() = 1))$ **then**
5: $k \leftarrow \min_i\{i : A[i] \notin I\}$ ... *top result in A not yet in I*
6: $I \leftarrow I + A[k];$ ..... *append it to I*
7: $TeamA \leftarrow TeamA \cup \{A[k]\}$ ..... *clicks credited to A*
8: **else**
9: $k \leftarrow \min_i\{i : B[i] \notin I\}$ ... *top result in B not yet in I*
10: $I \leftarrow I + B[k]$ ..... *append it to I*
11: $TeamB \leftarrow TeamB \cup \{B[k]\}$ ..... *clicks credited to B*
12: **end if**
13: **end while**
14: **Output**: Interleaved ranking I, $TeamA$, $TeamB$

limit ourselves to only the top ten documents[2]:

$$MAP@10(R) = \frac{1}{Q}\sum_{i=1}^{Q}\left[\frac{1}{n_i}\sum_{j=1}^{10} rel^b(d_j^i)\cdot Precision@j(R, q_i)\right]$$

where for query q_i there are n_i known relevant documents. We measure NDCG using an exponential gain and logarithmic decay based on the graded relevance judgments:

$$NDCG@5(R) = \frac{1}{Q}\sum_{i=1}^{Q}\left[\frac{1}{N_i}\sum_{j=1}^{5}\frac{2^{rel(d_j^i)}-1}{\log(j+1)}\right]$$

where N_i is the maximum possible DCG given the known relevant documents for q_i. Due to space constraints, we refer the reader to [5] for more details about these metrics.

3.3 Evaluation with Interleaving

Interleaved evaluation, originally proposed by Joachims [17], combines the results of two retrieval functions and presents this combination to users (essentially, alternating between the results from the two rankings while omitting duplicates). The users' clicks indicate a relative preference comparing the quality of two retrieval functions: the ranking that contributed the most clicked results is considered to be better. Radlinski et al. [10] showed that Joachims' interleaving approach, as well as a modified approach they introduced, detects changes in ranking quality much more reliably than other click-based metrics.

Our evaluation on real user traffic using interleaving involved showing the rankings produced for each experiment to a small fraction of users of a commercial search system until 220,000 impressions of non-adult queries with clicks had been observed. The experiments were performed in succession over two months, with each experiment run on the same days of the week (Tuesday through Friday) to avoid any weekday/weekend effects.

Producing Interleaved Rankings

We now describe our specific interleaving algorithm, the Team-Draft approach introduced by Radlinski et al. [10]. Let A and B be retrieval functions. Given the results for a query q, $A(q) = (a_1, a_2, \dots, a_n)$ and $B(q) = (b_1, b_2, \dots, b_m)$, Team-Draft interleaving combines these results into a single ranking. This algorithm is motivated by how sports teams are often assigned in friendly games: Given a pool of available players, two captains take turns picking the next preferred available player for their team. This approach treats $A(q)$ and $B(q)$ as team captains' preference orders. Subject to a coin toss after every pick, the rankings take turns "picking" the next available result for their "team", with the ranking shown to the user who issued the query being the pick order. An example ranking produced by Team-Draft interleaving, along with team assignments, is shown in Figure 5. The full algorithm is presented in Algorithm 1. For further details, we refer the reader to [10].

In addition, our implementation involves a minor modification to this algorithm due to the many near duplicate documents commonly found on the web. While each ranker avoided returning near-duplicates, each may return a different near-duplicate of the same result. Hence steps 5 and 9 were modified: when verifying the next result in the preference order of $A(q)$ and $B(q)$ was not already selected, we also skip over a result if it is very similar to one already selected, using the similarity measure described in [6].

Credit assignment

Given an interleaved ranking I produced by Algorithm 1 with team assignments $TeamA$ and $TeamB$, and clicks on these results, we must determine which ranking is considered better. To do this, we simply count how many distinct results were clicked on for each team. If one of the teams received clicks on more documents, this impression counts as a preference for that team. Otherwise it is a tie, and the impression is ignored. Note that the actual number of clicks is ignored, as is the order of clicking and the rank at which the clicked documents were presented. We will explore alternative credit assignment approaches in Section 7.4.

3.4 Research Questions

In the rest of this paper, we ask the following questions: (1) How many queries must be judged to obtain significant results for each metric given realistic ranking quality differences? (2) Does interleaving produce correlated results with judgment-based metrics? (3) How many impressions are needed to obtain comparable results? (4) How do interleaving algorithm design choices affect the outcome of the evaluations, and how can we extend interleaving analysis?

4. STABILITY OF JUDGMENT METRICS

In this section, we evaluate how the outcome of a comparison between ranking functions depends on the number of queries assessed when evaluated according to standard information retrieval metrics. We start with the previously described set of about 12,000 queries. From this set, we subsample n queries (with replacement) and measure the difference in the score of each input ranking according to NDCG@5, MAP@10 and Precision@5, repeating the sampling 1,000 times for each n. We then count the fraction of sampled sets of queries where each of the input rankings scored higher, ignoring cases when the scores were identical. The outcome of this evaluation is presented in Figure 1.

The top plot in Figure 1 shows the fraction of query samples for which the ranker hypothesized to be better (by the ranker developers) obtains a higher average NDCG@5 score than the other ranker, versus the number of queries in the query set evaluated. For very small query set sizes, each

[2]Deeper judging of documents was impractical due to the large number of queries assessed. However, since n_i is the number of known relevant documents, we are essentially assuming that anything not in the top 10 is unranked.

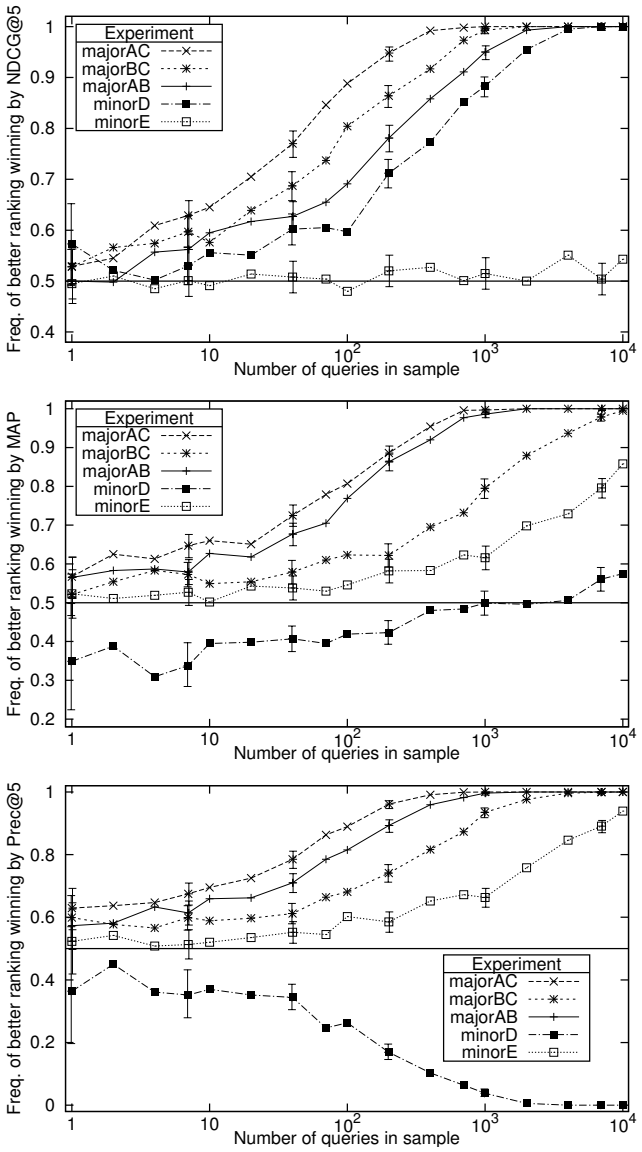

Figure 1: Query set size vs. the frequency with which "better" ranking scores higher.

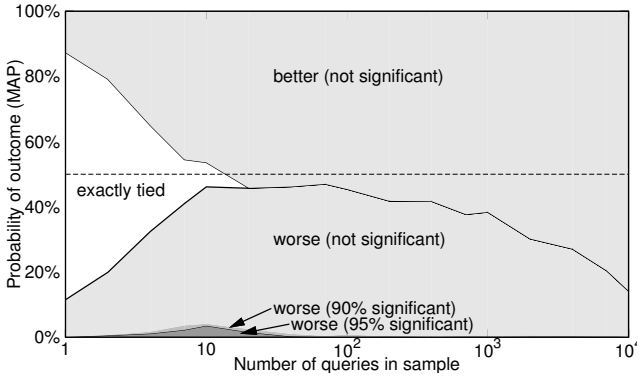

Figure 2: Preferred ranking drilling down by outcome for experiment *minorE* measured with MAP.

ranker has higher NDCG@5 roughly half the time. Once the query set size is comparable to typical TREC evaluation sets of 50 to 200 queries, there is a preference for the better ranker on between 50 and 90 percent of samples. Once the query set size approaches about 1,000 queries, one of the rankers tends to be consistently identified as better, with the exception of the *minorE* experiment. As would be expected, larger changes in ranking quality can be detected with smaller query set sizes. It is worth noting, however, that even with 10,000 queries in the sample, the outcome for experiment *minorE* is still uncertain using NDCG@5.

The middle plot in the figure shows the same results using MAP@10 to evaluate performance. Although the results are similar to those obtained with NDCG for the major experiments, this is not the case for the minor experiments, which measure smaller ranking changes. *MinorD* involves a change where we expect to significantly improve performance on a small fraction of queries. For small numbers of queries, the "better" ranker in fact performs worse according to MAP, while the opposite is true with large numbers of queries. This happens for two reasons. First, the binarization of the relevance judgments makes MAP scores behave differently than NDCG@5 scores. Second, according to the binarized scores, the improved ranker actually reduces performance slightly for many frequent queries, while improving some rare queries dramatically. If only one query is picked at random, it is usually a frequent query, and hence the unimproved ranker scores higher for most query sets. However, once the set of queries becomes large enough that at least one rare query is usually selected, the average change in MAP on the entire set of queries becomes positive. Experiment *minorE* sees consistent improvements with large query set sizes, unlike with NDCG. Also note that the relative differences in *majorAB* and *majorBC* are different when using MAP than when using NDCG. This can be explained by noting that perhaps *majorBC* involved more improvements in finding medium relevance documents, while *majorAB* involved more improvements in finding highly relevant documents (with MAP only sensitive to the latter).

The lower plot in Figure 1 shows the outcome as measured by Precision@5. Interestingly, the outcome for experiment *minorD* disagrees with both MAP and NDCG for large query set sizes. We hypothesize this difference to happen because improvements to rare queries often occur at lower ranks, with MAP and NDCG both less sensitive to such changes than Precision. A relevant document for a frequent query dropped out of the top 5 more often than a relevant document was added to the top 5 for a rare query.

We also tried taking the top *three* levels as relevant when binarizing our five levels of judgments (rather than the top *two* levels), in which case the plots for MAP and Precision@5 become more similar to those for NDCG@5. This suggests that the changes made in the minor experiments happen precisely to documents near this relevance threshold, and the choice of threshold is critical when evaluating ranking quality using metrics based on binary relevance. One could argue that this is evidence that the correct threshold for "relevant" is lower (so that all three metrics agree), yet perhaps one of the metrics better agrees with user behavior: we will study this in the next section. Our results also suggest that if only highly relevant documents are considered relevant (as found by [9], although based on judgments collected with very different judging guidelines), NDCG and MAP may disagree on the relative ordering of some ranking functions.

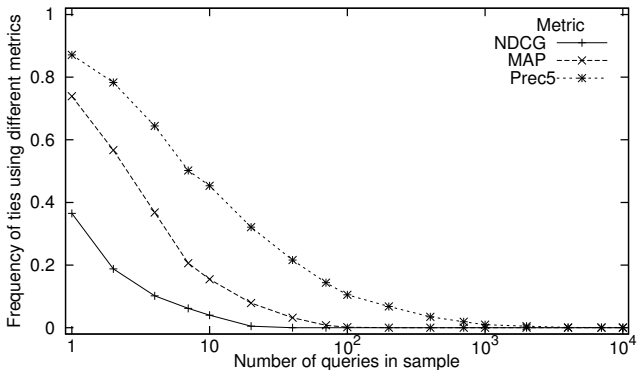

Figure 3: Frequency of ties vs. query set size

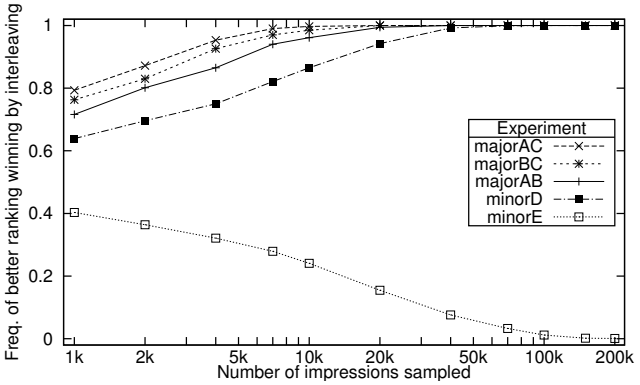

Figure 4: Number of query impressions during evaluation vs. the frequency with which hypothesized better ranking wins.

To further analyze the effect of queryset size on evaluation outcome, Figure 2 shows more detail for one metric (MAP) for one experiment (*minorE*). It shows the fraction of query samples for which a two-sided paired t-test indicates that one ranking is significantly better or worse (of 1,000 samples for each query set size). For small query set sizes, the performance on the set is usually exactly tied. As the query set size increases, more often than not the hypothesized better rankings is preferred (for up to about 85% of query sets of size 10,000). However, for small query sets from five to 30 queries, the *worse* ranking is sometimes statistically significantly better. At this significance level, this is not unusual (this incorrect conclusion with 95% confidence is drawn less than 5% of the time), but it is interesting that the difference is never significant the other way: For no query set, even consisting of 10,000 queries, does a t-test indicate that the hypothesized better ranking is better, although it is preferred by 85% of the selected query sets.

Finally, Figure 1 ignores queries where the scores are tied. As we saw that ties are very frequent at small sample sizes, Figure 3 shows how often each of the metrics were tied depending on the number of queries judged, averaged across the five experiments. As expected, ties are more common for small numbers of queries and for Precision@5 and MAP, which can take fewer values than NDCG@5.

5. SENSITIVITY OF INTERLEAVING

In this section, we perform a similar analysis to the previous section but with interleaving. From the 220,000 impressions observed for each experiment (except for the *majorAC*, where due to a misconfiguration only 190,000 impressions were collected) we sample impressions at random, obtaining from 1,000 to 200,000 sampled impressions. For each number of impressions, we evaluate which ranking was preferred by interleaving. This is repeated 1,000 times for each sample size, and the results are plotted in Figure 4. The figure shows the fraction of impression samples for which the ranking hypothesized to be better was indeed preferred by interleaving. The errors bars are too small to be visible.

We see a similar result as when sampling judged queries, with the major experiments agreeing with the hypothesized direction for even small numbers of impressions, and slower convergence for the minor experiments. However, even with just 1,000 impressions one ranking was consistently preferred 60% to 80% of the time. Moreover, as the number of impressions grows, the preferred ranking is always preferred for a larger fraction of samples, without the flipping behavior seen for MAP earlier. From the plot, interleaving results are 95% reliable after about 50,000 impressions, which corresponds to the standard IR metrics with about 5,000 judged queries for the major experiments, and over 10,000 queries with the minor experiments. Also, note that the outcome of *minorE* disagrees with the binarized judgment-based metrics (MAP, Precision@5) when the top two relevance levels are taken as relevant, and provides a statisticaly significant outcome whereas NDCG@5 does not.

5.1 Query level sensitivity

Given the consistency of interleaving with even a relatively small number of impressions, we now investigate whether it can be used to drill down further to analyze ranking performance. Figure 5 shows one example impression for the query *shaun cassidy ruby and the rockits* during experiment *majorAC* (the URLs shown are shortened to fit). This query appears 27 times with a non-draw outcome (at least one click, and not an equal number on each "team") in the 190,000 total impressions. In the example, we see that the rankings differed in where they returned the Wikipedia page for Shaun Cassidy. In fact, for 70% of clicked impressions, this result was clicked and determined the winner according to interleaving – and always preferred Ranking C. This suggests that this particular web result is most relevant for users who issued this query. Such an analysis can provide a detailed view not only of which ranker was preferred, but which results contributed to this preference for each frequent query.

Note that in addition to being useful as an evaluation tool, identifying particularly important differences in the rankings that affect user behavior could be used to generate training data for learning to rank. We leave this as future work.

6. CORRELATION BETWEEN METRICS

We can now address our second question: Does interleaved evaluation agree with standard information retrieval metrics in direction as well as in magnitude? We answer it by combining the results from the previous two sections. For each experiment, Figure 6 shows the relative NDCG@5, MAP@10 and P@5 difference versus the deviation from 50% observed with interleaving. For example if *rankerA* is preferred to *rankerB* for 52% of impressions in some experiment, we plot this as a 2% interleaving signal. The error bars on the judgment-based metrics indicate the 95% confidence intervals using 1000 samples of 10,000 queries as in Section 4.

team Presented	Ranking A	Ranking C
(A) facebook/ShaunCassidy_wall	(@1) facebook/ShaunCassidy_wall	X (@2) wikipedia/Shaun_Cassidy
X(C) wikipedia/Shaun_Cassidy	–(@2) wikipedia/Shaun_Cassidy	(@4) usatoday.com/life/television/...
(A) facebook/ShaunCassidy	(@3) facebook/ShaunCassidy	(@5) wikipedia/Ruby_&_The_Rockits
(C) usatoday.com/life/television/...	(@6) prime-time-...suite101.com/...	(@7) facebook/ShaunCassidy_wall
(C) wikipedia/Ruby_&_The_Rockits	(@7) thedeadbold.com/news/...	(@6) prime-time...suite101.com/...
(A) prime-time...suite101.com/...	(@8) buzzdash.com/polls/ruby-the..	(@10) puckettsprojects.com/2009/07/...

Figure 5: Example impression for query "shaun cassidy ruby and the rockits" from experiment *majorAC*. "X" indicates a click that counts, "–" the clicked URL on the other ranking, "@n" is the rank at which results from each input were shown. Of 27 impressions with a non-draw outcome, wikipedia/ShaunCassidy determined the outcome 19 times (70% of the time).

Table 1: Correlation between IR metrics and interleaving experiments.

Inter'l Scoring	IR Metric	Correlation	p-value
Per impression	NDCG@5	0.882	0.048
	MAP@10	0.689	0.198
	P@5	0.662	0.223
Per query	NDCG@5	0.910	0.032
	MAP@10	0.776	0.122
	P@5	0.733	0.159

The error bars on interleaving are 95% binomial confidence intervals given all the impressions for each experiment.

The figure shows that NDCG@5 is highly correlated with interleaving, with the other metrics being somewhat less correlated (although the difference is not statistically significant due to the small number of experiments). This suggests that interleaving is a reliable way to estimate the NDCG@5, MAP@10 and Precision@5 difference between pairs of rankers. Note that with the numbers of queries and impressions considered, the differences in all the interleaving experiments, and most of the judgment based evaluations, are statistically significant – despite the disagreements between metrics. The correlations corresponding to these plots are shown in the first three rows of Table 1.

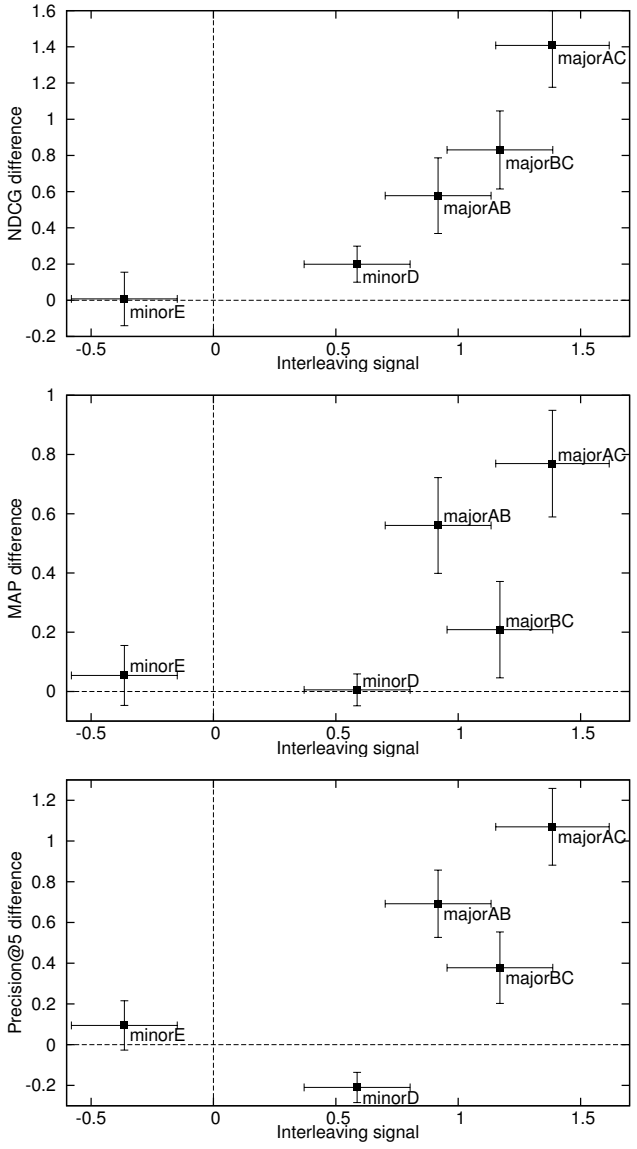

Figure 6: Correlation between IR metrics and interleaving experiments (corresponding to "per impression" row in Table 1).

7. INTERLEAVING DESIGN CHOICES

Thus far, we have used the team-draft interleaving method exactly as described by Radlinski et al. [10]. We now explore a number of possible variations of the analysis of interleaving.

7.1 Impression Aggregation

Interleaving credit assignment provides one "vote" to each impression, in effect allowing more frequent queries to contribute more to the outcome of an interleaving experiment. The alternative to is aggregate the preference by query: For each query, count how often each ranker is preferred, then aggregate per query and measure the fraction of *queries* for which each input ranker is preferred.

As shown in Table 1, this method provides a higher (although not statistically significantly so) correlation with all the judgment based metrics. This effect is surprising because the queries for evaluating the judgment-based metrics were sampled from a real workload, so we would expect interleaving to correlate more highly with the NDCG measured on this workload sample. We hypothesize this happens because the set of queries used for evaluating with standard IR metrics was sampled from the search workload a few years ago, thus has a different distribution than the current workload.

Table 2: Summary of interleaving and NDCG@5 evaluation for each experiment.

Experiment	NDCG@5			Interleaving		
	All	% qry	Δ	All	% imp	Δ
majorAC	1.41	82.3%	1.70	1.4%	40.3%	3.5%
majorBC	0.83	80.5%	1.03	1.2%	36.8%	3.2%
majorAB	0.58	78.8%	0.73	0.9%	38.5%	2.1%
minorD	0.20	16.0%	1.21	0.6%	7.1%	6.9%
minorE	0.01	63.5%	0.02	-0.4%	28.6%	-1.1%

7.2 Change frequency versus magnitude

In addition to providing a summary difference per experiment, both judgment based metrics and interleaving allow us to measure the fraction of queries for which the relevance of the two rankers differs, and the changes on just those queries. We present this analysis in Table 2. The left three columns show the mean NDCG@5 difference for each experiment (matching the NDCG@5 signal in Figure 6), as well as the average fraction of queries where NDCG@5 differs, and the mean NDCG@5 difference on just those queries.

To perform a similar analysis for interleaving, we must modify the credit assignment process. In Team Draft interleaving, each result is assigned to exactly one team, even if the rankers agree about the result order. If the input rankings are identical down to rank k, then the interleaved ranking will also share that top-k, and credit assignment is purely according to the coin toss. This is fair on average, over many queries, but is not informative.

In the modified credit assignment process, no credit is assigned to clicks in any such shared top-k. Lower clicks are treated as before. This does not change the mean interleaving signal (the shared results belong to each team equally often), but reduces the fraction of impressions that contribute to the outcome of the interleaving experiment. The last three columns of Table 2 show the mean interleaving signal, fraction of impressions where the click is on a non-shared result, and the mean signal from just these impressions.

We see that the fraction of queries where NDCG changes is much higher than the fraction of interleaving impressions where a click happens on a non-shared result. This is because changes in the relevance of any of the top 5 results (whether the user clicks on them or not) count as changes in NDCG@5, but do not count as changes in interleaving if the user only clicks on higher results. In fact, much of this difference is explained by navigational queries: When both rankers return the same top result, and users only click on that top result, any changes lower down are not considered meaningful by interleaving.

Second, note that the effect of experiment *minorD* becomes much clearer: a small fraction of queries/impressions changed, but the performance difference on these queries is large. The disagreement between NDCG@5 and interleaving on *minorE* persists, but whereas NDCG@5 seems to have changed only a very small amount on average, the signal on the impressions with changes in interleaving is now stronger.

7.3 Detecting Multiple Intents

Taking the analysis of interleaving impressions with changes further, we can look for queries with a particularly high or particularly low fraction of affected impressions and a signal far from 0% on those queries. Table 3 shows a sample of such queries from the *majorAC* experiment. Predominantly navigational queries that are answered well by both *rankerA* and *rankerC* have a low fraction of affected impressions. For example, for "facebook", 96% of impressions are followed by a click on the top result for both rankers, http://facebook.com/. The remaining impressions are followed by clicks on various results, the most commonly clicked one (usually presented around rank 5[3]) was a direct link to the facebook.com login page. For "usps", most users clicked on the top result, http://usps.com/. Of the 16% that did not, most clicked on the US Postal Service package tracking page. However, other queries saw big changes in ranking quality between rankerA and rankerC, resulting in almost all clicks being on non-shared results (although the rankers did sometimes share at least one top result, as the fraction of affected impressions is not 100%)[4].

Table 3: Sample queries from *majorAC* experiment.

Query	Impressions	Inter'l Signal	Affected impr. Fraction	Affected impr. Signal
facebook	5461	0.2%	4%	5.0%
myspace	1778	5.0%	12%	42%
usps	55	8.2%	16%	50%
cash for clunkers	58	36%	94%	39%
oprah denim makeovers	331	10%	97%	10%

Table 4: Effect of different credit assignment approaches on the consistency of interleaving outcome.

Impressions	Credit Assignment				
	constant	log(rank)	1/rank	top	bottom
1,000	70.2%	**72.2%**	66.5%	67.5%	71.1%
5,000	86.0%	**89.2%**	82.2%	82.5%	86.2%
10,000	91.3%	**93.5%**	88.6%	89.5%	92.3%
50,000	98.8%	**99.3%**	97.3%	97.9%	98.8%
100,000	99.8%	99.9%	99.0%	99.4%	99.7%
200,000	100%	100%	99.9%	100%	100%

7.4 Credit Assignment Alternatives

As a final analysis, we consider a credit assignment alternative where, unlike [17, 10], all clicks are not given an equal (constant) weight. This is motivated by the particularly strong bias web users have to click on top ranked search results. It may be the case that users who click on the top result are more likely to be clicking randomly than users who click further down the list. Alternatively, presenting the best result at the top of rankings is most important, so perhaps clicks at top positions should be weighted higher.

Table 4 shows the fraction of impression sample sets for which the ranking that was preferred overall was preferred on the sample (averaged across all five experiments). We compare the standard credit assignment (constant) with providing a score of $\log(rank)$ or $1/rank$ to each click before determining which input ranking is preferred. We also compare this to only considering the highest ranked click (top) or the lowest ranked click (bottom).

As in Figure 4, with more impressions interleaving is more consistent, and the choice of credit assignment has little effect. However, giving logarithmically more weight to lower

[3]The ranks changed as the web changed, and due to other instabilities inherent in web search result ranking.

[4]Note that if 100% of clicks on an interleaved ranking are on results from one of the input rankings, this translates to a signal of 50% in Table 3.

clicks improves consistency. In contrast, giving higher weight to higher clicks makes interleaving less consistent. The differences in bold are statistically significant improvements (with 95% confidence) over constant credit assignment.

Recently, in a completely independent study, Yue et al. [20] found that by learning a combination of related scoring alternatives, even larger improvements in sensitivity are possible.

8. CONCLUSION

In this paper, we have presented a detailed comparison between performance as measured by judgments-based information retrieval metrics and performance as measured by usage-based interleaving on five real pairs of web search ranking functions. We saw that performance measured by these methods is in agreement and, particularly in the case of NDCG@5 and interleaving, is highly correlated.

Using judgment-based metrics, we saw that realistic differences in ranking quality of about 1% by these metrics often were not reliably detected with queryset sizes below thousands of queries. We also saw that for some ranking improvements, particularly involving large changes to rare queries, it is possible for MAP measured on small query sets to disagree with MAP measured on large query sets. This suggests that small query set sizes may be impractical for measuring certain types of improvements in information retrieval research, or may even provide misleading results. NDCG appeared more reliable in this regard.

Evaluation with interleaving metrics was seen to require tens of thousands of user impressions to detect changes of this magnitude, with approximately 5,000 judged queries appearing to be similarly reliable to 50,000 user searches with clicks. Additionally, our results demonstrated that measuring the fraction of impressions where a click was made on non-shared results provides a better view of the changes in ranking quality than by identifying queries where NDCG changes: This separates changes to results which matter less to users from those that affect users more.

While 50,000 impressions per pair of rankers to evaluate may appear impractical for comparing tens of rankers, as are often evaluated during research, our results are consistent with interleaving outcomes being transitive (as was also seen by [10]), which we intend to investigate further in future work. In particular, if different rankers were interleaved with one or more standard baseline rankers, this would likely allow direct comparison between different ranking algorithms that were never compared directly.

Finally, we explored a number of alternative credit assignment modifications to interleaving. Our results suggested that placing more weight on lower clicks improves the consistency of the experimental outcome, thus making it more reliable with a small number of impressions.

Overall, we found a strong agreement between judgment-based and click-based evaluation, bolstering our confidence in both types of performance assessment. Moreover, our results show that the query volumes necessary to detect realistic changes in retrieval quality using interleaving require just tens to hundreds of regular search users, making them attainable in an academic environment.

9. ACKNOWLEDGMENTS

We would like to thank the developers of the Bing search engine, and in particular Rishi Agarwal, Nikunj Bhagat, Eric Hecht, Manish Malik and Sambavi Muthukrishnan for making this research possible.

10. REFERENCES

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

[2] Andrew Turpin and Falk Scholer. User Performance versus Precision Measures for Simple Search Tasks. In *Proc. of SIGIR*, 2006.

[3] Ben Carterette and Rosie Jones. Evaluating Search Engines by Modeling the Relationship Between Relevance and Clicks. In *Proc. of NIPS*, 2007.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 26(2):3–10, 2002.

[5] W. B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison Wesley, 2009.

[6] Dennis Fetterly, Mark Manasse, and Marc Najork. On The Evolution of Clusters of Near-Duplicate Web Pages. In *LA-WEB*, pages 37–45, 2003.

[7] Diane Kelly, Xin Fu, and Chirag Shah. Effects of rank and precision of search results on users' evaluations of system performance. Technical Report TR-2007-02., UNC SILS, 2007.

[8] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proc. of SIGIR*, 2002.

[9] Falk Scholer and Andrew Turpin. Metric and relevance mismatch in retrieval evaluation. In *Proc. of the Asia Information Retrieval Symposium*, 2009.

[10] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How Does Clickthrough Data Reflect Retrieval Quality. In *Proc. of CIKM*, 2008.

[11] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kraemer, L. Sacherek, and D. Olson. Do Batch and User Evaluations Give the Same Results? In *Proc. of SIGIR*, 2000.

[12] James Allan, Ben Carterette, and J. Lewis. When Will Information Retrieval be "Good Enough"? In *Proc. of SIGIR*, 2005.

[13] Mark Sanderson and Justin Zobel. Information Retrieval System Evaluation: Effort, Sensitivity and Reliability. In *Proc. of SIGIR*, 2005.

[14] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proc. of SIGIR*, 2008.

[15] Scott B. Huffman and Michael Hochster. How Well does Result Relevance Predict Session Satisfaction? In *Proc. of SIGIR*, 2007.

[16] P. Thomas and D. Hawking. Evaluation by comparing result sets in context. In *Proc. of CIKM*, 2006.

[17] Thorsten Joachims. Optimizing Search Engines Using Clickthrough Data. In *Proc. of KDD*, 2002.

[18] Text Retrieval Conference. http://trec.nist.gov/.

[19] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiments in Information Retrieval Evaluation*. MIT Press, 2005.

[20] Y. Yue, Y. Gao, O. Chapelle, Y. Zhang, and T. Joachims. Learning More Powerful Test Statistics for Click-Based Retrieval Evaluation. In *Proc. of SIGIR*, 2010.