Web Search Solved? All Result Rankings the Same?

Hugo Zaragoza
Yahoo! Research
Barcelona, Spain
hugoz@yahoo-inc.com

B. Barla Cambazoglu
Yahoo! Research
Barcelona, Spain
barla@yahoo-inc.com

Ricardo Baeza-Yates
Yahoo! Research
Barcelona, Spain
rbaeza@acm.org

ABSTRACT

The objective of this work is to derive quantitative statements about what fraction of web search queries issued to the state-of-the-art commercial search engines lead to excellent results or, on the contrary, poor results. To be able to make such statements in an automated way, we propose a new measure that is based on lower and upper bound analysis over the standard relevance measures. Moreover, we extend this measure to carry out comparisons between competing search engines by introducing the concept of disruptive sets, which we use to estimate the degree to which a search engine solves queries that are not solved by its competitors. We report empirical results on a large editorial evaluation of the three largest search engines in the US market.

Categories and Subject Descriptors

H.3.3 [**Information Storage Systems**]: Information Retrieval Systems

General Terms

Human Factors, Measurement

Keywords

Web search evaluation, editorial judgments, relevance

1. INTRODUCTION

Is web search solved? Is navigational search a commodity? Are all major web search engines roughly the same? Are tail queries the new battleground? In the past few years, there have been many qualitative statements about these and similar questions[1]. Despite lots of speculation, there are very few published research results on the topic [38]. We believe that this is mainly due to the ill-defined nature of such questions and the difficulties in the investigation process. Nevertheless, despite all the difficulties involved, we claim that it is still possible to conduct an analysis.

In this paper, to address the questions raised before, we propose a novel performance measure based on the idea of bounding standard IR performance measures. This measure enables making quantitative statements of the form "at least x% of queries lead to excellent results on search engine A".

Specifically, the contributions of the paper are as follows:

- We show that accurate estimation of query frequencies is crucial in aggregating relevance measurements as inaccurate estimations can cause a strong bias in obtained measurements. As a remedy, we propose a novel technique to correct sample query frequencies.
- We propose a new quality measure that is based on lower and upper bound analysis on the standard relevance measures in literature. We use this new measure to obtain absolute, quantitative statements about the result qualities of individual search engines.
- We extend the proposed measure by introducing the notion of disruptive sets[2]. We use this extended measure to compare competing search engines.

The proposed techniques are applied to a large, editorially judged query sample, obtained from the Yahoo! web search engine. The following are the findings of our work.

- In terms of query volume (i.e., the entire query traffic), most web search queries lead to satisfactory results on commercial search engines. At least 93% of the volume leads to excellent results on all search engines studied.
- Most navigational queries are solved by all search engines (at least 98% in terms of query volume and at least 81% in terms of unique queries). This statement also applies to frequent non-navigational queries.
- At least 23% of unique queries lead to poor results, but they constitute a very small fraction of the volume.
- The disruptive set size of each search engine is significantly large. Given any pair of search engines, about one-third of unique queries are well solved by one of the two search engines, but not by the other.

[1]As an example, Marissa Mayer (Google VP of search products and user experience) said in an interview in 2008: "Search is an unsolved problem. We have a good 90% to 95% of the solution, but there is a lot to go in the remaining 10%." (`http://latimesblogs.latimes.com/technology/2008/09/marissa-mayer-t.html`)

[2]We define the disruptive set of a search engine as the set of queries solved by that engine but not by its competitors.

CIKM'10, October 26–29, 2010, Toronto, Ontario, Canada.

These findings lead us to the conclusion that the only significant difference in search engines today is in the non-navigational tail, and in that different search engines are solving different regions of this tail. Many people had these intuitions before, but to the best of our knowledge, we provide the first empirical confirmation and quantification.

The rest of the paper is organized as follows. In Section 2, we provide the details of the query set used in the work. Section 3 discusses several caveats related to our data and the methodology followed. In Section 4, we propose a technique for frequency-based aggregation of judgments over a sample query set. The proposed performance measures are described in Section 5. Experimental results are presented in Section 6. In Section 7, we discuss some further issues related to our work. We provide a detailed survey of related literature in Section 8. The paper is concluded in Section 9.

2. QUERY SET

We sampled 1,000 queries[3] from the web search query traffic of Yahoo! in 2008. Sampling is done with replacement from the query traffic distribution. Therefore, popular queries had a much higher probability of being sampled, i.e., the same query might be sampled more than once, leading to several overlapping data points. The resulting frequency distribution in our sample set (without any query normalization) is as follows: 931 queries appear with frequency 1, 15 with frequency 2, 5 with frequency 3, 1 with frequencies 4, 5, and 15. An almost identical distribution is obtained if we normalize the queries by removing punctuation, lowercasing, and sorting query terms in alphabetical order.

Sample queries were issued in late 2008 to the three popular web search engines in the US market (Google, Microsoft Live Search, and Yahoo! Search), and the top 5 URLs returned by each search engine are recorded. A team of professional editors judged every recorded page against its query, assigning a five-graded relevance score[4] to each result: unrelated, related, relevant, very relevant, or perfect (denoted by `u`, `r`, `R`, `V`, and `P`, respectively). Along with this evaluation, the editors classified the queries as being navigational [10] or not. In evaluations, only algorithmic search results are considered, i.e., advertising, editorial shortcuts, web site grouping, and all other forms of shortcuts are ignored.

3. CAVEATS

In this section, we point at the difficulties faced in search quality evaluations and also note some caveats in our work.

Query sample size. Most research works use very small query sets (see Section 8). Moreover, sample queries are often selected by the authors themselves in an arbitrary fashion. Therefore, the sample queries typically represent only the popular or head queries. Since such queries are relatively easier to be answered by search engines, the quality estimations made by these works tend to be highly optimistic.

There are also some large-scale evaluations and average performance estimates made within the major search companies. Unfortunately, these cannot be used to answer the questions raised in this paper due to the following reasons. First, standard aggregate performance measures (e.g., the average NDCG measure [27]) cannot estimate in absolute terms how good or bad web search is. Second, since evaluations are conducted on a particular system and under specific conditions, they cannot be compared to the evaluations on other search engines. Third, evaluations tend to be biased since sample query frequencies are used for aggregation, without the correction proposed in this paper (see Section 4).

In our work, we use 1000 truly randomly sampled queries whose frequencies are corrected using a sample query log of 50 million queries. Hence, we expect to obtain a quite representative view of search engine behavior. Nevertheless, the question remains: can we represent the richness and complexity of today's web search queries by only 1000 queries[5]?

User sessions. A number of works pointed out that users tend to rewrite their queries several times before they reach their goal. This is especially true for difficult informational queries. For such queries, perhaps, the entire user sessions should be sampled and judged instead of individual queries. Unfortunately, detecting user sessions [21, 25] with sufficient accuracy is still an open research problem, and our query log does not have any information on sessions or query rewrites.

Human judges. Perhaps, editorial judgment is one of the main issues in relevance evaluation [3]. Since judges are not originators of queries, the intent or information need of users must be identified by the judges themselves. For certain queries, it is extremely difficult to identify the intent. Moreover, since the judges are not necessarily experts on many intents, it is difficult to accurately evaluate the quality of search results. To mitigate these effects, commercial search engine companies use highly trained full-time editors, who can achieve quite high agreement rates, as in our case.

Shallowness of evaluations. In this work, we consider only the top five results of each search engine, but it is possible that there are large differences between the engines below the fifth rank. It is known that, in web search, a user who does not see any relevant results in the top results is very likely to reformulate the query. Nevertheless, for very difficult queries, users are willing to check the results at lower ranks. In these cases, our estimations of user satisfaction are pessimistic since such queries are treated as unsolved.

Misspelled queries. All three search engines studied correct common misspellings automatically. Therefore, for misspelled queries, the returned results are related to the correct spelling (or to some assumed correct spelling) of the query. The editors in our study take this into account. If they believe that a query is misspelled, they judge the results with respect to the correct spelling of the query. Hence, it is possible that many queries that we treat as solved are, in fact, misspelled queries that, strictly speaking, would return poor or no results. Moreover, we should note that search engines are sometimes unable to correct the spelling, and they return poor or no results. In our sample, $2\% \pm 0.5\%$ of the queries return no results, and about two-third of them are clearly misspellings. Herein, we have treated all such queries as hard, but this is a debatable decision.

Relevance measure. Herein, because of two reasons, we use DCG [26] instead of NDCG [27] as our quality measure. First, DCG is shown to correlate better with user satisfaction than NDCG [1]. Second, we have very few judgments per query. Hence, for some queries, we may miss good results, obtaining only poor ones. NDCG can have fluctuation in such queries, rendering it not suitable to our purpose.

[3]Note that this number is quite large as editorial judgments are performed for all major search engines.

[4]Relevance is decided based on the web page content.

[5]According to the finding in [28], a query log with 650 queries is sufficient to reliably estimate significance.

4. AGGREGATE QUERY PERFORMANCE

In evaluating a search engine, one is typically interested in the expected performance, i.e., the average performance expected over a large set of queries. Unfortunately, estimating the expected performance based on the performance of individual queries is not trivial because query frequencies follow a power law distribution [44]. Taking this into account during query sampling is crucial and can have a dramatic impact on performance estimations. In this section, we discuss this problem and propose a performance aggregation technique, which yields a much better estimate of the expected performance. We also show that, without this technique, it is hard to make quantitative statements about the quality.

Query samples are weighted sets, in the sense that the same query can be sampled multiple times. Let $\mathcal{S}$ be a sample query set. For every $q \in \mathcal{S}$, we have a sample frequency $f_{\mathcal{S}}(q)$, i.e., the frequency of q in $\mathcal{S}$. Moreover, let us have

$$\widehat{P}_{\mathcal{S}}(q) = \frac{f_{\mathcal{S}}(q)}{\sum_{q' \in \mathcal{S}} f_{\mathcal{S}}(q')}, \tag{1}$$

as the relative frequency of q in $\mathcal{S}$. If $\mathcal{S}$ is an i.i.d. sample of a distribution of queries, $\widehat{P}_{\mathcal{S}}(q)$ is the empirical plug-in estimator of the prior probability $P(q)$ of q in the entire log. Also, let us assume that there is a discrete probability distribution Q^* of queries issued to the search engine. If we denote the relevance measurement[6] on q by $M(q)$, the expected performance of the search engine becomes

$$H(Q^*) = E_{Q^*}\{M(q) \colon q \in Q^*\} = \sum_{q \in Q^*} M(q) P_{Q^*}(q), \tag{2}$$

where $P_{Q^*}(q)$ is the true relative frequency of q. Unfortunately, (2) cannot be computed directly since it requires evaluating every query q and also knowing $P_{Q^*}(q)$, neither of which is feasible. Instead, it is standard to i.i.d. sample Q^* and obtain a relatively small sample query set $\mathcal{Q}^E$ (and sample frequencies). This set is then used to obtain a (maximum likelihood) empirical estimate of H. If the true relative query frequency distribution P_{Q^*} was known, we could approximate the true expected performance as

$$H(Q^*) \approx \widehat{H}(\mathcal{Q}^E) = \sum_{q \in \mathcal{Q}^E} M(q) P_{Q^*}(q). \tag{3}$$

In other words, we would sample as many queries as possible, and then use their true probability $P_{Q^*}(q)$ in aggregating. Unfortunately, we cannot evaluate (3) because $P_{Q^*}(q)$ is unknown. Therefore, we must estimate it from samples. To our knowledge, all previously published work use one of the following two approaches for this estimation.

`A.` Q^* is assumed to be uniform, and we compute

$$H(Q^*) \approx \widehat{H}(\mathcal{Q}^E) \approx \sum_{q \in \mathcal{Q}^E} \frac{M(q)}{|\mathcal{Q}^E|}, \tag{4}$$

which is implicitly obtained if repeated queries are removed from the evaluated sample, making $f_{\mathcal{Q}^E}(q) = 1$ for all q. This is a typical scenario in retrieval experiments and small search engine evaluations, where queries are artificially written by editors with no regard to their frequencies. We refer to this aggregation technique as `unique` since it can be naturally obtained from a sample where duplicates are removed.

[6]This can be any standard retrieval performance measure.

`B.` The plug-in estimator $\widehat{P}_{\mathcal{Q}_E}(q)$ is used to compute

$$H(Q^*) \approx \widehat{H}(\mathcal{Q}^E) \approx \sum_{q \in \mathcal{Q}^E} M(q) \widehat{P}_{\mathcal{Q}^E}(q). \tag{5}$$

This is the implicit approach taken by most large-scale search engine evaluations, where $\mathcal{Q}^E$ is obtained from a large i.i.d. sample of a very large query log (without uniquing). We refer to this technique as `sample` since it uses the query frequencies observed in the sample.

Although the above-mentioned approaches have their value, they are both problematic and considerably distort the true expected performance. The reason for this is that the frequency distribution is very far from uniform (as `A` assumes). In fact, it is well known that the query distribution follows a power law [44] (at least in web search). For this reason, plug-in estimators on small or medium size samples (as used in `B`) are also very inaccurate. Indeed, in power law distributions, the frequency plug-in estimator is not accurate unless the sample is extremely large, or the α coefficient is close to zero (making the distribution close to uniform).

In practice, the financial cost of evaluation prohibits search engines from having sufficiently large evaluated sets. This has the effect of making the plug-in estimator a very poor estimator, grossly underestimating the effect of head (frequent or popular) queries. We note, however, that frequency is independent of performance, and therefore we are not tied to the $\mathcal{Q}^E$ set to evaluate frequencies. Indeed, we can use any available large set (even if it is not evaluated) to obtain a better estimate of the relative frequencies. Since evaluated queries are obtained by sampling query logs, it is possible to obtain a second, much larger sample to estimate the query frequency. This leads to the option we propose:

`C.` We obtain a larger (non-evaluated) sample $\mathcal{Q}^U$ such that $|\mathcal{Q}^U| \gg |\mathcal{Q}^E|$. We can now use the plug-in estimator of this set, i.e., $\widehat{P}_{\mathcal{Q}^U}(q)$, which leads to

$$H(Q^*) \approx \widehat{H}(\mathcal{Q}^E) \approx \sum_{q \in \mathcal{Q}^E} M(q) \widehat{P}_{\mathcal{Q}^U}(q). \tag{6}$$

This is expected to be a more accurate estimator. We refer to this aggregation as `corrected` since it corrects the sample frequencies with better plug-in estimations.

To demonstrate the problem empirically, we use a query log of 50M queries and construct samples (i.i.d. with replacement) of sizes 1K, 100K, 10M, and 50M. In Fig. 1-a, for each sample, we plot query frequencies against ranks of queries. We observe that as the sample becomes smaller, its frequency range becomes flatter and flatter, approaching the uniform distribution. This demonstrates that the probabilities computed using small samples are far from real.

To show the impact of the proposed frequency correction approach, we perform the following experiment. We pick the queries in the 1K sample, and compute their frequencies separately in all four samples. This simulates the effect that, in reality, the query set is fixed in advance. Fig. 1-b shows the obtained frequencies, where the queries are sorted by their rank in the 50M query sample. We observe that increasing the size of the sample on which the frequencies are computed helps to recover the original power law distribution.

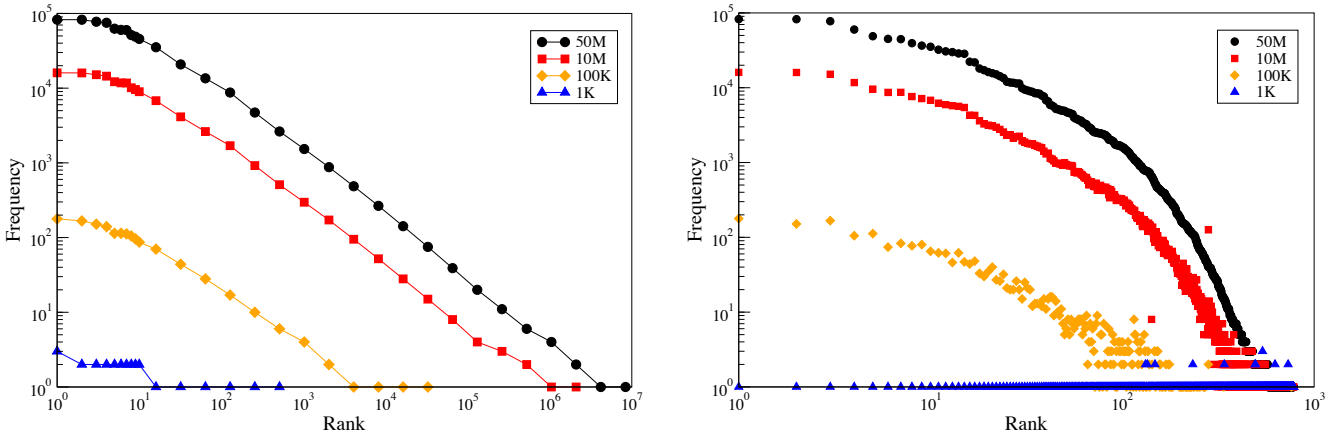

Figure 1: a) Query frequency distributions with varying sample sizes (the ranks after the tenth are sampled logarithmically. b) The effect of frequency correction.

Table 1: Set definitions

Set name	Definition
`solved`	$\mathcal{S}_{\text{solved}} = \{q \in \mathcal{Q}\colon M(q) \geq \delta_{\text{solved}}\}$
`hard`	$\mathcal{S}_{\text{hard}} = \{q \in \mathcal{Q}\colon M(q) \leq \delta_{\text{hard}}\}$
`two-engine solved`	$\mathcal{S}^2_{\text{solved}} = \{q \in \mathcal{Q}\colon M_{\text{I}}(q) \geq \delta_{\text{solved}} \wedge M_{\text{II}}(q) \geq \delta_{\text{solved}}\}$
`two-engine hard`	$\mathcal{S}^2_{\text{hard}} = \{q \in \mathcal{Q}\colon M_{\text{I}}(q) \leq \delta_{\text{hard}} \wedge M_{\text{II}}(q) \leq \delta_{\text{hard}}\}$
`tied`	$\mathcal{S}_{\text{tied}} = \{q \in \mathcal{Q} \cap \overline{(\mathcal{S}^2_{\text{hard}} \cup \mathcal{S}^2_{\text{solved}})}\colon \lvert M_{\text{I}}(q) - M_{\text{II}}(q)\rvert \leq \delta_{\text{tied}}\}$
`disruptive-I`	$\mathcal{S}_{\text{dis}-\text{I}} = \{q \in \mathcal{Q} \cap \overline{(\mathcal{S}^2_{\text{hard}} \cup \mathcal{S}^2_{\text{solved}} \cup \mathcal{S}^2_{\text{tied}})}\colon M_{\text{I}}(q) > M_{\text{II}}(q)\}$
`disruptive-II`	$\mathcal{S}_{\text{dis}-\text{II}} = \{q \in \mathcal{Q} \cap \overline{(\mathcal{S}^2_{\text{hard}} \cup \mathcal{S}^2_{\text{solved}} \cup \mathcal{S}^2_{\text{tied}})}\colon M_{\text{I}}(q) < M_{\text{II}}(q)\}$

5. BOUNDING WEB SEARCH QUALITY

In order to evaluate the quality of a search engine, most studies use averages over a relevance metric such as DCG. Typically, such retrieval performance measures are developed to evaluate the relative quality of a ranking function and are excellent devices for model comparison (i.e., choosing which ranking function is better in a pool) and model selection (i.e., tuning the parameters of a particular ranking function). However, these performance measures are not easy to interpret in absolute terms and are hard to relate to the kind of quantitative statements that we are after. For example, if a web search engine has an average DCG of 18.7, does this mean that the web search is solved by this engine?

Herein, we propose a measure that allows making lower and upper bound statements about the quality of a search engine that is evaluated by the standard retrieval performance measures. The proposed measure first constructs several query sets of interest (e.g., solved, hard, disruptive), using standard relevance metrics. The relative sizes of these interest sets are then used to make quantitative statements about search performance. Although this measure appears to be simple and intuitive, it is more interpretable, robust, and general than the standard relevance measures.

We now describe the proposed measure in more detail. Assume that we have evaluated a set of queries $q \in \mathcal{Q}$ by some standard performance measure $M(q)$. Also, assume that users are completely satisfied by the results when this measure is very high. More specifically, if the measure is higher than a fixed threshold δ_{solved}, the query is placed in the `solved` query set. Similarly, when the measure is lower than a fixed threshold δ_{hard}, users are completely unsatisfied and the query is placed in the `hard` query set. Given these `solved` and `hard` sets, the measure we propose obtains bounds on the number of solved and hard queries by simply computing the sizes of the respective sets. These bounds let us make quantitative statements of the kind "at least x% of queries are solved/hard with respect to this search engine" (see Section 6.2). The first two rows in Table 1 define the `solved` and `hard` sets. The idea is also illustrated in Fig. 2-a.

Next, we extend our measure to enable comparisons between a pair of search engines. The regions of interest used by the extended measure are defined in the last five rows of Table 1. We illustrate the process in Fig. 2-b, which shows a sample scatter plot of queries evaluated on engines I and II by some measure M. In the figure, by intersecting the two search engine's respective `solved` and `hard` sets, we obtain two corner regions. We call the region at the top right corner the `two-engine solved` set and the bottom left the `two-engine hard` set. These regions can be used to give bounds on the number of queries for which both search engines agree that they are solved (or hard). The top left region contains queries which are hard for engine II, but solved by engine I. We refer to this region as the `disruptive-I` set since, on queries in this region, engine I performs noticeably better than engine II. Hence, users who issue to search engine II a query that falls in this region may be compelled to switch to engine I. The symmetric region is similarly called the `disruptive-II` set. If we make no other assumptions about M, we can restrict the disruptive sets to these corners. However, we make the disruptive sets slightly larger by making

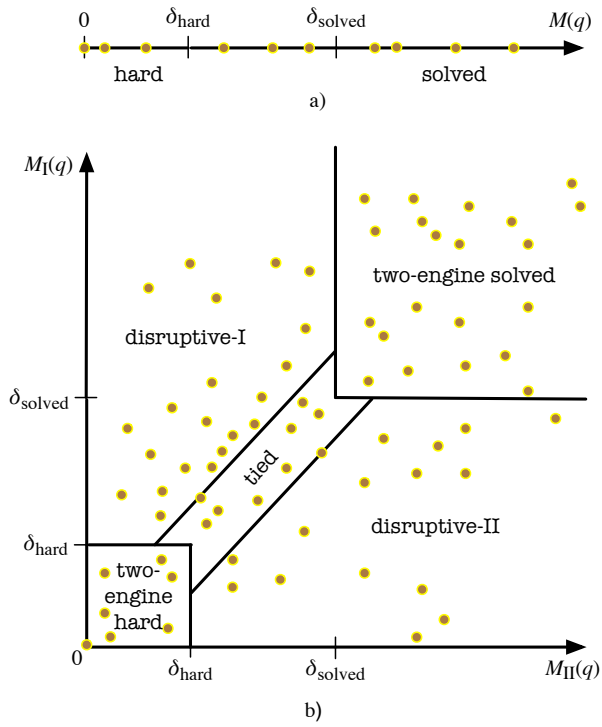

Figure 2: Generated query sets: a) for single-engine evaluation, b) for two-engine comparison.

a third assumption about M. Although we do not trust the relative values of M in the $(\delta_{solved}, \delta_{hard})$ interval, we assume that if M is much higher for engine I than for engine II (i.e., $M_I(q) - M_{II}(q) \geq \delta_{tied}$), then the user prefers engine I to II. This leads to the `tied` set, in which neither of the two search engines perform clearly better than the other.

6. EXPERIMENTAL RESULTS

6.1 Bound Selection

An important issue in our measures[7] is to accurately identify the three thresholds δ_{solved}, δ_{hard}, and δ_{tied}. Note that these three thresholds must be set to be able to specify the interest regions used by the measures. In our work, to identify the thresholds, we followed an ad-hoc approach. We analyzed the judgments of many queries with different DCG values, looking for reasonable thresholds. For the δ_{solved} threshold, we picked $\delta_{solved} = 9$ as a reasonable cut-off point: above this value all the queries we examined were clearly excellent, and the differences between search results above this value did not appear to be significant to us. We repeated this analysis looking for a DCG threshold under which a user would be upset with the results. We picked $\delta_{hard} = 2$ as a reasonable threshold. We note that these values are very conservative, but they are reasonable since the identified thresholds are used as lower bounds. Furthermore, the DCG distribution is linear around these values (see the discussion in Section 7). Hence, small changes in these thresholds may cause only small changes in the statistics obtained using them. Finally, to identify δ_{tied}, we determined at which DCG value the quality difference between the engine pairs becomes distinguishable. This threshold is much harder to set and is probably not constant[8], but its effect is small as it simply removes queries from consideration. Furthermore, as we will see in Section 7, the relative disruptive set sizes are stable with respect to the choice of this threshold. In our case, we set $\delta_{tied} = 1$, which has the effect of removing queries having DCGs within 2 points of each other.

6.2 Search Engine Evaluation

In this section, we analyze the aggregated performance of three search engines (Google, Microsoft Live Search, and Yahoo! Search)[9] by using the proposed quality measure (see Section 5) with `uniform`, `sample`, and `corrected` aggregation methods (see Section 4). The `solved` and `hard` set sizes, obtained by averaging the three engines' set sizes, are displayed in Table 2. Herein, we discuss only the `unique` and `corrected` aggregation methods since the `sample` aggregation method is always very close to `unique` and it is not informative anyway as it is too biased. In Table 2, when we look at unique queries (`unique`), we observe that a large fraction of queries remain unsolved for most search engines (23%), whereas less than half (47%) lead to excellent results on any engine. Moreover, we note that there is a high level of agreement between search engines on these numbers: all three evaluated search engines are within 3% of each other.

If we take into account the frequency of queries (`corrected`), however, this picture changes radically. In this case, we observe that there are almost no unsolved queries for any search engine ($.3\% \pm .1\%$), and almost all queries lead to excellent results for all search engines ($93\% \pm 1\%$). This is not very surprising as frequent head queries tend to be easier than infrequent tail queries. This is for several reasons, which are discussed in the next section.

In our opinion, the results obtained by both `unique` and `corrected` are interesting as they give different views of search engine quality. For a single user asking difficult queries, `unique` is a more interesting indicator of quality. Furthermore, it is more useful for comparing search engines as most search engines perform well in answering frequent queries. However, `corrected` depicts a better picture of the expected quality of a search engine in our daily lives. The results we obtained explain well why we depend on search engines so much: they work most of the time!

A comparison between the performance results obtained by `unique` and `corrected` shows that frequent queries are much better solved than infrequent queries. One explanation is the relative easiness of navigational queries [10], which are typically very frequent. Since our editors labeled navigational queries, we also apply our measure to only navigational and non-navigational query classes. According to Table 2, we observe that 30% of unique queries are navigational and they constitute 87% of the volume. Indeed, looking at the sizes of the `solved` set, we confirm that search

[7]In our experiments, we use DCG@5 as the standard measure. In computation of this measure, some works use linear gain functions with few levels (e.g., $\ell \in \{0,1,2\}, g(\ell) = \ell$ in [1]). Others give higher relative weight to excellent results (e.g., $\ell \in \{0,\ldots,4\}, g(\ell) = 2\ell - 1$ [49]). In our work, we use $g(\ell) \rightarrow \{0, .5, 3, 7, 10\}$ for $\ell \in \{0,\ldots,4\}$ and the standard logarithmic discount function $d(r) = 1/\log_2(1+r)$.

[8]It may be a function of the absolute DCG of the engines.

[9]In order to avoid a debate over which search engine is better, throughout this paper, we report the obtained results without identifying individual search engines. In all plots, we use letters A, B, and C (arbitrarily mapped to engines).

Table 2: Single engine comparison (all values are percentages)

	Query class								
	All			Navigational			Non-navigational		
Type	`unique`	`sample`	`corrected`	`unique`	`sample`	`corrected`	`unique`	`sample`	`corrected`
Query class	100	100	100	30	33	87	70	67	13
`solved`	47 ± 2	49 ± 2	93 ± 1	$81 \pm .2$	$83 \pm .4$	98 ± 1	33 ± 3	33 ± 3	65 ± 4
`hard`	23 ± 3	22 ± 2	$.3 \pm .1$	$9 \pm .4$	$8 \pm .3$	$.2 \pm 0$	29 ± 4	29 ± 4	$1 \pm .6$

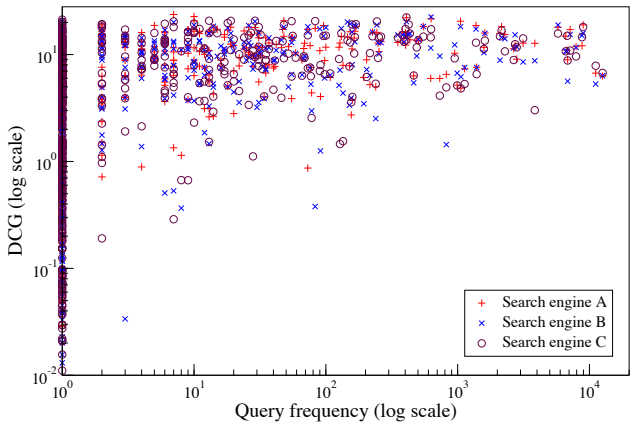

Figure 3: The log-log scatter plot showing the frequencies of queries against their DCG values.

engines excel in solving navigational queries. We observe that 81% and 98% of such queries lead to excellent results in all search engines for `unique` and `corrected`, respectively. This implies that search engines succeed, to a large degree, in solving navigational queries, not only the popular ones. This is also confirmed by the fact that the `hard` set sizes are 9% and .2% for `unique` and `corrected`, respectively. Moreover, we note that the small spread (.4% and 0%, respectively) indicates that all three search engines have close performance. In conclusion, we can state that navigational search is mostly solved. It remains to investigate what is not solved in the remaining 9% (`unique`) of navigational queries.

Our findings are somewhat contrary for non-navigational queries. We first note that there are many more unique non-navigational queries (70%) than navigational queries. However, they constitute only 13% of the query volume (`corrected`). The performance on non-navigational queries is much worse than that on navigational queries. With `unique`, only 33% of non-navigational queries lead to excellent results, and more than 29% lead to poor results. We note that there is higher search engine variance here: the spread of these measurements is 3% and 4%, respectively, meaning that the performance difference between search engines is larger for non-navigational queries. With `corrected`, we observe that the `solved` set size is much higher (65%), but is still far from the performance on navigational queries. More interestingly, the `hard` set size with `corrected` is 29 times smaller than with `unique`, indicating that frequent non-navigational queries are much better solved by search engines than less frequent ones. This challenges the previous explanation that search engines work because they have solved navigational queries: they work because they have solved all frequent queries, whether they are navigational or informational! This is further demonstrated in Fig. 3, where we show the DCG scatter plot for non-navigational queries. According to the figure, only low frequency queries lead to poor results. Again, the spread between search engines is high (4%), indicating that some engines can cope with non-navigational queries significantly better than the others. This is clearly where the web search engine innovation battles are being played out in the relevance field.

6.3 Pairwise Search Engine Comparison

In this section, we compare the performance of a pair of search engines on a query per query basis. This allows us to investigate the question of whether search engines achieve similar result qualities for the same queries, or on the contrary, some engines are better than the others for certain queries. To be able to perform this analysis, we employ the extended measure proposed in Section 5.

This analysis is interesting for a number of reasons. First, we are interested in knowing whether the `hard` set of engines is fundamentally hard, i.e., all search engines perform poorly on the queries in this set, or instead, there are queries that are hard for one search engine but not for another. Second, we are interested in verifying whether the `solved` set is not only equal in size for all search engines (as shown in the previous section), but is also composed of the same queries for all search engines. Finally, we want to see if there are important differences between search engines.

Note that we are not interested here in trying to decide which of today's commercial search engines is better. Instead, we want to derive quantitative statements about the similarity or dissimilarity of search engines with respect to result quality. As in the previous section, in order to avoid a controversy generated by any direct comparison of two commercial search engines, we report only the averages computed over three possible search engine pairs and the spreads. The `two-engine solved`, `two-engine hard`, and `tied` sets are symmetric. However, the `disruptive-I` and `disruptive-II` sets are not symmetric. Hence, in pair-wise comparisons, we order the search engines such that the size of `disruptive-I` is always larger than that of `disruptive-II`, i.e., the first one is always the better search engine.

Table 3 summarizes aggregate results over all queries. We observe that the size of the `two-engine solved` set is 40% with `unique` (91% with `corrected`) and has a small spread. This is not far from the 47% (`unique`) and 93% (`corrected`) values observed in single-engine evaluation (see Table 2 and the discussion in Section 6.2). This shows that queries solved by one engine tend to be solved also by the other two, i.e., all three engines perform perfectly on similar queries (at least 40% with `unique` and 91% with `corrected`). Similarly, there is a high overlap in the `hard` set. At least 17% (`unique`) of queries lead to poor results in all three engines. Since the single-engine `hard` set size was 23%, this constitutes an overlap of poor results in search engines of at least 74%. However, these queries have very low frequencies. Hence, the `two-engine hard` set size is only .2% with `corrected`. Given

Table 3: Two-engine comparison (all values are percentages)

	Query class								
	All			Navigational			Non-navigational		
Type	`unique`	`sample`	`corrected`	`unique`	`sample`	`corrected`	`unique`	`sample`	`corrected`
`two-engine solved`	40 ± 2	43 ± 2	91 ± 1	75 ± 1	78 ± 1	96 ± 2	26 ± 2	26 ± 2	61 ± 2
`two-engine hard`	17 ± 2	16 ± 2	$.2 \pm 0$	$7 \pm .4$	6 ± 0	$.2 \pm 0$	21 ± 3	21 ± 3	$.1 \pm 0$
`tied`	8 ± 1	8 ± 1	3 ± 1	$2 \pm .4$	2 ± 0	$.2 \pm .2$	11 ± 1	11 ± 1	17 ± 10
`disruptive-I`	21 ± 3	20 ± 3	$4 \pm .2$	9 ± 1	8 ± 1	2 ± 2	26 ± 4	26 ± 4	14 ± 9
`disruptive-II`	13 ± 1	13 ± 1	2 ± 1	8 ± 1	7 ± 1	2 ± 1	16 ± 2	16 ± 2	8 ± 11

that these are only lower bounds, we can only conclude that there is an important body of queries (at least 17% with `unique`) that are not solved by any of the three engines.

Herein, we also try to identify any other significant differences between search engines. According to Table 3, the size of `disruptive-I` is quite large (21% with `unique` and 4% with `corrected`), larger than the `two-engine hard` set. We can therefore claim that there is a large number of unique queries that are significantly better solved by the superior search engine. Interestingly, the size of `disruptive-II` is also large (13% with `unique` and 2% with `corrected`). If there was a search engine that solved all the queries that the other engines could not solve, then we would expect at least one disruptive set size to be close to zero, and hence the spread should be larger than average. However, this is not the case. This leads to the conclusion that there is always a significant number of queries that are better answered by one search engine than by the other, for all three search engines considered. Furthermore, these queries have non-negligible frequencies. Taking this further, we can conclude that there are significant differences in the three search engines in terms of their ranking functions or their crawls. These differences are specially notable for low-frequency queries and non-navigational queries, but we can see in Table 3 that also minor differences exist for navigational queries.

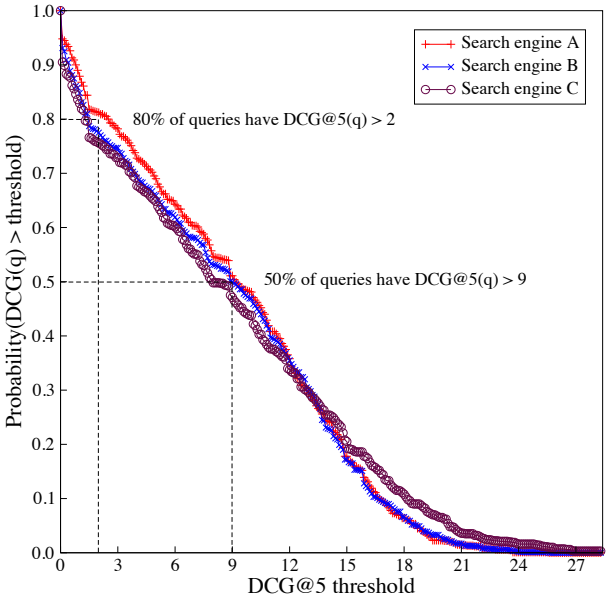

Figure 4: Cumulative probability distribution of DCG(q), reversed to emphasize which % of queries have at least some DCG value.

7. DISCUSSION

Inadequacy of standard measures. We already mentioned that correlating standard measures with absolute search quality is very difficult. It is even more difficult to interpret relative differences between search engines when they obtain similar results on these measures. We explain the problem by some examples. A ranking of the form `Puuuu` (perfect followed by unrelated) has a DCG@5 of 10, whereas `uPuuu` only 6.3. The following rankings have DCG@5 close to 8.5: `RVRuu`, `uVRRR`, `uPrur`. But how good (or bad) are these results in absolute terms? And why do we as users perceive some engines as being better than the others?

DCG values depend on a gain constant and a discount function, accumulated over all ranks. This makes it very hard to make absolute quantitative statements about search quality by using DCG results. Fig. 4 shows the cumulative probability distribution obtained in our evaluation[10] with respect to DCG@5. In Fig. 4, we observe that 50% of queries have a DCG@5 close to 9 or higher, which are excellent results. But, we note that the values are not concentrated around their mean at all. The DCG@5 increases almost linearly from .1 to 16, indicating that almost all values are equally common. We can draw two conclusions based on this observation. First, because of the shape of the DCG distribution, the mean is not prototypical, i.e., queries have low likelihood of being near it. Second, if we consider that the mean is close to our notion of an excellent result (or a solved query), we see that DCG does not have the right resolution, i.e., it over emphasizes relative differences in very good results (going from 9 to 27) while underemphasizing differences in poor results. This is an indication that we should be more interested about the logarithmic DCG, or in other words, we should pay more attention to the geometric mean rather than the arithmetic mean[11].

Stability of set sizes. An important issue in our analysis is the sensitivity of the obtained set sizes to selection of thresholds. In particular, we are most worried about the stability of results with respect to $|\mathcal{S}_{\text{tied}}|$ since the `tied` set corresponds to a dense region of the scatter plot and small variations in δ_{tied} may have a strong impact on the results. However, as we will now illustrate, this is not the case in practice. In Fig. 5, we display the ratio of the disruptive sizes of engines (i.e., $|\mathcal{S}_{\text{dis}-\text{I}}|/|\mathcal{S}_{\text{dis}-\text{II}}|$) for increasing values of δ_{tied}. As usual, we show both the average (for the three pair-wise comparisons that are possible) and the $\pm$ region in which the three averages lie. We observe that the computed

[10]Average DCG@5 range (`unique`) of all engines is $8.5 \pm .3$.

[11]When we computed the geometric mean, we indeed saw slightly larger differences between the three search engines, but they were still ranked in the same order.

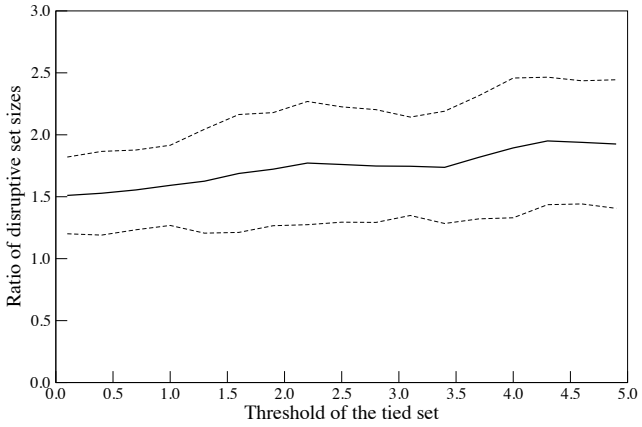

Figure 5: Stability of disruptive set sizes.

ratio is quite stable, increasing very slowly between 1.5 and 2, regardless of the δ_{tied} threshold chosen. Similar curves can be drawn for δ_{hard} and δ_{solved} with similar results. These are less interesting since these two thresholds are more robust and cover less dense areas of the scatter plot.

Generality of thresholds. We believe that it is possible to find reasonable thresholds for most measures commonly used in retrieval evaluation. In order to verify this, we tried a measure different from DCG and observed what the resulting `solved` and `hard` set sizes are. We need to note that it is quite hard to find an alternative to DCG since the data is multi-valued and only the top 5 documents are evaluated. Furthermore, when binarizing relevance, measures like precision and MRR on the top 5 results are very coarse, having only 5 possible values. Nevertheless, we found that, using reasonable threshold values for these measures, we can obtain `solved` and `hard` set sizes similar to those obtained with DCG. As an example, we replaced the previously used DCG@5 metric with P@5 by mapping perfect, very relevant and relevant labels to relevant, and the rest to irrelevant. With these binary labels, when we set $\delta_{\text{solved}} = 0.5$ and $\delta_{\text{hard}} = 0$, the `solved` and `hard` set sizes we obtained are $48\% \pm 2\%$ and $21\% \pm 3\%$ (`unique`), respectively. Both values are within 2% of the values obtained using DCG.

Other factors effecting results. Besides query frequency, there are other factors that have an impact on our results. One factor we investigated, although we do not report any results, is the query length. We found that shorter queries are indeed much better solved than longer ones. Similarly, performance of the engines varied depending on query classes (e.g., person, organization, place queries). We should note that human factors, which are outside the scope of this study, may also have a strong effect on the results. The first factor that comes to mind is human adaptability [46]: as users learn to use search engines, the intuition is that they stop asking queries that they do not think the engine will be able to solve. This promotes a rich gets richer behavior, where frequent queries by definition obtain better results.

Prioritizing investment. Commercial web search engines make significant investments[12] to improve the quality of their search results. Based on the proposed quality measure, an investment can be made on improving the queries in one or more of the five sets identified (i.e., `two-engine solved`, `two-engine hard`, `tied`, `disruptive-I`, `disruptive-II`)[13]. It is important to prioritize these sets for investment so that the highest search quality increase can be achieved with the least amount of investment. One option is to invest on the queries in the `two-engine solved` set as they are relatively easy queries. However, these queries are already well-solved, and hence it may be difficult to improve them further. Even if they can be improved, users may not be able to notice any quality difference. Another option is to improve the queries in the `disruptive-I` set as some of them may have room for improvement. However, it is perhaps better not to prioritize them for investment since they do not form an immediate threat, as the rival search engine performs relatively poor on these queries. A third option is to try to optimize the queries in the `two-engine hard` set. If these queries are improved, the benefit can be significant. However, these queries are really hard as indicated by the fact that the rival search engine so far could not solve them either. Therefore, investment on such queries is risky and may not be cost-effective. In our opinion, the primary target for investment should be the `disruptive-II` set (followed by `tied`). The queries in this set have room for improvement as the rival search engine is known to have them solved. Moreover, solving these queries prevents user switches to the rival search engine. In summary, the following prioritization of sets may be the most feasible: `disruptive-II` > `tied` > `two-engine hard` > `disruptive-I` > `two-engine solved`. Ideally, the variation in set sizes needs to be followed in time, and necessary actions should be taken accordingly. For example, the `hard` set should be targeted only after the `disruptive-I` and `tied` set sizes are sufficiently small.

8. RELATED WORK

Evaluation measures. There are two lines of research in search evaluation measures. The first line of research investigates the measures for relevance of search results. Despite the wide range of proposals (e.g., P-R [26], ESL [17], RHL [9], RR [9], ASL [37], SR [42]), only a few measures are commonly used in traditional IR evaluations (e.g., precision [52], recall [52], and MAP [11]). In case of availability of graded relevance judgments, which is mainly the case in internal search engine evaluations, CG [26], DCG [26], and NDCG [27] are the preferred measures. The latter two measures take into account the ranks of documents. More recent measures such as RBP [39] and ERR [14] also consider the information gained by the user after viewing each rank.

The second line of research aims to develop measures to compute the distance between two given rankings. Two widely used measures are Spearman's footrule [47] and Kendall-Tau [29] measures, which compute the distance between two full rankings. These measures are later extended to distance computation between partial rankings [19] and top-k lists [20]. More recent works adapt these measures to handle graded judgments [31] and incorporate a DCG-like decay with increasing rank [13, 31, 54].

User studies. There have been numerous user studies for evaluating and comparing result qualities of search engines. Most of these works concentrate on evaluating the relevance of search engine results that are judged by humans, according to an average performance measure [5, 16, 18, 22, 23,

[12]In this context, investment may refer to financial costs, time spent, or the amount of human resources allocated.

[13]During the discussion, we assume that `disruptive-II` is the disruptive set of a rival search engine company.

32, 34, 43, 51]. Unfortunately, most of these studies are very small-scale (the largest has 100 queries [34]) and are obsolete due to the fast-changing nature of search engines. Therefore, it is hard to draw useful conclusions from them.

There are a few user studies in which search engines are evaluated explicitly for user satisfaction [1, 7, 36]. In [36], a user study (1000 queries) is carried out with students. It is found that Google is significantly better than Yahoo! and MSN Live at navigational queries and slightly better at informational queries. It is noted, however, that the gap started closing in 2007. The results we report are worse than those reported in [36] (perhaps, due to the fact that we are looking at lower bounds), but we confirm that the gap has indeed closed. In [1], a methodology is proposed to measure user satisfaction by observing user's actions after the results are presented. In that study, user actions (e.g., copy-pasting, printing, saving, emailing) are used to estimate the level of importance of a result for the user. A user study is conducted in [1] to investigate the relationship between user satisfaction and retrieval effectiveness measures. This study showed that user satisfaction correlates better with the CG and DCG measures [26] than the NDCG measure [27].

Automated evaluation. Several studies investigated how search engines can be evaluated in the absence of relevance judgments [12, 15, 35, 41, 45]. A thread of research papers [35, 41, 45] suggested downloading the content pointed by the search results and assigning relevance scores via standard query-document similarity computations. Obtained similarity values are then used in estimating relevance of documents and evaluating search engines. In another thread [15, 41], some pseudo-relevant documents are automatically identified and used in evaluation. In [15], pseudo-relevant documents are identified by matching queries with the documents in the open directory project. Search engines are then evaluated by their ability to return these documents in high ranks. In [41], multiple search engine rankings are merged via rank aggregation and a certain fraction of top ranking documents are assumed to be relevant.

Other evaluation criteria. Besides relevance, there are a number of works that compare search engines across other criteria. Some of these works evaluate content-related issues, such as web coverage and speed in indexing newly published content [30], index freshness [33], index size and overlap [8], consistency in hit count estimates [50], and bias in web coverage [53]. Other works evaluate issues related to presented results, such as result page structure [24], change of results in time [6], uniqueness of results [48], result similarity [4], bias in results [40], and the coverage of domain names [50]. These studies are interesting as they highlight the fact that the search engine quality does not depend only on the ranking of results, but rather on many interrelated factors. In this work, we concentrated only on the result quality.

Search difficulty. We are aware of two works that ask questions similar to ours [2, 38]. In [2], a user study is conducted to understand the relationship between effectiveness of users in achieving a task and the accuracy of the retrieval system. The experiments indicate that there is no significant difference in utility unless the search accuracy is improved by a large amount. In [38], difficulty of search is estimated by measuring the entropy of the query logs. This study concludes that users often find the documents they are looking for in the top 10 results of search engines. However, the study does not provide a quantitative analysis as we do.

9. CONCLUSIONS

Throughout the paper, we pointed out many caveats and open issues in search quality evaluations. We demonstrated the important problem of query frequency aggregation and proposed a technique to correct it. In our opinion, this can have an important impact in many future evaluation studies. We developed novel measures to make quantitative statements on lower and upper quality bounds of web search engine rankings. We tested the proposed measures on a large, real-life, professionally evaluated web search query sample. We provided bounds on what fraction of queries are solved or hard. Moreover, we showed that all three major search engines solve navigational and frequent non-navigational queries, but there are differences in how they treat infrequent non-navigational queries. Interestingly, each engine has a non-negligible disruptive set, on which it performs significantly better than the other two engines. We hope to extend this work to different query classes and connect it to related research areas (e.g., user session and click analyses).

10. REFERENCES

[1] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *Proc. 30th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 773–774, 2007.

[2] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In *Proc. 28th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 433–440, 2005.

[3] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proc. 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 667–674, 2008.

[4] J. Bar-Ilan. Comparing rankings of search results on the Web. *Information Processing & Management*, 41(6):1511–1519, 2005.

[5] J. Bar-Ilan, K. Keenoy, E. Yaari, and M. Levene. User rankings of search engine results. *Journal of the American Society for Information Science and Technology*, 58(9):1254–1266, 2007.

[6] J. Bar-Ilan, M. Mat-Hassan, and M. Levene. Methods for comparing rankings of search engine results. *Computer Networks*, 50(10):1448–1463, 2006.

[7] M. M. Beg. A subjective measure of web search quality. *Information Sciences*, 169(3-4):365–381, 2005.

[8] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems*, 30(1-7):379–388, 1998.

[9] P. Borlund and P. Ingwersen. Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In *Proc. 21st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 324–331, 1998.

[10] A. Broder. A taxonomy of web search. *ACM SIGIR Forum*, 36(2):3–10, 2002.

[11] C. Buckley and E. Voorhees. *Retrieval system evaluation. In TREC: Experiment and Evaluation in Information Retrieval*, chapter 3, pages 53–75. The MIT Press, 2005.

[12] F. Can, R. Nuray, and A. B. Sevdik. Automatic performance evaluation of web search engines. *Information Processing & Management*, 40(3):495–514, 2004.

[13] B. Carterette. On rank correlation and the distance between rankings. In *Proc. 32nd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 436–443, 2009.

[14] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc. 18th ACM Conf. on Information and Knowledge Management*, pages 621–630, 2009.

[15] A. Chowdhury and I. Soboroff. Automatic evaluation of World Wide Web search services. In *Proc. 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 421–422, 2002.

[16] H. Chu and M. Rosenthal. Search engines for the World Wide Web: a comparative study and evaluation methodology. In *Proc. 59th Annual Meeting of the American Society for Information Science*, pages 127–135, 1996.

[17] W. Cooper. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19(1):30–41, 1968.

[18] W. Ding and G. Marchionini. A comparative study of web search service performance. In *Proc. 59th Annual Meeting of the American Society for Information Science*, pages 136–142, 1996.

[19] R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, 20(3):628–648, 2006.

[20] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. In *Proc. 14th Annual ACM-SIAM Symp. on Discrete Algorithms*, pages 28–36, 2003.

[21] D. Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179(12):1822–1843, 2009.

[22] M. Gordon and P. Pathak. Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing & Management*, 35(2):141–180, 1999.

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

[24] N. Hochstotter and D. Lewandowski. What users see – structures in search engine results pages. *Information Sciences*, 179(12):1796–1812, 2009.

[25] B. J. Jansen, A. Spink, C. Blakely, and S. Koshman. Defining a session on web search engines. *Journal of the American Society for Information Science and Technology*, 58(6):862–871, 2007.

[26] K. Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In *Proc. 23rd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 41–48, 2000.

[27] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[28] E. C. Jensen, S. M. Beitzel, O. Frieder, and A. Chowdhury. A framework for determining necessary query set sizes to evaluate web search effectiveness. In *Special Interest Tracks and Posters of the 14th Int'l Conf. on World Wide Web*, pages 1176–1177, 2005.

[29] M. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[30] Y. S. Kim and B. H. Kang. Coverage and timeliness analysis of search engines with webpage monitoring results. In *Proc. 8th Int'l Conf. on Web Information Systems Engineering*, pages 361–372, 2007.

[31] R. Kumar and S. Vassilvitskii. Generalized distances between rankings. In *Proc. 19th Int'l Conf. on World Wide Web*, pages 571–580, 2010.

[32] D. Lewandowski. The retrieval effectiveness of web search engines: considering results descriptions. *Journal of Documentation*, 64(6):915–937, 2008.

[33] D. Lewandowski. A three-year study on the freshness of web search engine databases. *Journal of Information Science*, 34(6):817–831, 2008.

[34] D. Lewandowski. The retrieval effectiveness of search engines on navigational queries. *ASLIB Proceedings*, 62, 2010. In press.

[35] L. Li and Y. Shang. A new method for automatic performance comparison of search engines. *World Wide Web*, 3(4):241–247, 2000.

[36] B. Liu. Personal evaluations of search engines: Google, Yahoo! and Live (MSN). `http://cs.uic.edu/~liub/searchEval/2006-2007.html`, 2007.

[37] R. Losee. *Text Retrieval and Filtering: Analytic Models of Performance*. Kluwer Academic, 1998.

[38] Q. Mei and K. Church. Entropy of search logs: how hard is search? With personalization? With backoff? In *Proc. Int'l Conf. on Web Search and Data Mining*, pages 45–54, 2008.

[39] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):1–27, 2008.

[40] A. Mowshowitz and A. Kawaguchi. Measuring search engine bias. *Information Processing & Management*, 41(5):1193–1205, 2005.

[41] R. Nuray and F. Can. Automatic ranking of information retrieval systems using data fusion. *Information Processing & Management*, 42(3):595–614, 2006.

[42] S. Pollack. Measures for the comparison of information retrieval systems. *American Documentation*, 19(4):387–397, 1968.

[43] H. Qin and P. P. Rau. Relevance measurement on Chinese search results. In *Proc. 12th Int'l Conf. on Human-Computer Interaction*, pages 981–988, 2007.

[44] C. Saraiva, E. de Moura, N. Ziviani, W. Meira, R. Fonseca, and B. Ribeiro-Neto. Rank-preserving two-level caching for scalable search engines. In *Proc. 24th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 51–58, 2001.

[45] Y. Shang and L. Li. Precision evaluation of search engines. *World Wide Web*, 5(2):159–173, 2002.

[46] C. Smith and P. Kantor. User adaptation: good results from poor systems. In *Proc. 31th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 147–154, 2008.

[47] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[48] A. Spink, B. J. Jansen, C. Blakely, and S. Koshman. A study of results overlap and uniqueness among major web search engines. *Information Processing & Management*, 42(5):1379–1391, 2006.

[49] M. Taylor, H. Zaragoza, N. Craswell, S. Robertson, and C. Burges. Optimisation methods for ranking functions with multiple parameters. In *Proc. 15th ACM Conf. on Information and Knowledge Management*, pages 585–593, 2006.

[50] M. Thelwall. Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology*, 59(11):1702–1710, 2008.

[51] S. Tongchim, V. Sornlertlamvanich, and H. Isahara. Improving search performance: a lesson learned from evaluating search engines using Thai queries. *IEICE Transactions on Information and Systems*, E90-D(10):1557–1564, 2007.

[52] C. J. van Rijsbergen. *Information Retrieval, 2nd ed.* Butterworths, 1979.

[53] L. Vaughan and M. Thelwall. Search engine coverage bias: evidence and possible causes. *Information Processing & Management*, 40(4):693–707, 2004.

[54] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proc. 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 587–594, 2008.