

Exploring Cost-Effective Approaches to Human Evaluation of Search Engine Relevance

Kamal Ali, Chi-Chao Chang, and Yun-Fang Juan

Yahoo Search, 701 First Avenue, Sunnyvale, CA 94089, USA.
{kamal, chichao, yunfang} @yahoo-inc.com

Abstract. In this paper, we examine novel and less expensive methods for search engine evaluation that do not rely on document relevance judgments. These methods, described within a proposed framework, are motivated by the increasing focus on search results presentation, by the growing diversity of documents and content sources, and by the need to measure effectiveness relative to other search engines. Correlation analysis of the data obtained from actual tests using a subset of the methods in the framework suggest that these methods measure different aspects of the search engine. In practice, we argue that the selection of the test method is a tradeoff between measurement intent and cost.

1 Introduction

In classical IR, the most common measures of the retrieval engine are based on human judgments of document relevance – is the document relevant to the query? The predominant methodology – the Cranfield [1] technique – compares IR systems over a set of topics, a set of documents for each topic, and a set of relevance judgments for each document. Researchers and business intelligence groups have adapted the Cranfield method to evaluate search engines (for example, topics become queries) with some success.

The task of a search engine is to accept a query and return a ranked list of references to documents that are relevant for that query to the user issuing the query. For an overall evaluation, one needs a representative sample of queries, which usually yields a representative sample of users across a representative set of documents. In addition, human judgments are idiosyncratic and vary depending on the judges. These are fundamental factors that were already present in classical IR evaluation studies and have been addressed extensively in the literature [4].

The main difficulty with the Cranfield-like approach is the cost of obtaining reliable and complete judgments. The average TREC collection contains about 800,000 documents spanning across 50 topics. Voorhes [4] estimates that nine person months are required to fully cover these documents. Pooling techniques—judging a subset of the documents rather than the entire set [5, 6]—and recent efforts in

formulating robust metrics in the presence of incomplete judgments [5] and in term-based evaluation [2] are reasonable attempts to mitigate the cost factor.

Besides the cost of collecting document relevance judgments, we are motivated by the following factors:

- **Advent of Domain-Specific Search Engines.** Domain-specific engines such as shopping, news and particularly image search pose specific evaluation challenges.
- **Improved Document Summarization.** Search engines typically return document abstracts that do a fairly good job of summarizing the underlying document. This opens the possibility of judging these summaries in lieu of judging the full document. We explore the relationship between document (landing-page) relevance and abstract relevance.
- **Diversity of Query Intent and Content.** The information needs of search engine users range from navigational, that is, users rely on search engines as a trampoline to specific documents and sites, to single-answer queries (simple question-answer sessions) such as “What is the capital of Afghanistan?” to research, that is, users searching for a set of documents for browsing. For many of these “types” of queries, it is unclear if the thorough perusal of the documents retrieved is preferred to simply looking for the right document or answer in the set of results. And more often than not, this can be accomplished by simply evaluating titles, abstracts, and URLs displayed.
- **Business Intelligence and Competitive Metrics.** Search engine evaluation serves two purposes: to improve the quality and relevance of the results as well as to gather data and metrics to understand the competitive landscape. Relative judgments, directly comparing two sets of results, can be more a more cost-effective approach than judging two sets of results separately.
- **Search Engine User Interface Features.** Engines such as Yahoo! and Google offer a multitude of features along with the search results, such as advertisements, spelling and related search suggestions, and opportunities for personalization and customization. We believe that the overall quality of the search experience is not the sum of its parts—a more holistic approach is needed.

In this paper, we detail our experiences with novel testing methodologies arranged along three axes of a methodology framework:

1. Judging the document summary (i.e. title, abstracts) versus the actual document. Perceived relevance versus ‘actual’ (landing-page) relevance.
2. Judging sets of results rather than each result individually. (We will use the terms ‘item’ and ‘result’ interchangeably.)

3. Judging relative relevance (between two search engines) rather than absolute relevance.

We experimented with these methodologies in a practical setting, evaluating several domain-specific search engines, namely image search, news search, ads search, as well as web search. The results show that there is no “silver-bullet” methodology—correlation measures between two different evaluation methodologies for a given domain are not high—which means that each methodology is sensitive to certain aspects of relevance that others are oblivious to.

2. Related Work

In [10], Mizzaro et al. proposed a framework with three key dimensions of relevance evaluation: information needs (and their levels of expressiveness), information resources (which includes documents as well as their surrogates such as titles and abstracts), and information context (which is the context surrounding the search activity). Their framework illustrates that judgments within a cell in this 3-D space are not necessarily applicable to other cells, which is consistent with our results. Mizzaro’s framework does not cover the dimension pertaining to absolute and relative judgments nor does it cover the effects of set-level versus item-level judgments.

Amento *et al.* [13] correlated editorial document relevance judgments from expert judges with automated evaluation metrics such as the in-degree, PageRank [15], page size, etc, of linked web documents. The results show that these metrics are good predictors of human relevance, although no particular metric stood out. Amento *et al.* reported that variations in human judgments are typically understated. Harter [14] had earlier warned that researchers take relevance judgments variations for granted and that judgments should be collected based on the specific needs and goals of an evaluation, which limited the ability to re-use judgments. Mizzaro [10] pointed out that high rate of disagreement can be attributed to poor testing set-up or to the inherent difficulty in relevance evaluation. In our setting, we ensure that all tests are subject to QA and audit process. Nevertheless, disagreement between judges (regardless of the methodology) is measurable in our systems but is outside the scope of this paper [11].

3. Judgment Elicitation Methods

There appear to be at least broad classes of judgments:

1. **Implicit/Behavioral:** measurement of click-rates, dwell-times, patterns of clicking and returning to earlier results, etc.

2. **Explicit:** Ask a judge which engine is better. Such judgments are more expensive than click-rates but don't suffer the ambiguity of click-rates: higher click-rates don't necessarily mean the results were more relevant.

Furthermore, there are at least three sources of judgments:

1. **Live users:** Users who happen to come to the search engine. Survey data may be collected from such judges.
2. **Volunteer Panelists:** These are ideally random Internet users who have agreed to participate in tests where they may be given queries and asked to compare search engines. They are usually monitored by client-side software and given some monetary reward for their participation.
3. **Editors or domain experts:** These judges have extensive knowledge of web, offline and proprietary (e.g. Deep Web) resources in a particular domain. For our experiments we use domain-expert editors conforming to well-defined judgment guidelines. Internal work we have done has shown their judgments to be reliable with respect to click behavior of average Internet users [11].

Now we consider the advantages and disadvantages of each test type and judgments source. In an ideal scenario, we would like to elicit articulate, patient direct judgments from a perfectly random sample of live users such that that elicitation would not affect their subsequent use of the search engine. This ideal is unattainable for the reasons listed below. Since both implicit and explicit methods have disadvantages and since both measure different aspects of relevance, we need to use both methods.

- Elicitation of direct judgment involves asking the user; this may affect their subsequent searches. To avoid disturbing users, we can consider indirect measures of relevance such as click-through rates.
- The set of users that agree to give their judgments in an online survey is probably not random – it could easily be that busier people are unlikely to say yes to a survey. To avoid such a non-random sample we can again use indirect measures of relevance such as click-rates.
- Users are not perfectly articulate: their behavior may differ from their explanation of it. For this reason, one might again prefer measuring user behavior metrics such as click-rates rather than asking the user.
- Click-behavior is cheap and plentiful to obtain but it is ambiguous. A user may perform more clicks because she likes the results or simply because she is lost.

In practice, it is necessary to combine all these approximations to the ideal in order to build a better joint picture of relevance. In this first paper, we will concentrate on expert editors giving explicit judgments. Future papers will focus on explicit judgments from panelists, contrasting those with editors, and on assessing relevance using implicit (user-behavior) attributes such as click-rates, dwell-times and so on.

4. Framework of Methodologies

Having decided on the distribution of users, queries and documents, further experimental design questions need to be answered. Three of these dimensions form the basis of our framework for this paper:

1. Perceived relevance versus landing-page relevance. The relevance of the web-site (“landing page”) is mediated by the relevance of its presentation (abstract, title, URL) in the Search Engine Results Page (SERP). A site may well be very relevant, but if its presentation attributes are constructed carelessly, users may not click on the result. Judgment made of a landing-page using only its presentation is called *perceived* relevance. Search engines may differ in how well they summarize the underlying page so perceived relevance is a separate relevance metric from landing-page relevance. The following is a partial list of presentation factors, which we will refer to as the <T,A,U> triplet:

- **T: Title:** Sites may not have informative titles. Search engines that automatically construct better titles using the body of the document will get higher perceived relevance scores.
- **A: Abstract:** The abstract is the short paragraph describing the site that appears in the SERP. Abstracts fall into two categories: query-specific and query-independent. Query-specific abstracts (also called *dynamic*) are automatically generated and provide a summary of the site in the context of the user query. Abstract generation engines such as the ones found in Yahoo! and Google are generally of high quality. Static abstracts are often supplied by editors such as Yahoo! Directory or ODP; they tend to be carefully chosen short sentences that may not contain the user query.
- **U: URL:** A given web page may have several URLs as proxies. Search engines that select the URL that appears more relevant for a given query will receive higher perceived relevance scores. For example, for the query ‘Disney’, it would be better to display the alias ‘www.disney.com’ rather than the alias ‘disney.go.com’.

2. Item-level versus Set-level relevance. In order to see the significance of this dimension, one only needs to ask whether ten repeats of an excellent result would constitute an excellent *set* of results. Since the answer is emphatically “No!”, an excellent set should contain excellent individual results but should also have considerations about the diversity of the results or whether different senses of the query are addressed.

3. Absolute relevance versus Relative relevance. This dimension refers to the method of measurement rather than to an entity whose relevance is being judged (by contrast, set-level, item-level and T,A,U are all entities whose relevance is being judged). Joachims [9] has postulated that it is easier to elicit comparative or relative judgments from users “Which engine is better: left or right?” rather than elicit an absolute measure of relevance on a fixed scale without reference to an alternative.

4.2 Advantages and Disadvantages

Each of the test types has its advantages and disadvantages as shown in Table 1. The main advantage of judging at the item level is that those judgments together with a “roll-up” function such as DCG (Jarvelin et al. [8]) that combines item scores into a set score, can be algorithmically re-applied when the ordering of the items is changed. So when a search engine needs to change its ranking function, we don’t need to elicit a new set of judgments. In fact, one can hill-climb through the space of ranking functions to maximize rolled-up DCG score. A ‘rolled-up’ DCG score for a query is simply a position-weighted sum of item-level scores for all items for that query. This all assumes the existence of a good roll-up function. DCG is not an ideal roll-up function in that it does not penalize sets that have duplicates or lack of diversity.

Table 1. Advantages and Disadvantages of different test types.

	Set-level	Item-level
Advantages	Takes duplicates and diversity into account	Recomposable under ranking function changes
Disadvantages	Not recomposable	Doesn’t take duplicates and diversity into account

There are also advantages and disadvantages of judging the presentation rather than the landing page. The presentation algorithm is an independent component of the search engine and should be judged separately. It is important to optimize presentation but this should be done relative to the relevance of the landing-page – the presentation should give a fair assessment of the relevance of the landing-page in response to the query. It should not over-sell or under-sell the landing page.

There are also advantages and disadvantages of relative rather than absolute relevance measurements. Joachims [9] has stated that relative measurements are more reliable in the sense that given the task several times, judges would be more consistent than if they were asked to give absolute measurements. However, the advantage of absolute measurements is that (if they are reliable) they can be used for all kinds of unanticipated purposes. For instance, if a third search engine arises, absolute measurements need only be taken on the new engine and then can readily be compared to existing measurements for the existing engines. Alternatively, longitudinal analysis (trending over time) can be done and each engine’s scores at one date can be compared to its performance one quarter later. Relative measurements also have the disadvantage that if A was judged to be better than B, and then later, A and B are judged to be the same, we do not know if A has gotten worse or if B has gotten better.

5. Experimental Setup

In Section 3 we listed different types of judgments and different sources of judgments. These are attributes of the test methodology. Conversely, the set of queries and results to be judged can also be characterized along several dimensions. The following is a partial list of these:

User distribution. In order to measure the relevance of the engine, one must decide what population or distribution of users to use in the evaluation. Should the evaluation be done over random Internet users or some more specific class such as advanced users.

Query distribution. Queries have changed in distribution since the early days of search engines [12]. Earlier queries tended to be shorter and contain prepositions; now users have realized that the engines are not paying attention to prepositions and so have adapted by accepting a less precise formulation of their need by constructing queries that are sets of keywords. At first glance, it may appear that the user factor is completely mediated by the query in that given the query, the engine can respond without knowing more about the user. This, however, ignores the fact that two users may construct the same query (e.g. ‘jaguar’) for completely different information needs. One user may intend the car, another the MAC operating-system and another the animal. For the experiments in this paper, we use random queries selected from our web server logs. Other sets we could consider include “Tail” queries, commercial queries and ambiguous queries. After the query set is selected, the judges are allowed to “self-select” queries from it. Thus the judge does not have to provide a judgment on an unfamiliar query.

Document Distribution. Search engines with different underlying indices retrieve different documents, which result in difference relevance scores. This dimension will be controlled for by selection of a random set of queries.

There are three kinds of tests we will explore in this paper. These correspond to three points out of the possible eight in the three-dimensional framework we presented in Section 4.

1. Per-set. These tests require the judge to give a single judgment for the entire set of results for the query. By ‘entire set’ we actually mean only the top 10 or 20 results. The ranking of the engine is preserved; judges see the <T,A,U> triplet per result.

2. Per-item. The second test type is the item-level or per-item (PI) test. Here, the judge gives a judgment on each result. The results are presented in a random order so the judge is truly judging the relevance of the result, not the ranking order. In a PI test, judges may or may not be presented the presentation attributes. Depending on the details of the test, they may be required to give a judgment using just the Title, just the Abstract or both. Afterwards, they may be required to click through to the landing page and then render a second judgment on the landing page.

3. Side-by-side (SBS). Judges see two sets of results: each result is presented using its <T,A,U> triplet and rank from its search engine is preserved. The URL is

usually clickable to they can check out the landing page before giving their judgment. In addition to giving a score, they sometimes record free text reasons supporting their judgment and these have proved very useful. They don't know which side corresponds to which search engine. Sides are randomized so each engine gets 50% of the queries on the left side.

6. Set-level versus Item-level Judgments

In this section, we explore the relationship between Set-level and Item-level Judgments for two domains: Image Search and News Search. For each domain, we look at the correlation between set-level and item-level judgments and characterize what kinds of result-sets receive high set-level but low rolled-up item-level scores.

6.1 Image Search

In the set-level test, the judges were shown 20 images in a 5-by-4 matrix with the 1st row being the images ranked highest by the search engine. 299 queries were judged by 2 or 3 judges each. The judgments were given on a scale of '1' being best and '3' being worst. The queries were self-selected by judges. 24 judges were involved in this test. In the item-level test, the images for a given query were presented one at a time, in randomized order. 282 queries were judged for a total of 6856 judgments. 198 queries were found in common between the tests (see grand total in bottom right cell of Table 2).

Table 2. Image Search: Contingency matrix for per-set versus rolled-up item-level judgments. For instance, there were 130 queries with SET=1 and DCG=1.

	SET=1	SET=2	SET=3	Marginal
DCG=1	130	18	1	149; 75%
DCG=2	16	13	7	36; 18%
DCG=3	3	5	5	13; 7%
Marginal	149; 75%	36; 18%	13; 7%	198; $r=0.54$

To analyze the differences between set-level and item-level judgments, we only considered queries for which the search engine returned the full 20 images on page one. In order to do a query-level analysis between set-level and item-level scores, we had to, for each query, roll-up its 20 item-level scores to produce a single set-level score. To do this roll-up, we used the DCG position-weighted average. The rolled DCG score forms the *DCG* random variable in Table 2. For the set-level test, if the query was judged by several judges, we just used their average score to produce the

Set random variable. We discretized the DCG scores so that the bucket boundaries would reflect the proportions seen in the three levels of the *Set* variable. In interpreting the table, recall that for *Set* and *DCG*, that ‘1’ is the best score.

The Pearson correlation between Set-level judgments and rolled-up Item-level judgments is a middling 0.54. One can thus conclude that these two variables are measuring different kinds of image-relevance; judges are measuring different underlying factors in the Set-level test than they are in the Item-level test. To get a better idea of what kinds of search quality factors each measurement methods was sensitive to, we looked at a few outlier queries: queries which scored good scores on one axis and bad scores on the other. The query “hollow man” (after a movie) received a high set-level score but poor rolled-up DCG score. Looking at detail at the judgments for this query, we saw that most of the items were irrelevant but a couple of them were about the movie. At the set-level, the user only wants perhaps one icon or gif/jpeg to use; he does not much care if it is at a lower position. This is especially true for image search in which the images are scanned very quickly by the eye. At the item-level however, the judge saw that most of the items were irrelevant images so she gave poor item-level scores and a poor rolled-up DCG score. So we can conclude that for image searches *where just one or a few relevant items is sufficient to satisfy the user* that there will be a discordance between item-level and set-level scores.

Another effect occurring for this query was that the images with the highest individual scores were at the bottom of the set and thus the rolled-DCG score was low. So poor *ranking* can lead to a low rolled item-score but why did the set-level judge not penalize the set that had its best images as the bottom? We believe this is a idiosyncrasy of image search: all the images are scanned in parallel by the eye so ordering is not so important for image sets. Had this been web-search, the set-level judgment would have been a poor score.

At the other extreme we saw the query ‘Slam Dunk’ which had a poor set-level score but a good rolled-up item-level score. The set-level judge gave a poor score because about 90% of the images were about the video-game ‘Slam Dunk’ and only one was an actual photo of a slam dunk in a basketball game. The judge expected there to be many more real photos and judged the set as a whole to have essentially missed the most important sense of the phrase ‘Slam Dunk’. The reason that this got good item-level scores was that the items are presented to the judge in a random order. The judge only needs to make individual judgments on each item and since each item was either about the video-game or a real photo, each item was scored well. The item-level judge did not maintain a memory of the *distribution* of real photos to video-game images.

6.2 News Search

For News Search, the set-level test involved 23 judges giving 150 judgments over 148 queries – only 2 queries were judged more than once. The item-level test involved 25

judges providing 1284 judgments over 128 queries for an average of about 10 judgments per query.

Table 3 shows results comparing news per-set and news PI. The correlation between news per-set and PI was 0.29, lower than it was for image search. Duplication of the same story from different sources seems to be a leading cause of the difference. Since the average number of results in News search is only 3 to 5, duplicates are more strongly penalized in News Search than in Image Search. Ranking also seems particularly important for news. A bad first result can mar the entire set. For example, for the query ‘Mary Kate Olsen’, the first result is actually about the debut of Jenna and Barbara Bush – Mary and Kate Olsen were tangentially mentioned in the article. This set got a terrible rating whereas the average PI score was high.

Table 3. News Search: Contingency table for per-set versus rolled-up item-level judgments.

	SET=1	SET=2	SET=3	Marginal
DCG=3	9	3	11	23; 18%
DCG=2	17	12	4	33; 26%
DCG=1	45	19	7	71; 56%
Marginal	71; 56%	34; 27%	22; 17%	127, r=0.29

6.3 Cost Analysis

We use number of judgments as the primary metric of cost. The cost of a judgment is a function of the type of the judgment — PI or per-set — as well as the type of the result — image or text (news). To compute the relative cost of PI versus per-set in image search, let R_{pi} be the cost of one PI image judgment and R_{ps} be the cost of one per-set judgment. There were 6856 item-level and 302 set-level judgments. The relative cost at the test level is $(6856 * R_{pi}) / (302 * R_{ps})$. If we assume that each per-set judgment takes as much time as N PI judgments, then per-set will be less expensive than the PI as long as N is less than 22.7. Our experience indicates that N is on the order of 3 to 5 because it costs little to scan through the set and the judge need not scan through the entire set. This implies that per-set is about 5 to 7 times more cost-effective than PI. Similarly, for News search, we get $1784 * R_{pi} / (150 * R_{ps})$. With N being on the order of 4 to 6 (out of 10), we estimate that news per-set is about 2 to 3 times more efficient.

In summary, for image search, we identified two main sources of differences between set-level and item-level: poor rankings and missed important meanings. Item-level did not pick up these factors. However, as we pointed out in section 4.2, PI is still useful for computing tuning and ranking-function changes. For news search, we saw that duplication and poor first result were the main causes of differences between per-set and item-level evaluation.

7. Perceived Relevance versus Landing-Page Relevance

In this section we compare perceived relevance versus landing-page relevance for item-level judgments for advertising results and news search results. That is, we hold constant the ‘size’ of the object being judged to be at the item-level. As in all PI tests, the results were presented in random order to the judges. The presentation of the item consisted of two parts: the title and the abstract (the URL was not presented). The judge had to render two Boolean judgments before clicking-through: one for whether the title was relevant (random variable *Title*) and another for whether the abstract was relevant (*Abstract*). After clicking-through, they could see the landing page behind the result and were asked to render another judgment (*LandingPage*).

7.1 Advertising Results Search

Since the advertising results section typically has four to six results, each judge had to judge fewer results per query than she had to for web results. The landing-page judgments ranged from ‘1’ being perfect to ‘5’ being poor. For our correlation analysis, we recoded this variable to Boolean by ignoring the neutral ‘3’ score, by coding ‘1’ and ‘2’ to ‘1’ and by recoding ‘4’ and ‘5’ to ‘-1’. The test was done over 470 random commercial queries by 32 judges yielding 2003 judgments. In this statistical formulation, we obtained a Pearson correlation of $r = 0.63$ ($r^2 = 0.40$) between the landing-page score and the title relevance. We obtained a lower $r = 0.55$ ($r^2 = 0.30$) correlation for abstract-relevance. We also wanted to create a compound variable that captured both aspects of perceived relevance (title and abstract). For this, we summed title-relevance and abstract-relevance and ignored the results with sum 0 (15% of all results). Combining these factors produced a higher correlation ($r=0.77$) but it may throw away the hard cases. If we re-include the ‘0’s, we get a correlation of $r=0.66$. Tables 4 through 6 present these results; ‘NR’ stands for not-relevant, ‘R’ stands for relevant.

Table 4. Title (Presentation Factor) Relevance versus Landing-Page Relevance

	landingPg=NR	landingPg= R	marginal
title= NR	53	20	73; 15%
title= R	28	369	397; 85%
Marginal	81; 17%	389; 83%	470; $r=0.63$

Table 5. Abstract (Presentation Factor) Relevance versus Landing-Page Relevance

	landingPg=NR	landingPg= R	marginal
--	--------------	--------------	----------

abstract=NR	64	58	122; 26%
abstract=relev.	17	331	348; 74%
Marginal	81; 17%	389; 83%	470; r=0.55

Table 6. Overall Perceived (Presentation Factor) Relevance versus Landing-Page Relevance

	landingPg=NR	landingPg=R	marginal
perceived=NR	53	9	62; 16%
perceived=R	17	320	337; 84%
Marginal	80; 20%	389; 80%	399; r=0.77

Looking in detail, we found a number of queries where the landing-pages were good but the title or abstract were not. In the first, the query was ‘world war 2’, the title was ‘Perilous Fight on VHS and DVD: Save 15%’ and the abstract was ‘Publicvideostore.org offers a vast selection of offerings from the BBC...’. This is an example of the title/abstract essentially advertising the provider rather than being sensitive to the query. As the contingency matrices in Tables 4 through 6 imply, the converse was rarer: finding good titles and abstracts that led to poor landing pages. One class of these involves landing pages that generate HTTP Not-Found 404 errors. Another rare class involves over-advertising. For example, for the query ‘DMV’, the title was ‘Access DMV Records’ but the landing page did not lead to dmv.org; instead it directs users to an intermediary or broker: www.public-record-searches.com.

7.2 News Search

For news search, Tables 7 through 9 below summarize the results. We see higher correlations between perceived (title,abstract) and landing-page relevance than we did for advertisement search because news titles and abstracts are carefully written to describe the underlying documents, and not to advertise the provider.

Table 7. News Search: Title relevance versus landing page

	landingPg=NR	landingPg= R	marginal
title= NR	659	383	1042; 30%
title= R	23	2301	2324; 70%
marginal	682; 20%	2684; 80%	3366; r=0.72

Table 8. News Search: Abstract relevance versus Landing-Page relevance

	landingPg=NR	landingPg= R	Marginal
abstract=NR	512	80	592; 18%
abstract=relev.	170	2604	2774; 82%
marginal	682; 20%	2684; 80%	3366; $r=0.76$

Table 9. News Search: Perceived relevance versus Landing Page relevance

	landingPg=NR	landingPg=R	Marginal
perceived=NR	511	62	573; 20%
perceived=R	22	2283	2305; 80%
marginal	533; 19%	2345; 81%	2878; $r=0.91$

7.3 Cost Analysis

For perceived versus landing page relevance, we want to measure the cost of reading the title and abstract versus the cost of reading the landing page. One proxy for this is the number of words. In advertisement search, we estimate that the number of words in the title and abstract is less than 200. The average number of readable words in a advertisement landing page is about 500 words. This yields a 2 to 3 fold reduction in judgment cost. In news search, the reduction is more significant as the number of readable words in the landing-page is around 800.

8. Real-World Test: Absolute versus Relative Judgments

In addition to the tests above that explore the effect of varying one factor at a time, we wanted to simulate the real-world condition where some users click-through and hence provide landing-page judgments whereas others provide perceived judgments. For this experiment we compared web results from two search engines and did two tests: Side-By-Side (Set-level, mixture of perceived and landing-page) and PI: Per-Item (Item-level, landing-page). For the set-level test we used 36 judges judging 887 randomly chosen queries. For the item-level test we used 40 judges judging 847 queries with up to 10 results each. Retaining only queries that were self-selected in both tests, and that yielded at least 10 results we ended up with 658 queries. For PI, the DCG rollout function was computed separately for each engine to yield two rolled-up scores: x , y . Then a relative DCG number was computed as $(x-y)/(x+y)$.

Figure 1 shows a weak correlation ($r^2 = 19\%$) between the SBS scores and the PI (relative DCG) scores. Outlying queries in the figure corresponded to queries that received a high rolled-up PI score, but low SBS score because the query had many duplicates or whose results were poorly ordered.

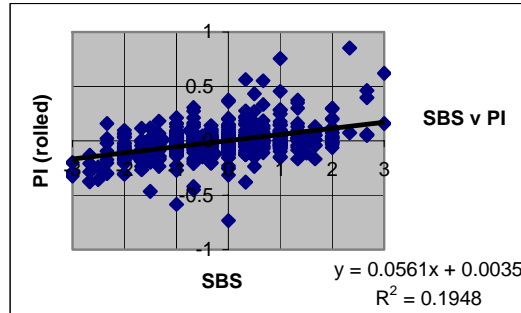


Fig. 1. Correlation between Set-level Relative and Item-level Absolute Judgments.

There are other reasons for this low correlation. The SBS test allows some users to base judgments on perceived relevance, others on landing page relevance. Another reason is that, as previous sections showed, PI versus per-set correlation is already low so the correlation to SBS will be even lower. In related work [11] we have observed higher correlation by only considering queries with multiple judges and non-adult queries.

9. Conclusions

This paper presents a methodological framework for evaluating search engine relevance. We have experimented with a subset of methods that we found to be practical and cost-effective over four different types of search engines: image search, advertising results search, news search and web results search.

We have shown that set-level judgments are capable of measuring aspects (poor ranking, missed important senses) of relevance missed by item-level evaluation. We have presented results comparing perceived relevance versus “actual” (landing-page) relevance and shown that there is a moderate correlation between the two. The factors causing differences are poor title and abstract construction. We have also evaluated domain-specific search engines. For image search we found that ranking is less critical than it is for web search as long as the relevant image is somewhere in the first page. For advertising results, we found that query-insensitive titles and abstracts were under-selling the target web-sites. For news search we found a particularly high correlation between perceived relevance and landing page relevance. We conclude that overall our experiments suggest that there is no single method for

comprehensively measuring search relevance. The methodology to be chosen depends on the search domain, the measurement intent (perceived or actual) and the cost of the available editorial resources.

Acknowledgments: Thank you to Jan Pedersen at Yahoo! and the editorial team for hundreds of hours of work on judging search results.

References

1. Cleverdon, C. The significance of the cranfield tests on index languages. Proceedings of the SIGIR Conference on Research and Development in Information Retrieval, pages 3-12, 1991.
2. Amitay, E., Carmel, D., Lempel, R., Soffer, A. Scaling IR-System Evaluation using Term Relevance Sets. In Proceedings of SIGIR 2004, pages 10-17, Sheffield, UK.
3. Buckley, C., Voorhees, E., Retrieval Evaluation with Incomplete Information. In Proceedings of SIGIR 2004, pages 25-32, Sheffield, UK.
4. Voorhees, E., The philosophy of information retrieval evaluation. In Proceedings of the Second Workshop of the Cross-Language Evaluation Forum (CLEF 2001), pages 355-370, 2001.
5. Buckley, C., Voorhees, E., Evaluating evaluation measure stability. In Proceedings of SIGIR 2000, pages 33-40.
6. Zobel, J., How reliable are the results of large-scale information retrieval experiments? In Proceedings of SIGIR 1998, pages 307-314, Melbourne, Australia.
7. Gabrieli, S., and Mizzaro, S., Negotiating a Multidimensional Framework for Relevance Space. In Proceedings of MIRA Conference, 1999.
8. Jarvelin, K. and Kekalainen, J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (ACM TOIS) 20(4), 422-446.
9. Joachims, T. Evaluating Retrieval Performance Using Clickthrough Data, Proceedings of the SIGIR Workshop on Mathematical/Formal Models in Information Retrieval, 2002.
10. Mizzaro, S. How Many Relevances in Information Retrieval? Interacting With Computers, 10(3):305-322, 1998.
11. Chang, C. and Ali, K. How much correlation is there from one judge to another? Yahoo! Technical Report, 2004-12.
12. Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. Analysis of a Very Large AltaVista Query Log, SRC Technical Note #1998-14.
13. Amento, B. Terveen L. and Hill W. D. "Does 'Authority' Mean Quality? Predicting Expert Quality Ratings of Web Sites". Proceedings of SIGIR 2000 (Athens, Greece).
14. Harter, S. Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness." JASIS, 47(1):37-49, 1996.
15. Brin, S. and Page L. 1998. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1-7): 107-117.