Evaluation by Highly Relevant Documents

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
ellen.voorhees@nist.gov

ABSTRACT

Given the size of the web, the search engine industry has argued that engines should be evaluated by their ability to retrieve highly relevant pages rather than all possible relevant pages. To explore the role highly relevant documents play in retrieval system evaluation, assessors for the TREC-9 web track used a three-point relevance scale and also selected best pages for each topic. The relative effectiveness of runs evaluated by different relevant document sets differed, confirming the hypothesis that different retrieval techniques work better for retrieving highly relevant documents. Yet evaluating by highly relevant documents can be unstable since there are relatively few highly relevant documents. TREC assessors frequently disagreed in their selection of the best page, and subsequent evaluation by best page across different assessors varied widely. The discounted cumulative gain measure introduced by Järvelin and Kekäläinen increases evaluation stability by incorporating all relevance judgments while still giving precedence to highly relevant documents.

1. INTRODUCTION

The information retrieval research community has a well-established evaluation methodology for comparing the relative effectiveness of different retrieval methods. Unfortunately, little rigorous evaluation of the effectiveness of various web search engine strategies has been performed. This has in part been due to the lack of a test collection specifically developed for web retrieval, coupled with skepticism on the part of the search engine industry that existing test collections are appropriate for the task. Existing collections do not contain the link structure inherent in the web. Furthermore, most existing collections use binary relevance judgments, which the industry considers to be unsuitable for evaluating web searches [11]. Instead, they believe that engines should be evaluated by their ability to retrieve the very best pages.

The Text REtrieval Conference (TREC) has had a web track for the past two years to address the need for a web test collection. The documents and queries in the collections built in the track were taken from the web (see below), and the relevance judgments were produced using pooling as in other TREC collections [15]. Unlike other TREC collections, the most recent web track used a three point relevance scale: *not relevant*, *relevant*, and *highly relevant*. In addition, assessors also selected the best document from the pool.

This paper examines the effect of incorporating highly relevant documents into retrieval system evaluation. By comparing the ranking of systems produced when evaluating systems by highly relevant documents only against the ranking produced when evaluating systems by generally relevant documents, we show that the relative effectiveness of different web track runs differs depending on which judgments are used. Differences in system rankings were found for all evaluation measures computed, including mean average precision and precision at ten documents retrieved. This finding confirms the hypothesis that different retrieval techniques should be used for retrieving highly relevant documents.

Section 3.2 examines the discounted cumulative gain (DCG) evaluation measure introduced by Järvelin and Kekäläinen [10]. DCG incorporates multiple relevance levels into a single measure. Using the web track data, we show how the system rankings vary between the two extremes of using all relevant documents and using only highly relevant documents as the ratio of the weights given to highly relevant documents vs. generally relevant documents varies. A larger ratio mimics evaluation by relevant documents only, yet is more stable since all relevant documents are used in computing scores.

Section 4 examines the effects of restricting the evaluation to only the best documents. Three different TREC assessors were asked to select the best page for a topic. The assessors frequently chose different best pages, and when they did agree on a page they often selected it for different reasons. The rankings of systems produced by evaluating runs using different assessors' best pages varied widely, indicating that best page evaluation is unsuitable as an evaluation metric for the type of queries used in the track.

2. THE TREC-9 MAIN WEB TASK TEST COLLECTION

The TREC web track is a successor to the very large collection (VLC) track [7]. Both the VLC and web tracks were coordinated by David Hawking of CSIRO, with assistance from his colleagues in the ACSys project at the Australian

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA..

National University. The document sets used in the tracks were various subsets of a spidering of the web produced by the Internet Archive in 1997 [1]. Hawking and his colleagues created the subsets by balancing competing desiderata including target size, selecting at least some content-heavy pages, having naturally defined sub-collections contained within the set, and having a good set of closed hyperlinks [9]. The TREC-9 main web track used a 10.3 gigabyte subset of the spidering (named WT10g) chosen to include a significant number of inter-server links. WT10g does not include images contained within the pages, and pages that were obviously not text or obviously non-English were also eliminated[1].

TREC provides statements of information need called *topics* rather than queries. The topic statements allow researchers to experiment with different query formulation methods as well as provide a means for describing how the topic will be judged. TREC organizers wanted to use realistic web queries in the web track while maintaining the benefit of standard TREC topic statements. To this end, NIST relevance assessors retrofitted TREC topics around seeds mined from a log of Excite queries[2]. Assessors were given a list of 260 queries taken from the log and asked to select one that interested them. Each assessor then wrote a one-sentence "Description" that fit the query and a "Narrative" that explained how the topic would be judged. In general, the queries were short, and thus could be broadly interpreted. NIST instructed the assessors to pick some more specific interpretation of a broad seed as the subject of the topic. For example, a seed of "cats" might have a topic description relating to information on the care of house pets or a topic description on finding tickets for the musical. The so-called "Title" field of the topic statement contained the Excite query, exactly as it appeared in the log. Fifty topics were created for the TREC-9 web track.

Web track participants were encouraged to submit to NIST at least one run that used only the "Title" field during query formulation, and were permitted to submit up to six runs. Runs could use both page content and link structure, or page content alone. A maximum of two content-only and two content-link runs of the participant's choosing were added to the judgment pools. The relevance assessor who retrofitted the topic statement to the Excite query judged the pool for that topic.

As mentioned in the introduction, the assessors were asked to make three-way judgments for the track. The assessors were given the following instructions before they made any judgments:

> Assume that you have the information need stated in the topic and that you are at home searching the web for appropriate material. If the document contains information that you would find helpful in meeting your information need, mark it relevant. If the document directly addresses the core issue of the topic, mark it highly relevant. Otherwise, mark it not relevant.

The total number of relevant documents of either type found across the 50 topics in the topic set was 2617, with a mean of 52.3 and a minimum number of relevant for a topic of 1 and a maximum of 519. For highly relevant documents only, the total across all topics was 246, with a mean of 4.9 and a minimum of 0 (3 topics) and a maximum of 32.

The relevance assessors were also asked to select the best document from the pool. NIST defined the best document as the page the assessor would most like to see returned in response to the topic (i.e., the criteria used to determine the best document was left up to the individual assessor). So that assessors would not spend too long agonizing over decisions, they were allowed to select multiple bests or to decide no document was best.

After all the TREC-9 assessing was completed, the assessors were given the set of documents they had marked either relevant or highly relevant to a topic (the *relevant pool*) and were asked to fill in the questionnaire shown in Figure 1 for that topic. (If there were more than 100 documents that were judged relevant, the relevant pool contained only the highly relevant documents.) The purpose of the questionnaire was to gain insight into how and why the assessor chose a particular document as the best. Two other assessors were independently given the topic statement and the relevant pool for that topic, and were asked to complete the questionnaire.

3. EVALUATION BY MULTIPLE RELEVANCE LEVELS

Retrieval runs are generally evaluated using *precision* and *recall*, and other measures derived from precision and recall. Precision is the proportion of retrieved documents that are relevant, while recall is the proportion of relevant documents that are retrieved. These measures are set-based measures, where the retrieved set is defined by a cut-off level. A cut-off level of 10, for example, says that the retrieved set is the first 10 documents in the ranked list returned by the retrieval system. Mean average precision is a commonly used measure that is computed over the entire ranking. The average precision for a single topic is the mean of the precision obtained after each relevant document is retrieved. The mean average precision for a run consisting of multiple topics is the mean of the average precision scores of each of the topics in the run.

3.1 Using only highly relevant documents

Evaluation by these traditional measures assumes binary relevance judgments. With the ternary judgments made for the web track, we can evaluate web track runs considering both relevant and highly relevant documents as relevant (call this set R1), or considering only the highly relevant documents as relevant (R2). To see if there is a difference between the two evaluations, we can compute a measure of association between the system rankings produced when using the two relevance judgment sets. A system ranking is an ordered list of the TREC-9 web track runs sorted by decreasing value of the effectiveness measure used to evaluate the runs. A low correlation between the system rankings produced by using R1 and R2 would indicate that some retrieval strategies are better at retrieving highly relevant documents while others are better at retrieving generally relevant documents.

In earlier work, we used this methodology of comparing

[1] Such pages are frequently not obvious, so WT10g still contains some of this material.

[2] The log was fetched by anonymous ftp from `ftp.excite.com/pub/jack/Excite_Log_12201999.gz` and represents a few hours of queries on December 20, 1999.

1. What document(s) did you pick as the best document(s) for this topic from the pool given?
2. What is it about this document that makes it the best?
3. Was picking a best document a difficult choice? (Yes No)
   If so, why?
   (a) too many good documents
   (b) no document was particularly great
   (c) there was a set of very similar documents
   (d) different documents were better than one another in different ways
   (e) other (what?)
4. Please use the following responses for question 4.
   1: much worse 2: worse 3: about the same 4: better 5: much better X: can't tell
   How does the best document compare to other relevant documents in terms of the following attributes:
   (a) Quality of content?
   (b) Authoritativeness of site?
   (c) Breadth of content coverage?
   (d) Recency of content?
   (e) Presentation of material?
5. If you had the information need stated in the topic and a search engine returned this document to you ranked first in a list of documents, what would you do next?
   (a) Stop, the document provides all I want to know.
   (b) Follow some of the links contained in this document.
   (c) Look at the next document in the ranked list.
   (d) Try another search.

Figure 1: Best document questionnaire

system rankings to show that the relative effectiveness of different retrieval systems remains the same when systems are evaluated over different sets of relevant documents [14]. In that study, multiple relevance assessors judged each topic in the TREC-4 and TREC-6 topic sets, and the Kendall τ correlation [12] was computed among rankings of systems produced by evaluating runs using the judgment sets from different assessors. While the overlap between different judgments sets averaged .45, indicating significant disagreement among assessors, the correlations between system rankings were almost always greater than .9. However, that study focused on different assessors' opinions, and did not distinguish between highly relevant and generally relevant documents. The goal in this study is to determine whether the type of relevance judgment affects system performance.

Figure 2 gives the system rankings of the 40 title-only web track runs when evaluated using both mean average precision (MAP) and precision with a cut-off of 10 (P[10]). Each run was assigned a one letter identifier in the ranking produced by MAP using set R1, and the same identifier is used for that run in each ranking. Some runs, such as run S perform much better relative to the other runs when evaluated by set R2.

We use a correlation based on Kendall's τ to quantify the difference between a pair of rankings. Kendall's tau computes the distance between two rankings as the minimum number of pairwise adjacent swaps to turn one ranking into the other. The distance is normalized by the number of items being ranked such that two identical rankings produce a correlation of 1.0, the correlation between a ranking and its perfect inverse is -1.0, and the expected correlation of two rankings chosen at random is 0.0. Table 1 gives the correlations between the system rankings when the systems were evaluated using mean average precision, precision with a cut-off of 10, and precision with a cut-off of 100. The third column in the table gives the number of interchanges the τ represents given that there are 40 runs being ranked. The highest correlation is .8, which is noticeably smaller than the correlations found in the earlier study, and supports the conclusion that different retrieval systems are better at finding the highly relevant documents than those that are better at finding generally relevant documents.

Table 1: Kendall τ correlation and corresponding number of pairwise differences between system rankings when evaluated using relevant sets R1 (both relevant and highly relevant) and R2 (highly relevant only).

Measure	Kendall τ	# interchanges
MAP	.8000	78
P[10]	.7051	115
P[100]	.6744	127

How is this conclusion reconciled with the earlier study that showed the relative effectiveness of retrieval systems is stable despite differences in the relevance judgment sets?

MAP using R1:	ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmn
MAP using R2:	DSBAFCIPEKQLGHOMNJVTRYZXUbWcdefaihgljnkm
P[10] using R1:	HCBPEGMANQDJFOILKRTUSVYXZcdefWbahnkiljgm
P[10] using R2:	HCPFOUSBAQIXJKMNLTVYREGDZcdefbWhaignljkm

Figure 2: System rankings produced using different relevance judgment sets. Each system is assigned a one-letter identifier that remains the same across rankings.

The robustness of retrieval evaluation across different sets of human assessors' judgments is due mostly to the effect of averaging over many topics [14]. Retrieval effectiveness does change for individual topics when relevance sets vary, but the differences caused by using different human judges are uncorrelated with retrieval system functionality, and average effectiveness remains the same. The ability to separate highly relevant documents from generally relevant documents evidently is correlated with system functionality, and thus differences among systems are reflected in the average scores.

In their evaluation of web search engines, Gordon and Pathak noted that the Infoseek search engine scored much better relative to the other search engines when the evaluation was based on "strict" (highly relevant) judgments as opposed to when evaluation was based on "lenient" (generally relevant) judgments [6]. They hypothesized that perhaps Infoseek was tuned for high early precision (i.e., tuned such that the documents retrieved at the very top of the list are relevant, possibly at the expense of lower recall), and that was why its performance differed relative to the other engines. However, the TREC-9 results demonstrate that retrieving highly relevant documents is not the same as emphasizing high early precision—Table 1 shows the correlation between system rankings for P[10] is only .7051. As examples, consider systems 'F', 'M', and 'P' in Figure 2. System 'P' appears to be tuned for high early precision since its performance relative to the other systems is much better when evaluated by P[10] than when evaluated by MAP; it does relatively better when evaluated by judgment set R2 (highly relevant only). On the other hand, system 'M' is also ranked higher when evaluated by P[10] than when evaluated by MAP, but its relative performance worsens when evaluated using R2. Finally, system 'F' is ranked higher when evaluated by MAP, but it is better relative to the other systems when evaluated by R2.

The number of relevant documents in the R2 judgment set is very much smaller than the number of relevant documents in the R1 set. However, it is unlikely that the number of relevant documents alone could cause the observed differences in relative scores. The earlier study found no difference in the relative effectiveness when evaluating by the intersection of different assessors' relevance sets, and the intersection was also much smaller than the original sets. Unfortunately, the small number of highly relevant documents is a problem for a different reason. Retrieval evaluation measures are inherently unstable when a topic has only a few relevant documents in that a small change in a document ranking can produce a large change in the evaluation score [14]. The instability makes it more difficult to reliably conclude that two systems are significantly different from one another [4]. For the TREC-9 judgments, 39 of the 50 topics had 5 or fewer highly relevant documents, and an additional 5 topics had between 6 and 10 highly relevant documents. That is, only 6 topics had more than 10 highly relevant documents. It is possible that the differences seen in the system rankings are caused by the inherent instability of the evaluation measures with this few relevant documents rather than the distinctiveness of highly relevant documents.

Assuming the difference between levels of relevance is real, the instability inherent in using small numbers of relevant documents makes it difficult to use highly relevant document evaluation. The *discounted cumulative gain* (DCG) measure introduced by Järvelin and Kekäläinen should be more stable since it exploits all available relevance information while rewarding systems for retrieving highly relevant documents first [10].

3.2 Discounted cumulative gain

In this section we explore the behavior of DCG as the relative weight given to highly relevant documents varies. We show that by manipulating this weight we can closely approximate either evaluation by all relevant documents or evaluation by highly relevant documents only. By using intermediate gain values, we produce an evaluation that lies between the two extremes.

The discounted cumulative gain measure was introduced in a paper by Järvelin and Kekäläinen [10]. In that paper, the authors define the measure and list several advantages over similar measures such as average search length [5] and ranked half life [2]. The chief advantage for the current study is that DCG incorporates multiple relevance levels into a single measure. The vast majority of retrieval evaluation measures, including average search length, assume binary judgments. Ranked half life can incorporate different relevance levels, but is based on *relative relevance*, which combines a system-dependent probability of being retrieved with a user-assessed degree of relevance.

Discounted cumulative gain is computed as follows. Given a ranked list of documents of length N and a set of weights or *gain values* associated with each of the different relevance levels, the ranked list is first converted to a gain vector, $\vec{G}$, by replacing document ids with the appropriate gain value. For example, assume that the gain values are 0, 1, and 2 corresponding to not relevant, relevant, and highly relevant documents respectively. If $N = 5$, and the first document is relevant, the second and fourth documents are highly relevant, and the third and fifth documents are not relevant, the corresponding gain vector is $\vec{G} = \langle 1, 2, 0, 2, 0 \rangle$. The discounted cumulative gain vector $D\vec{C}G$ is then defined as

$$DCG[i] = \begin{cases} G[1], & \text{if } i = 1 \\ DCG[i-1] + G[i]/\log_b i, & \text{otherwise} \end{cases}$$

In this formula, b, the base of the logarithm, is a parameter of the measure and controls how much a document retrieved at a lower rank is penalized. In the experiments reported here, the natural log is used (and since $\ln 2 > 1$, no discount is applied for the first two ranks). The DCG vector for a run consisting of multiple topics is the mean of the individual topic vectors.

Järvelin and Kekäläinen included a case study in their paper that shows how incorporating multiple relevance levels into retrieval evaluation provides better discrimination among a set of retrieval runs. They also demonstrated how changing b allows the DCG measure to model the persistence of a user reviewing a ranked list. For example, a small value of b causes the value of a relevant document to decrease quickly, reflecting an impatient user who will never read poorly ranked documents. For the current study, we keep b constant and investigate the effect of using different gain values on the measure.

The best possible DCG can be computed by assuming a system first retrieves all highly relevant documents followed by all relevant documents for each topic. The average DCG for the best possible run and four TREC-9 title-only web track submissions is plotted against document rank for ranks 1–100 in Figure 3. For each plot, the gain value used for not relevant documents is 0, and the gain value used for relevant documents is 1. The gain value used for highly relevant documents varies from 1 through 1000.

The shape of the best possible curve changes as the relative weight of highly relevant documents increases, reflecting the scarcity of highly relevant documents. When relevant and highly relevant documents are equally important, each of the relevant documents affects the cumulative gain, and the best possible DCG continues to increase even after rank 100. When highly relevant documents are much more important than relevant documents, the DCG reaches its maximum as soon as all highly relevant documents are retrieved. Since a large majority of the topics have fewer than 10 highly relevant documents, the best possible DCG is close to its maximum by rank 10. The shape of the individual runs' curves does not change as dramatically because they do not retrieve as many highly relevant documents.

The relative effectiveness of the TREC-9 runs can be seen to change in Figure 3 as the importance of highly relevant documents increases. The run plotted with a '+' is the run labeled 'S' in Figure 2. This run is clearly beneath the others when the gain value for highly relevant documents is 1 or 2, is approximately equal to all but the top run when the gain is 5 or 10, and is above all but the top run when the gain is 100 or greater. Table 2 gives the Kendall τ correlations for the rankings over the entire set of title-only runs when evaluated at rank 10 and rank 100 using the six different gain values for highly relevant documents. The correlations confirm that the relative effectiveness of the whole set of runs, not just run S, changes as more emphasis is placed on highly relevant documents.

It would be useful to have a critical value for the τ correlations such that correlations less than the critical value reflect important differences in the evaluation. However, the τ value is a measure of how similar two rankings are, not a measure of significance. The value is affected by the set of runs being compared: a set of runs each of which have roughly similar effectiveness will tend to have smaller correlations than a set of runs with widely varying effectiveness because the widely varying runs will sort in the same order regardless of the particular evaluation scheme used. Nonetheless, correlations computed over the same set of runs can be meaningfully compared. A smaller correlation implies more differences in the relative effectiveness of the systems being compared, and hence is a stronger indication that the two evaluation schemes measure different things. As a practical matter, the correlation between rankings produced by using different relevance assessors is generally greater than .9. Thus other evaluation schemes that produce correlations of at least .9 should be considered equivalent since it is not possible to be more precise than this. Correlations less than .8 generally reflect noticeable changes in the rankings, not simply inversions among neighbors, and suggest that the evaluation schemes have different emphases. (Any evaluation scheme is based on rewarding systems that retrieve good documents, so the rankings are not random and correlations much greater than 0.0 are to be expected.)

The correlations given in Table 2 show that treating all relevant documents equally (gain ratio of 1:1) is basically equivalent to giving highly relevant documents a small emphasis (gain ratio of 2:1) since the correlation is approximately .95. Giving yet more weight to highly relevant documents (5:1 ratio) results in a ranking that is different from both no emphasis and very strong emphasis. Relative importances of 10:1 or greater produce rankings that are similar to one another and markedly different from equal importance rankings.

These changes in rankings when the relative importance of highly relevant documents varies correspond to the differences found using the traditional measures when evaluating by all relevant (R1) vs. highly relevant only (R2). Table 3 gives the correlations for rankings produced by mean average precision, precision at cut-off 10, and average discounted cumulative gain at cut-off 10. For Prec[10] and DCG[10], the R1 and low-emphasis rankings are highly correlated with one another, as are the R2 and high-emphasis rankings. The mid-emphasis rankings fall between the two extremes. The same general trend is seen with MAP, though DCG[10] is a high precision measure so all of the correlations are weak.

The strong correlations between the R2 and high-emphasis rankings and the shape of the best possible curves in Figure 3 show that the DCG measure is completely dominated by the highly relevant documents when large gain ratios are used. This implies that the DCG measure with large gain ratios is also unstable since it, too, is based on very few relevant documents. Using DCG with smaller ratios (say, 3–5) will incorporate all relevance information into the score to increase stability, while still rewarding systems for retrieving highly relevant documents first.

4. BEST DOCUMENT EVALUATION

The previous section investigated the problem of evaluation when the task is to retrieve highly relevant documents. The motivation for that investigation is the belief that web users will be better served by systems that retrieve highly relevant documents rather than generally relevant documents. Taking this reasoning a step further, some have argued that web search engines should actually be evaluated on their ability to retrieve the very best page. This section examines the feasibility of best page evaluation. Since the small number of highly relevant documents was a concern in the previous section—and there are even fewer best

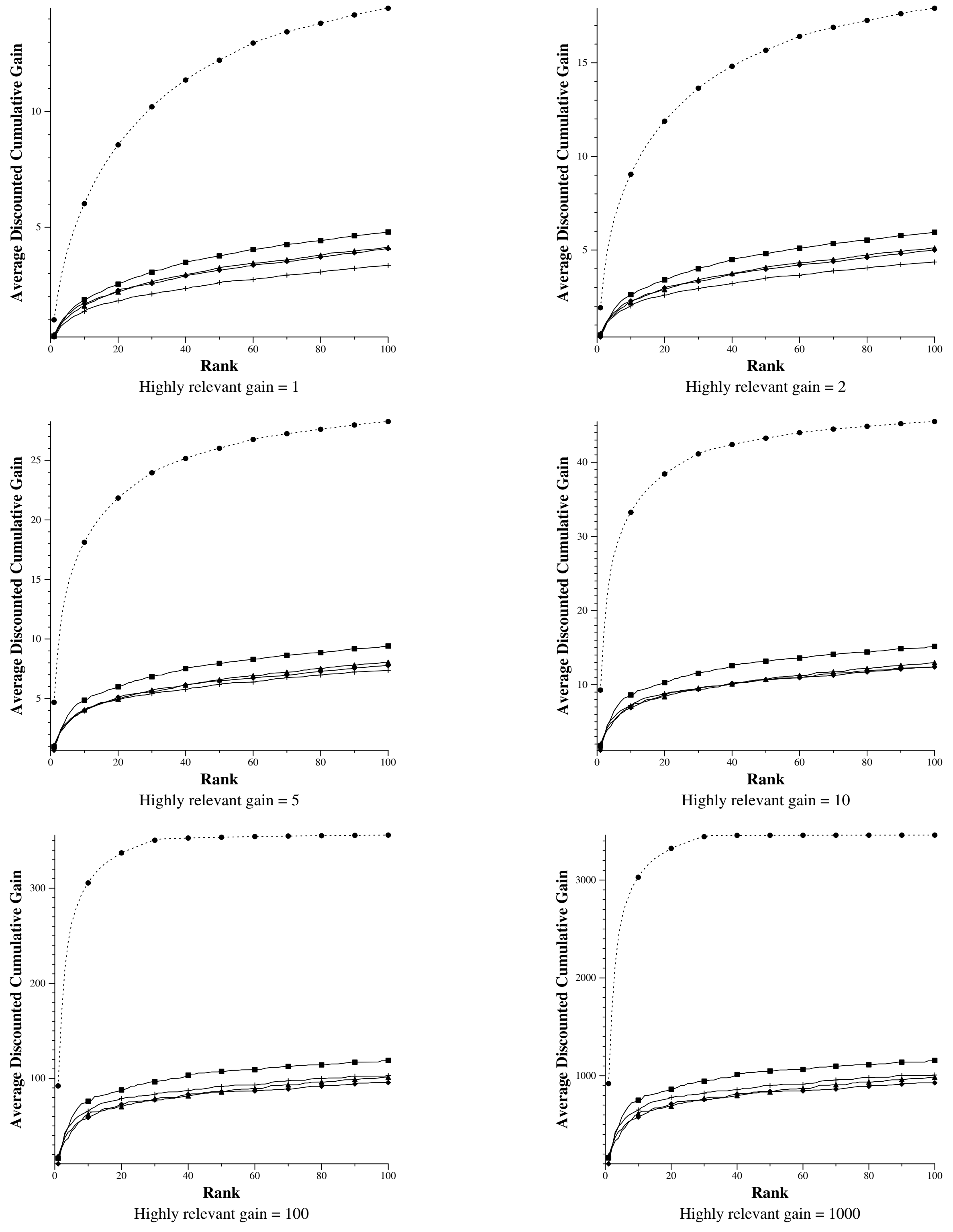

Figure 3: Mean discounted cumulative gain between ranks 1-100 for the best possible run (dotted line) and a selection of TREC-9 runs. The relative importance of highly relevant documents varies in each plot from being equally important as relevant documents to being 1000 times as important. The run plotted with '+' is the run labeled 'S' in Figure 2. Its position relative to the other runs changes as highly relevant documents assume more importance. The shape of the best possible curve also changes as highly relevant documents increase in importance. Since there are relatively few highly relevant documents, the best possible DCG reaches its limit by 30 documents retrieved when highly relevant documents are much more heavily weighted than generally relevant documents.

Table 2: Kendall τ correlations between system rankings as evaluated using DCG and different ratios of the gain values for highly relevant and relevant documents.

	Average DCG[10]					Average DCG[100]				
	2:1	5:1	10:1	100:1	1000:1	2:1	5:1	10:1	100:1	1000:1
1:1	.9487	.8436	.7641	.7205	.7128	.9692	.8538	.7949	.7231	.7154
2:1		.8949	.8154	.7718	.7641		.8846	.8256	.7538	.7462
5:1			.9205	.8769	.8692			.9410	.8692	.8615
10:1				.9564	.9487				.9282	.9205
100:1					.9923					.9923

Table 3: Kendall τ correlations between system rankings as evaluated using mean average precision, precision at cut-off 10 and average discounted cumulative gain at cutoff 10 for different relevance levels.

	Average DCG[10]					
	1:1	2:1	5:1	10:1	100:1	1000:1
MAP_{R1}	.7436	.7487	.7103	.6564	.6282	.6205
MAP_{R2}	.6974	.7282	.7513	.7179	.7102	.7077
P[10]_{R1}	.9333	.9128	.8333	.7590	.7154	.7077
P[10]_{R2}	.6949	.7410	.8410	.9103	.9282	.9359

documents—the main issue is the reliability of conclusions drawn from best document evaluation.

As described in Section 2, three different TREC assessors selected the best page(s) for a topic. The original assessor selected the page while judging the entire pool for the topic; the other two assessors were given only the documents the original assessor judged relevant or highly relevant (unless there were more than 100 relevant). The secondary assessors knew the original assessor had judged the documents in their pool relevant, but they did not know which documents had been judged highly relevant. In addition to selecting the best page, all three assessors filled out the questionnaire in Figure 1 to explain why they chose the page they did.

There was significant disagreement among the assessors as to the best page for a topic. Since assessors were allowed to choose multiple pages as the best, agreement was defined using set intersection: if the best document selections for two assessors had a non-empty intersection, the two assessors were said to agree. Even with this liberal definition of agreement, all three assessors disagreed with one another for 17 of the 50 topics. All three assessors agreed with one another for only 13 topics, and 5 of those 13 topics had three or fewer documents in its relevant set.

The questionnaire asked the assessor whether it was difficult to select a best page. Slightly more than half the time (80/150 or 54%) the assessor answered no. When the answer was yes, the most common reasons given were that there was no particularly good document (20 responses), that there was a set of very similar documents (20 responses), and that different documents were better in different ways (15 responses). The questionnaire also asked the assessor how their selection of best page compared to other pages on various dimensions. No particular pattern is apparent in the responses. As a rule, each assessor rated their best page better than the other pages on all dimensions (except recency, which was more frequently "Can't tell" as expected given the nature of the document set), including the topics for which assessors had chosen different best pages. The comments section makes it clear that different assessors were looking for different things in their best pages. For example, topic 461 asked for documents that discuss the origin or operation of lava lamps. One assessor chose document WTX004-B02-375 as the best page because "It explained the scientific principle behind the connection current of the rising lava.", but admitted document WTX091-B19-269 was a close second that "focuses on the reasons for the resurgence of interest in the lamps...[and] addresses the probable origin question". Another assessor chose WTX091-B19-269 as the best page, commenting

> "Interesting, some discussion of history and comments from people who own them now. Only other choice was technical/boring discussion on principle behind lava lamps."

Unlike traditional retrieval system evaluation, best document evaluation is not robust against the changes caused by using different assessors to select best documents. We used both mean average precision and precision at cut-off 10 to evaluate the TREC-9 title-only runs. When there is exactly one best document per topic, MAP is mathematically equivalent to the mean of the reciprocal of the rank at which the best document is found, and P[10] is the percentage of times the best document appears in the top 10 documents. The usual definitions of the measures gracefully accommodate the multiple best documents selected by some assessors. The average pairwise Kendall τ correlation between system rankings produced using different relevance assessors is .7368 for MAP and .7761 for P[10], demonstrating considerable variation in system rankings as relevance assessors change.

Researchers at AltaVista have classified web search requests into three broad categories: informational, navigational, and transactional [13]. A user with an informational request is assumed to be seeking information as in a traditional IR query, while a navigational request seeks a specific page on the web (e.g., using the search request "TREC" to get to the TREC home page) and a transactional request seeks a web page where a specific function can be performed (e.g., buying flowers). The 50 TREC-9 web track topics were all informational requests. It is likely that there would be more agreement among assessors as to the best page for navigational requests than for informational requests[3]. Additional test collections that incorporate these other request types are required to determine if best page evaluation is appropriate for them. Nevertheless, it is clear that best page evaluation is not viable for informational requests—and real users do make informational requests on the web.

5. CONCLUSION

The world wide web continues to grow in importance as a source of information, and it is therefore increasingly important that appropriate retrieval technologies be developed for the web. Such development is facilitated by infrastructure that permits reliable evaluation of search techniques [8]. This infrastructure includes test collections that adequately model distinctive portions of the task and evaluation metrics that distinguish better retrieval quality from poorer retrieval quality.

This paper addressed one aspect of web retrieval that has been a concern in the web community: the desire to evaluate search engines by their ability to retrieve highly relevant documents. The experiments reported in the paper show that the relative effectiveness of retrieval systems does change when evaluated by highly relevant documents as compared to when evaluated by generally relevant documents. Since each topic was judged by only one assessor, the effect of inter-assessor disagreements could not be quantified for the highly relevant document evaluation. However, retrieval measures are known to be unstable when computed over small numbers of relevant documents. The gain values used in the discounted cumulative gain measure can be manipulated to reward systems for retrieving highly relevant documents while exploiting all known relevance information, thereby increasing stability.

Multiple assessors selected the best document for each topic. Assessors frequently disagreed on which document was best, and the relative effectiveness of systems when evaluated by different assessors changed markedly. Thus, consistent evaluation of systems by their ability to retrieve the best document is not possible for informational requests.

While web searching is similar in many respects to searching any large, heterogeneous collection, it is distinctive in other respects. The massive link structure is one such distinctive feature. Another distinctive feature is the lack of editorial control. The lack of editorial control not only means that the quality of information varies dramatically, but also means that the text contained within a document is not necessarily a reliable indicator as to what a document is about (spam) [3]. Web pages are also becoming increasingly active. The literal content of a page may be a script that generates what the user sees based on arbitrary other data, possibly including customization such that different users see different content. How best to incorporate these aspects of web searching into a common evaluation methodology is the subject of future work.

[3]Navigational requests can be ambiguous: is the TREC home page the home page of the Text REtrieval Conference (`http://trec.nist.gov`) or of the Texas Real Estate Commission (`http://www.trec.state.tx.us`)?

6. REFERENCES

[1] Internet Archive. The Internet Archive: Building an 'Internet Library'. `http://www.archive.org`.

[2] Pia Borlund and Peter Ingwersen. Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In W. Bruce Croft, Alistair Moffat, C.J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 324–331, Melbourne, Australia, August 1998. ACM Press, New York.

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypretextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, `http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm`, 1998.

[4] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In N. Belkin, P. Ingwersen, and M.K. Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Rertrieval*, pages 33–40, 2000.

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

[6] Michael Gordon and Praveen Pathak. Finding information on the world wide web: the retrieval effectiveness of search engines. *Information Processing and Management*, 35:141–180, 1999.

[7] David Hawking, Nick Craswell, and Paul Thistlewaite. Overview of TREC-7 very large collection track. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 91–103, August 1999. NIST Special Publication 500-242. Electronic version available at http://trec.nist.gov/pubs.html.

[8] David Hawking, Nick Craswell, Paul Thistlewaite, and Donna Harman. Results and challenges in web search evaluation. In *Proceedings of the Eighth International World Wide Web Conference*, `http://www8.org/w8-papers/2c-search-discover/results/results.html`, 1999.

[9] David Hawking, Ellen Voorhees, Nick Craswell, and Peter Bailey. Overview of the TREC-8 web track. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 131–150, 2000. NIST Special Publication 500-246. Electronic version available at http://trec.nist.gov/pubs.html.

[10] Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, July 2000.

[11] Chris Sherman. Special report—the 5th annual search engines meeting. http://websearch.about.com/internet/websearch/library/blsem.htm. Section "The Fireworks Fly".

[12] Alan Stuart. Kendall's tau. In Samuel Kotz and Norman L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 4, pages 367–369. John Wiley & Sons, 1983.

[13] Bob Travis and Andrei Broder. The need behind the query: Web search vs classic information retrieval. http://www.infonortics.com/searchengines/sh01/slides-01/sh01pro.html.

[14] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

[15] Ellen M. Voorhees and Donna Harman. Overview of the sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, 36(1):3–35, January 2000.