# Empirical Justification of the Gain and Discount Function for nDCG

Evangelos Kanoulas *
e.kanoulas@shef.ac.uk
Department of Information Studies
University of Sheffield
Regent Court, 211 Portobello Street
Sheffield S1 4DP, UK

Javed A. Aslam
jaa@ccs.neu.edu
College of Computer & Information Science
Northeastern University
360 Huntington Ave, #202 WVH
Boston, MA 02115, USA

## ABSTRACT

The nDCG measure has proven to be a popular measure of retrieval effectiveness utilizing graded relevance judgments. However, a number of different instantiations of nDCG exist, depending on the arbitrary definition of the gain and discount functions used (1) to dictate the relative value of documents of different relevance grades and (2) to weight the importance of gain values at different ranks, respectively.

In this work we discuss how to empirically derive a gain and discount function that optimizes the efficiency or stability of nDCG. First, we describe a variance decomposition analysis framework and an optimization procedure utilized to find the efficiency- or stability-optimal gain and discount functions. Then we use TREC data sets to compare the optimal gain and discount functions to the ones that have appeared in the IR literature with respect to (a) the efficiency of the evaluation, (b) the induced ranking of systems, and (c) the discriminative power of the resulting nDCG measure.

**Categories and Subject Descriptors:** H. Information Systems; H.3 Information Storage and Retrieval; H.3.3 Information Search and Retrieval:Retrieval models

**General Terms:** Experimentation, Measurement, Reliability

**Keywords:** Evaluation, nDCG, Generalizability Theory

## 1. INTRODUCTION

The evaluation of retrieval systems has been a significant area of research in IR. Evaluation measures play a critical role in the development of retrieval systems either as metrics in comparative evaluation experiments, or as objective functions to be optimized in a learning-to-rank fashion. Due to their importance, dozens of measures have appeared in IR literature, with average precision being the dominant one.

One of the main criticism traditional evaluation measures, such as average precision, have received is due to the assumption they make that retrieved documents can be considered as either relevant or non-relevant to a user's request. In other words, traditional measures treat documents of different degrees of relevance as equally important. Naturally, however, some documents are more relevant to a user's request than others and therefore more valuable to a user than others.

The nDCG measure [10, 9] has proven to be one of the most popular measures of retrieval effectiveness that utilizes graded relevance judgments. The underline model of user search behavior on which nDCG is based makes two assumptions: (1) highly relevant documents are more valuable to the user than marginally relevant documents, and (2) the greater the rank a relevant document appears the less valuable to the user that document is.

In the framework used to define nDCG, first relevance scores are mapped to relevance grades, e.g. a score of 3 is given to highly relevant documents, a score of 2 to fairly relevant documents and so on. Relevance scores are viewed as the *gain* returned to a user when examining the document. Thus, the relative value of relevance scores dictates how much more valuable for instance a highly relevant document is to a user than a marginally relevant. Even though, relevance scores were used directly as gains when nDCG was originally introduced, alternative gain functions that map gain values to relevance scores have appeared in the literature. To account for late arrival of relevant documents, gains are then discounted by a function of the rank. The discount function is viewed as a measure of the patience of a user to step down the ranked list of documents. As in the case of gains, a number of different discount functions has appeared in the literature. The discounted gains are then summed progressively from rank 1 to $k$ and this discounted cumulative gain is normalized to range from 0 to 1, resulting in the *normalized discounted cumulative gain* (nDCG).

Hence, nDCG can be considered as a functional of a gain and a discount function. By utilizing different gain and discount functions one is able to accommodate different user search behavior patterns on different retrieval scenarios.

Even though nDCG offers a flexible family of measures so far the selection of the gain and discount functions has been done rather arbitrarily, based on speculations of the

*We gratefully acknowledge the support provided by NSF grants IIS-0533625 and IIS-0534482 and by the European Commission who funded parts of this research within the Tripod project under contract number IST-FP6-045335.


*CIKM'09,* November 2–6, 2009, Hong Kong, China.

search behavior of an average user and speculations of the correlation of the measure to user satisfaction. For instance, Burges et al. [7], introduced an exponential gain function ($2^{\text{rel(r)}} - 1$, where rel(r) is the relevance score of the document at rank r) to express the fact that the gain of a highly relevant document is not just twice the gain of a relevant one for an average user. Further, the logarithmic discount function ($1/log(r+1)$) dominated the literature compared to the Zipfian one ($1/r$) based on the speculation that the gain a user obtains by moving down the ranked list of documents does not drop as sharply as indicated by the Zipfian discount.

Despite these reasonable speculations, Al-Maskari et al. [1] exhibited that cumulative gain without discounting (CG) is more correlated to user satisfaction than discounted cumulative gain (DCG) and nDCG (at least when computed at rank 100). This result not only questions the overall utility of the discount function but most importantly underlines the need for a methodological selection of gain and discount functions. However, given the infinite number of possible gain and discount functions, the vast differences in user search behavior, the many possible retrieval tasks and the difficulty in measuring user satisfaction, a complete analysis of the different gain and discount functions with respect to user satisfaction is prohibitively expensive, if at all possible.

In order to methodologically reduce the space of possible gain and discount functions, a number of correlation studies has been conducted to examine whether different functions lead to equivalent nDCG measures regarding the induced ranking of systems. Voorhees [14] utilized nDCG in TREC Web Track evaluation weighting highly relevant documents by factors 1 to 1000 in relation to marginally relevant documents and she concluded that varying the gain function leads to different ranking of systems. In a similar study, Kekäläinen [11] also examined how different weighting schemes of relevance scores affect the ranking of systems and similarly to Voorhees [14] concluded that the larger the relative difference between relevance grades, the more the ranking of systems deviates from that in the binary case. Furthermore, by comparing the rankings of systems induced by the discounted cumulative gain (DCG) and cumulative gain without discounting (CG), she demonstrated that discounting the gain values also alters the induced ranking of systems.

Given the fact that nDCG variations result in different rankings of systems and thus they evaluate different aspects of retrieval effectiveness along with the difficulty in studying gain and discount functions with respect to user satisfaction, one can compare different variations of nDCG based on other desirable properties of the resulting measure. For instance for different gain and discount functions, one can investigate how informative the resulting variations of nDCG are, i.e. how well do they summarize the relevance of the underline ranked list of documents [2], how discriminative they are, i.e. how well do they discriminate good from bad systems [12], or how stable they are, i.e. how different the rankings of systems are over different sets of queries [6]. Sakai [13] compared the effect of a number of different gain and discount functions on the discriminative power of nDCG.

In this paper, we adopt the variance component analysis framework proposed by Bodoff and Li [4] to measure the stability/efficiency of the resulting nDCG measure when different gain and discount functions are utilized. Based on this framework, we define a stability- or efficiency-optimal gain function by treating gain values of relevance grades as unknown variables and optimizing for the aforementioned stability/efficiency measure. We compare the resulting function to both the linear and the exponential variates that have appeared in the literature, both in terms of stability/efficiency and induced rankings of systems. Similarly, we also define a stability- or efficiency-optimal discount function and compare it against the Zipfian, the log and a linear function. Further, we also define a Pareto optimal combination of gain and discount function, i.e. the combination of gain and discount function that maximizes the minimum stability. Finally, we explore whether the stability- (efficiency-) optimal gain and discount functions lead also to an nDCG measure with high discriminative power [12].

The rest of the paper is organized as follows: In Section 2 we describe the methodology used to obtain stability- or efficiency-optimal gain and discount functions, in Section 3 we present the results of our methodology and conclude in Section 4.

## 2. METHODOLOGY

In this section, we describe a methodology to numerically derive stability- (efficiency-) optimal gain and discount functions. First, we adopt the methodology used by Bodoff and Li [4] to assess the reliability of an IR test collection.

Given an evaluation measure, a number of retrieval systems, a set of queries and a document collection, Bodoff and Li considered two sources of variability in the observed system scores of the retrieval systems when they are run over the given queries, (a) the actual performance differences between systems, and (b) differences in the nature of the queries themselves. This way, Bodoff and Li [4], quantified the quality of the test collection as the proportion of the total variability observed in the scores of the retrieval systems that is due to actual performance differences among these systems.

In a similar manner, different sources of variability can be considered and quantified. For instance, earlier than Bodoff and Li, Banks et al. [3] considered, as an additional to the systems and queries source of variability, the judges that assess the relevance of the documents to the queries.

In this work, we consider the evaluation measure itself as a source of variability. In particular, given a number of retrieval systems, a set of queries and a document corpus, we consider gain and discount function of nDCG as unknown variables and we select the ones that maximize the proportion of variability due to actual performance differences among systems. The proportion of variability reflects the stability of the evaluation measure, and thus by maximizing this proportion we maximize the stability of the measure. Furthermore, the more stable a measure is the less queries it requires to reliably evaluate the retrieval systems. Thus, by maximizing stability we also maximize efficiency in terms of required queries.

We numerically computed the stability optimal gain and discount function by employing (a) variance decomposition analysis of the nDCG scores [3, 4, 5] and (b) optimization. In the following subsections, we describe both components of our methodology in details.

### 2.1 Variance component analysis

Assume an experimental design that involves $n_s$ systems run over a sample of $n_q$ queries resulting in a set of $n_s * n_q$

ranked lists of documents. Further assume that each list of documents is evaluated by nDCG and the overall quality of a system is captured by averaging the nDCG values over all topics. Systems, then, are ranked by their mean scores, i.e. mean nDCG.

Hypothetically, if a second set of topics was available, the systems could be run over this new set of topics and new mean nDCG scores (and consequently new ranking of the systems) would be produced. The question that naturally arises is, how many topics are necessary to guarantee that the mean nDCG scores do not change radically when two different query sets are used, or alternatively how many topics are necessary to guarantee that the mean nDCG scores of systems reflect their actual performance?

Given different sets of topics one could decompose the amount of variability that occurs in mean nDCG scores (as measured by variance) across all sets of topics and all systems into three components: (a) variance due to actual performance differences among systems—*system variance*, (b) variance due to the relative difficulty of a particular set of topics—*topic variance*, and (c) variance due to the fact that different systems consider different sets of topics hard (or easy)—*system-topics interaction variance*. Note that among the three variance components, only the variance due to *systems* and *system-topics interactions* affect the *ranking* of systems—it is these two components that can alter the relative differences among mean nDCG scores, while the topic variance will affect all systems equally, reflecting the overall difficulty of the set of topics.

Ideally, one would like the total variance in mean nDCG scores to be due to the actual performance differences between systems rather than the other two sources of variance. If this would be the case, running the systems over different topic sets would result in each system having identical mean nDCG scores regardless of the topics used, and thus mean nDCG scores over a single set of topics would be 100% reliable in evaluating the quality of the systems.

In practice, retrieval systems are run over a single given set of topics. The decomposition of the total variance into the aforementioned components in this case can be realized by fitting an ANOVA model into nDCG scores [3, 4, 8]. Given the variance components tools from Generalizability Theory [5] can be used to quantify the stability of the evaluation.

## 2.2 Stability coefficients

There are two coefficients that predominate in Generalizability Theory to quantify the stability of the evaluation, the generalizability coefficient and the dependability coefficient, with the former reflecting the stability of the system rankings and the latter the stability of the system effectiveness scores. They both lie in a zero to one range.

The former coefficient is the ratio of the system variance and the variance in relative nDCG scores (i.e. in system rankings), that is the summation of the system and system-topic interaction variance,

$$E\rho^2 = \frac{\sigma^2(\text{system})}{\sigma^2(\text{system}) + \frac{\sigma^2(\text{system:topic})}{\#\text{ of topics}}} \quad (1)$$

and it can be interpreted as an approximation to the squared correlation between the relative mean nDCG scores observed over the given set of topics and the relative mean nDCG scores that would be observed if infinite number of topics was available.

The dependability coefficient, $\Phi$, is the ratio of the system variance and the total variance,

$$\Phi = \frac{\sigma^2(\text{system})}{\sigma^2(\text{system}) + \frac{\sigma^2(\text{topic})+\sigma^2(\text{system:topic})}{\#\text{ of topics}}} \quad (2)$$

and it can be interpreted as an approximation to the squared correlation between the mean nDCG scores observed over the given set of topics and the mean nDCG scores that would be observed if infinite number of topics was available. Note that both $\Phi$ and $E\rho^2$ decrease with the topic set size. Further note that $E\rho^2$ is always larger than $\Phi$. In our experiments we employ only the latter coefficient since stable scores infer stable rankings.

Also note that the computation of the two coefficients is done independently of the estimation of the variance components. That is, first the variance components are estimated over a set of available topics (50 topics in our experiments). Then, the two aforementioned coefficients are using these estimates to project reliability scores to topic sets of any size. The topic set size in the computation of the coefficients does not need to be the same as the topic set size used to estimate the variance components (it can even be larger).

As mentioned before, in this work we consider the gain values for different relevance grades and the discount factors for different ranks used in nDCG as unknown variables. Given a fixed-size topic set we would like to obtain the gain values/discount factors that maximize the stability of the mean nDCG scores of the systems.

## 2.3 Optimization

In the optimization process employed, we use $\Phi$ as the objective function to maximize with respect to the gain values/discount factors employed in nDCG.

Note that nDCG is a scale-free measure with respect to both the gain values and the discount factors in the sense that multiplying either the gain or the discount with any number does not affect the nDCG score. For this reason, we enforced the gain values to be a probability distribution over relevance grades and the discount factors to be a probability distribution over ranks. This way we limit the range of values both for the gain and the discount within the $[0, 1]$ range and reduce the unknown parameters by one. Furthermore, it so happens that there maybe some fluctuation in the values of the optimal discount factors, e.g. the discount factor on a certain rank may happen to be larger than the one on a lower rank. This is not justifiable from an IR perspective and thus, we also enforce that the discount factors are non-increasing with the rank. The same may be true for the gain values, hence we enforce them to be non-decreasing with the relevance grade. Further, we set the gain value for non-relevant documents equal to zero.

Moreover, note that the coefficient $\Phi$ in Equation 2 is a monotonically non-decreasing function of the number of queries. In other words, the gain or discount function that is optimal for $n$ queries is also optimal for $n + 1$ queries. Therefore, in the optimization process we set the number of queries equal to 1.

The optimization setup for the gain/discount function is mathematically expressed in Figure 1. When we optimize for the discount factors we consider the gain values as given,

$$\underset{\{gain(grade_j)\}}{\text{argmax}} \frac{\sigma^2(\text{sys})}{\sigma^2(\text{sys}) + \sigma^2(\text{topic}) + \sigma^2(\text{sys:topic})}$$

Subject to:

1. $\sum_{j=1}^{k} gain(grade_j) = 1$
2. $gain(grade_j) - gain(grade_{j+1}) \leq 0 \forall j : 1 \leq j \leq k-1$

where $k$ is the number of relevance grades.

$$\underset{\{disc(rank_r)\}}{\text{argmax}} \frac{\sigma^2(\text{sys})}{\sigma^2(\text{sys}) + \sigma^2(\text{topic}) + \sigma^2(\text{sys:topic})}$$

Subject to:

1. $\sum_{r=1}^{r} disc(rank_r) = 1$
2. $disc(rank_r) - disc(rank_{r+1}) \leq 0 \forall j : 1 \geq j \leq N-1$

where $N$ is the cut-off rank at which nDCG is calculated.

**Figure 1: Optimization setup for gain values and discount factors, respectively.**

while when we optimize for gain values we consider the discount factors as given. We also perform a multi-objective optimization to simultaneously optimize for the gain and the discount function. For the purpose of the optimization, we used the **fmincon** MATLAB function for the normal optimization and the **minimax** MATLAB function for the multi-objective optimization. Both functions employ Sequential Quadratic Optimization.

# 3. RESULTS

The afore-described optimization framework was applied to the TREC 9 and 10 Web track collections and the TREC 12 Robust track collection. The number of participating systems for the three collections is 105, 97 and 78 respectively. All systems were run over 50 queries[1]. Documents returned as a respond to these queries were judged by a single assessor in 3 relevance grades scale: highly relevant, relevant and non-relevant. The task in all tracks was the usual ad-hoc retrieval task.

For each one of the three test collections, we calculated the optimal discount factors (given a linear gain function), the optimal gain values (given a logarithmic discount function) and the Pareto-optimal gain values and discount factors. We compared the optimal gain and discount functions with the a number of commonly used gain and discount functions both with respect to the stability/efficiency of the resulting nDCG measure and with respect to the induced by the resulting nDCG ranking of systems.

## 3.1 Optimal discount function

In this section we present the results of the optimization for the discount function. The gain values were set to 0, 1 and 2 for non-relevant, relevant and highly relevant documents respectively and they were treated as constants during the optimization. We performed the optimization over the TREC 9, 10 and 12 data sets for nDCG computed at rank 10, 20 and 100 and we report the results in Figure 2. We compare the optimal discount factors – blue solid curve with circles as markers in the figure – (a) with the Zipfian discount function (1/rank) – green solid curve with plus signs as markers, (b) with the log discount function ($1/\log_2(1+\text{rank})$) – dark blue solid curve with triangles as markers and (c) with the linear discount function ( (cut-off rank + 1 - rank) / cut-off rank) – magenta solid curve with crosses as markers. For comparison purposes, we transformed the linear, log and Zipfian discount factors to probability distributions over the ranks.

As it can observed in Figure 2, the optimal discount function is the least steep one among the discount function considered. The log discount function is the one closest to the optimal, while the Zipfian drops much faster than the optimal. The linear discount also appears to be close to the optimal one, at least when only the top ranks are considered..

Looking at the right-most plots for each TREC, that is the plots corresponding to nDCG at rank 100, one can observe that the top ranks are the ones that mainly matter and thus they are given higher discount factor, while the rest of the ranks are given a rather small and constant discount factor. The number of the top-most ranks that really matter seems to be collection dependent, with the top 10 ranks being the important ones for TREC 9 and 12 and the top 20 ranks being the important ones for TREC 10. A further observation one can make is that, even though the rest of the ranks are given a rather constant discount factor, this constant is far from zero (or at least farther than the discount factors the log and the linear discount function assigns to those ranks) suggesting that documents lower at the ranked list may also be useful in discriminating systems efficiently. This further suggests that computing nDCG at top ranks is sub-optimal since computing nDCG at some cut-off rank implicitly assigns zero discount factors to the ranks below that cut-off.

For the purpose of completeness, Figure 3 illustrates the results when we optimized the stability (efficiency) of nDCG without enforcing the non-increasing constraint for the discount factors. We only report results for nDCG at rank 20. One may observe that the optimal un-constraint discount factors are not strictly decreasing with the rank. Intuitively, these fluctuations are due to the fact that often times similar systems return relevant documents at the top ranks and thus the only way to discriminate them is by looking deeper in the ranked list of documents. Thus, once again, this indicates that lower ranks may very well help in discriminating systems.

Figure 4 illustrates the stability of the nDCG measure (i.e. the fraction of the variance in the mean nDCG values due to actual performance differences between systems) when computed using (a) the optimal, (b) the log, (c) the Zipfian, and (d) the linear discount function. As expected, the optimal discount function eliminates all variance components other

[1]The TREC 12 Robust track collection includes 100 queries, however the first 50 of them were obtained from TREC 6, 7 and 8, where documents were judged as either relevant or non-relevant. For this reason, we did not use these 50 queries in our studies.

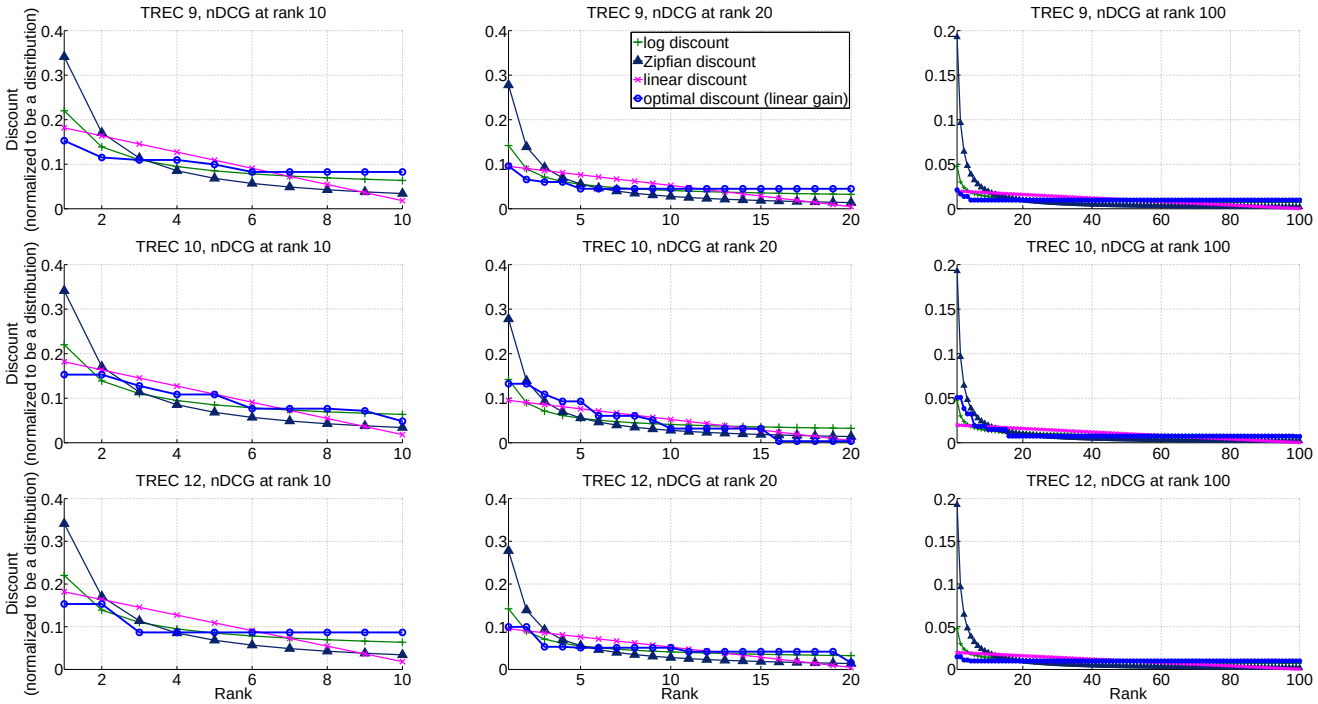


**Figure 2: Discount factors at each rank for (a) the optimal discount function, (b) the Zipfian discount function (1/rank), (c) the log discount function ($1/\log_2(1+\text{rank})$) and the linear discount function ((cut-off rank+1-rank)/(cut-off rank)).**

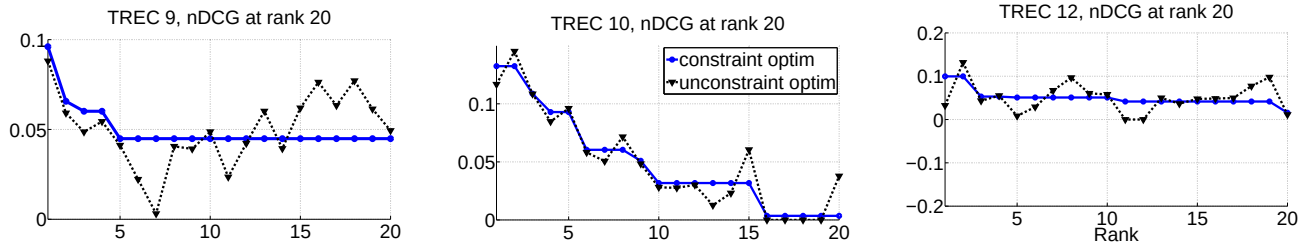


**Figure 3: Discount factors for the optimal un-constraint discount function. The optimal constraint discount function is also included for comparison purposes.**

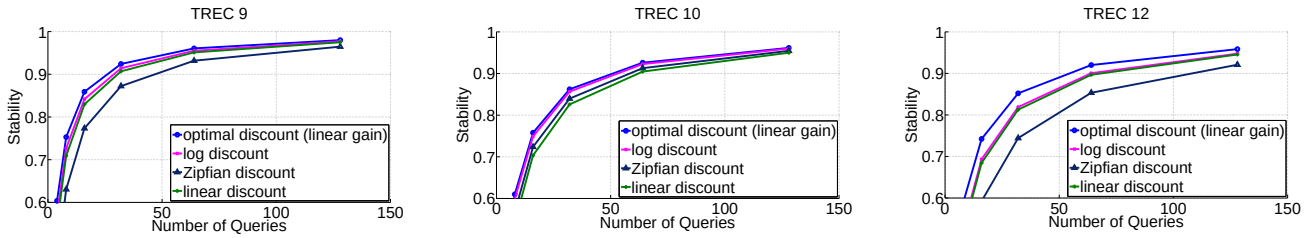


**Figure 4: The stability of nDCG at rank 100 (or the fraction of the variance in the mean nDCG values due to actual performance of system) when nDCG is computed with the (a) optimal, (b) log, (c) Zipfian, and (d) linear discount function.**

than the one due to systems, faster (in terms of queries) than the rest of the discount functions. The log discount function is the second most stable one while the Zipfian and the linear lead to the least stable nDCG measure.

Finally, in Table 1 we compare the efficiency of the nDCG when the optimal discount function is employed with the efficiency of nDCG when the log, the Zipfian and the linear discount functions are employed. To calculate the efficiency, we fit an ANOVA model into the resulting nDCG scores, for each one of the discount functions. Then, setting the value of $\Phi$ equal to 0.9, that is 90% of the total variance in the nDCG scores being due to the actual performance differences between systems, and using Equation 2 we computed the necessary number of queries to reach the given stability. As expected the log discount function is the closest to the optimal one.

To conclude, the stability- (efficiency-) optimal discount function is less steep than any of the commonly used discount functions. The widely used log discount function is the one closest to the optimal discount function, while the Zipfian and the linear ones are the least stable. Furthermore, the optimal discount factors over low ranks are far from zero which suggests that looking further down at the ranked list of documents (regardless of the underline user search behavior and patience to step down the ranked list) can improve the reliability of system comparisons.

**Table 1: Number of queries required to achieve 0.95 stability in evaluation.**

| $\Phi \geq 0.95$ | Zipfian | linear | log | **Optimal** |
|---|---|---|---|---|
| TREC 9 | 45 | 31 | 29 | **25** |
| TREC 10 | 58 | 64 | 51 | **49** |
| TREC 12 | 104 | 70 | 67 | **53** |

## 3.2 Optimal gain function

We also performed an optimization for the gain values assigned to the different relevance grades of documents. In this case, the discount factors were treated as constants. The log discount function, the closest to the stability- (efficiency-) optimal discount function, was utilized. Further, we set the gain value for the non-relevant equal to zero and optimized for the gain values of the relevant and highly relevant documents. We performed the optimization over TREC 9, 10 and 12 data sets for nDCG at ranks 3, 10, 20, 100 and 200. Instead of the gain values themselves, we report the ratio between the gain value assigned to the highly relevant documents and the gain value assigned to the relevant ones. The results can be viewed in Table 2. As in the case of the discount function, we performed both an un-constraint and a constraint optimization. In the constraint optimization we enforced the gain value of the highly relevant documents to be greater than or equal to the gain value of the relevant ones. The optimal gain value ratios for the un-constraint optimization are reported in the first column of Table 2, while the ones for the constraint optimization are reported in the second column. The last two columns show the ratio of the gain values when the linear and exponential gain functions are utilized.

By comparing the first two with the last two columns of the table one can observe that the utility of relevant documents in comparative evaluation of systems is underrated by the commonly employed gain functions. The optimal ratio of the gain values for highly relevant and relevant documents is in most of the cases much smaller than 2 or 3. Intuitively, this means that relevant documents are almost equally discriminative to the highly relevant ones. Good systems will retrieve both highly relevant and relevant documents while bad systems will have difficulties in retrieving either highly relevant or relevant documents. Thus, discriminating systems regarding their performance can be similarly done with either relevant or highly relevant documents. Note that this is true for the particular TREC collections under study and the systems run over these collections and it may not be true in the general case.

In the un-constraint optimization column, highly relevant documents still appear more discriminative than relevant documents for most of the cases. However, there are cases, e.g. in TREC 12 with nDCG computed at low ranks, that relevant documents appear to be more discriminative than highly relevant documents. An intuitive explanation of this behavior may be given by fact that the total number of highly relevant documents retrieved by systems in TREC 12 is quite small and highly relevant documents tend to appear at the very top of the ranked lists, while they are almost absent from the deeper ranks. Thus when deeper ranks are considered, highly relevant documents lose some of their discriminative power. The percentage of relevant and highly relevant documents on average (over all queries) at each rank for TREC 12 can be viewed in Figure 5.

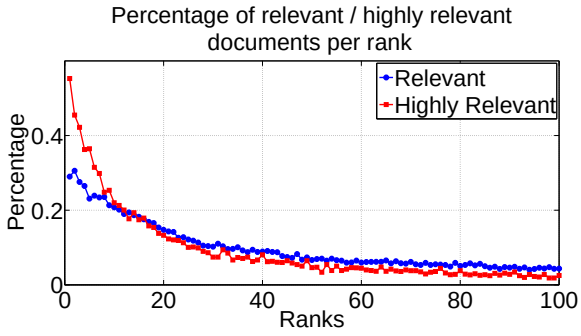


**Figure 5: The percentage of documents that are relevant and the percentage of documents that are highly relevant on average (over all queries) at each rank for TREC 12.**

Finally, for both TREC 9 and 10, one can observe a trend in the optimal ratio between the grades for relevant and highly relevant documents, with the ratios originally increasing by the rank nDCG is computed at and then dropping. This phenomenon needs to be further explored.

In Table 3 we compare the efficiency of the nDCG measure calculated at rank 100 when the optimal gain function is employed with the efficiency when the linear or the exponential gain functions are employed. As in the case of discount functions, to calculate the efficiency of each measure, we fit the ANOVA model into the resulting nDCG scores, for each on of the discount functions. Then, setting the value of $\Phi$ equal to 0.95, that is 95% of the total variance in the nDCG scores is due to the actual performance differences between systems, and using Equation 2 we compute the necessary number of queries to reach the given stability.

**Table 2: Ratio among the gain values of highly relevant and relevant documents for TREC 9, 10 and 12**

| TREC9 | optimal ratio (unconstraint) | optimal ratio (log discount) | optimal ratio (optimal discount) | hrel/rel | $(2^{\text{hrel}}-1)/(2^{\text{rel}}-1)$ |
|---|---|---|---|---|---|
| nDCG@3 | 1.1 | 1.1 | 1.1 | 2 | 3 |
| nDCG@10 | 1.3 | 1.3 | 1.4 | 2 | 3 |
| nDCG@20 | 1.5 | 1.5 | 1.6 | 2 | 3 |
| nDCG@100 | 1.2 | 1.2 | 1.9 | 2 | 3 |
| nDCG@200 | 1.1 | 1.1 | 2.7 | 2 | 3 |

| TREC10 | optimal ratio (unconstraint) | optimal ratio (log discount) | optimal ratio (optimal discount) | hrel/rel | $(2^{\text{hrel}}-1)/(2^{\text{rel}}-1)$ |
|---|---|---|---|---|---|
| nDCG@3 | 1.2 | 1.2 | 1.3 | 2 | 3 |
| nDCG@10 | 1.6 | 1.6 | 1.8 | 2 | 3 |
| nDCG@20 | 2.0 | 2.0 | 2.0 | 2 | 3 |
| nDCG@100 | 1.8 | 1.8 | 1.8 | 2 | 3 |
| nDCG@200 | 1.5 | 1.5 | 1.6 | 2 | 3 |

| TREC12 | optimal ratio (unconstraint) | optimal ratio (log discount) | optimal ratio (optimal discount) | hrel/rel | $(2^{\text{hrel}}-1)/(2^{\text{rel}}-1)$ |
|---|---|---|---|---|---|
| nDCG@3 | 1.2 | 1.2 | 1.2 | 2 | 3 |
| nDCG@10 | 1.2 | 1.2 | 1.1 | 2 | 3 |
| nDCG@20 | 1.0 | 1.0 | 1.0 | 2 | 3 |
| nDCG@100 | 0.8 | 1.0 | 1.0 | 2 | 3 |
| nDCG@200 | 0.7 | 1.0 | 1.0 | 2 | 3 |

Interestingly, the values in the table are almost identical for all gain functions for TREC 9 and 10, while only for TREC 12 the optimal gain is significantly better than the linear or the exponential ones in terms of efficiency.

Comparing Table 3 with Table 1 one can observe that the choice of the discount function affects much more the efficiency (stability) of the resulting nDCG measure than the choice of the gain function. As mentioned before, intuitively this means that at least in these particular collections when a system is good it retrieves both many highly relevant and many relevant documents, while when a system is bad it fails to retrieve either. Even thought, this is true for the given test collections, this may not be the case for other test collections and in particular for collections with more than three relevance grades, where for instance retrieving enough marginally relevant documents may not necessarily mean that the system can also retrieve enough excellent documents (where excellent is more than 1 relevance grade away from marginally relevant). Unfortunately, currently we do not possess any such collection and thus we leave this as a future work.

**Table 3: Number of queries required to achieve 0.95 stability in evaluation.**

| $\Phi \geq 0.95$ | exp | linear | Optimal |
|---|---|---|---|
| TREC 9 | 28 | 29 | **30** |
| TREC 10 | 52 | 51 | **51** |
| TREC 12 | 72 | 67 | **63** |

## 3.3 Pareto-optimal gain and discount functions

Finally, we performed multi-objective optimization in order to optimize efficiency (stability) for both the gain and the discount functions simultaneously. To do so, we utilized the minimax MATLAB function, which produces the Pareto optimal discount and gain functions. That is, the discount and gain functions that maximize the worst case value of nDCG stability. We performed the optimization over TREC 9, 10 and 12, concluding that the Pareto optimal gain and discount functions are very close to the optimal gain and discount functions when the optimization is done independently for gains and discounts. The multi-objective optimal discount function for TREC 9 when nDCG is computed at rank 20 is shown in Figure 6. For comparison reasons, the optimal discount function when linear gain is used is also shown in the figure. As it can be observed in all cases the discount factors obtained from the multi-objective optimization are almost equal to the ones obtained with linear gains used. The multi-objective optimal ratio between highly relevant and relevant documents is reported in the third column of Table 2. As it can be observed, except for the case of TREC 9, when nDCG is computed at very low ranks, the multi-objective optimal ratio is very close to the one obtained with the log discount function. This may be an indication that gain and discount functions independently affect the stability of the measure. Similar plots are obtained for all TREC's and all ranks nDCG is computed at.

## 3.4 Correlation study

Different gain and discount functions employed in the calculation of nDCG may result in different mean nDCG values and therefore different rankings of the systems. To investigate how gain and discount functions affect the nDCG score and thus the induced ranking of systems, we calculated the mean nDCG at rank 100 for different gain and discount functions and computed the Kendall's $\tau$ between the induced rankings.

The scatter plots in Figure 7 illustrate the mean nDCG

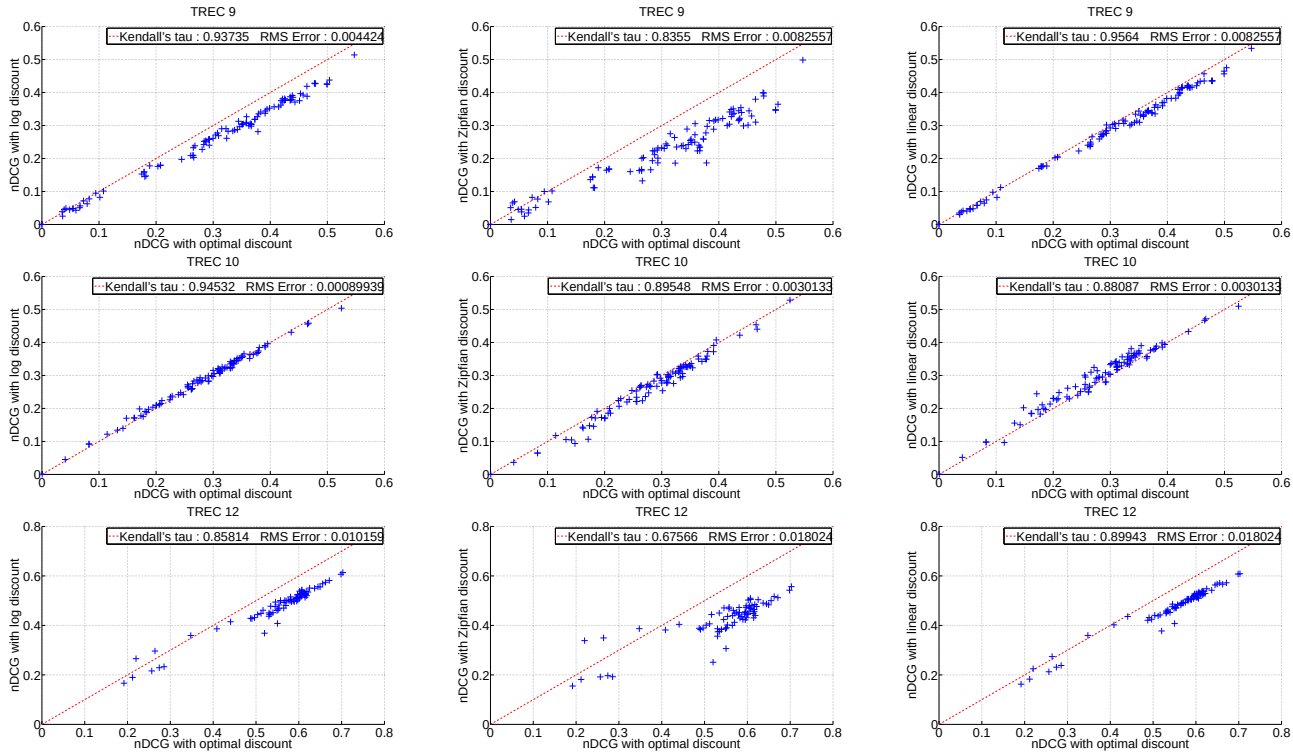


**Figure 7: Scatter plots of the mean nDCG scores for the optimal discount function versus the log, Zipfian and linear discount function.**

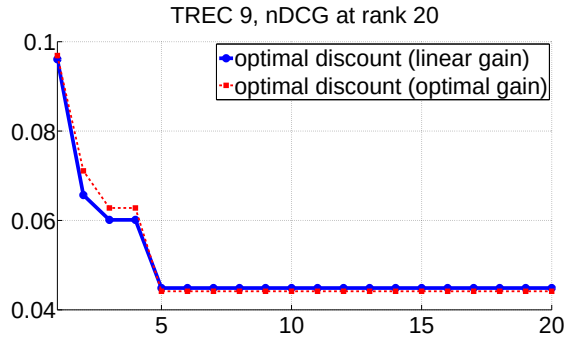


**Figure 6: The multi-objective optimal discount function along with the optimal discount function when linear gains are used for TREC 9 and nDCG computed at rank 20.**

**Table 4: Kentall's $\tau$**

| | | *Zipfian* | *linear* | *log* | Optimal |
|---|---|---|---|---|---|
| TREC 9 | *Zipfian* | 1.0000 | 0.8315 | 0.8960 | 0.8355 |
| | *linear* | 0.8315 | 1.0000 | 0.9282 | 0.9564 |
| | *log* | 0.8960 | 0.9282 | 1.0000 | 0.9374 |
| | Optimal | 0.8355 | 0.9564 | 0.9374 | 1.0000 |
| TREC 10 | *Zipfian* | 1.0000 | 0.7886 | 0.8625 | 0.8955 |
| | *linear* | 0.7886 | 1.0000 | 0.9184 | 0.8809 |
| | *log* | 0.8625 | 0.9184 | 1.0000 | 0.9453 |
| | Optimal | 0.8955 | 0.8809 | 0.9453 | 1.0000 |
| TREC 12 | *Zipfian* | 1.0000 | 0.7136 | 0.8149 | 0.6757 |
| | *linear* | 0.7136 | 1.0000 | 0.8828 | 0.8994 |
| | *log* | 0.8149 | 0.8828 | 1.0000 | 0.8581 |
| | Optimal | 0.6500 | 0.8994 | 0.8581 | 1.0000 |

scores for the optimal discount function ($x-axes$) computed at rank 100 against the mean nDCG scores for the log, Zipfian and linear discount functions respectively ($y-axes$) for TREC 9, 10 and 12. The RMS Error and Kendall's $\tau$ are reported in the plots. The Kendall's $\tau$ between the rankings of systems induced by any two discount functions are also reported in Table 4.

By inspecting both the scatter plots in Figure 7 and the Kendall's $\tau$ values in Table 4 one can observe that both the rankings by the linear discount function and the rankings by the log discount function are very close to the rankings by the optimal discount function. As illustrated in Figure 2, these two discount functions are the closest to the optimal one. The rankings by the Zipfian discount function are widely different than the ones by the optimal discount function, especially in TREC 12.

This wide difference between the induced rankings by the optimal discount function and the Zipfian one can be explained by revisiting Figure 4. As it can be observed, for the Zipfian discount function, only 80% of the differences in the mean nDCG scores over a set of 50 queries (which is the case

in all scatter plots here), is due to actual performance differences between the systems, while the corresponding percentage for the optimal discount function is about 90%. The corresponding percentages for TREC 9 (where the ranking of systems for the two discount functions are closer to each other) are 90% and 95% respectively, while for TREC 10 (where the ranking of systems are almost identical) the percentages are around 88% and 90%, respectively. Therefore, the ranking by the Zipfian discount in TREC 12 incorporates a lot of noise which is reduced in the case of TREC 9 and 10.

The scatter plots in Figure 8 illustrate the mean nDCG scores computed at rank 100 for the optimal gain function ($x-axes$) against the mean nDCG scores for the exponential gain function ($y-axes$) for TREC 9, 10 and 12. The RMS Error and Kendall's $\tau$ are also reported in the plots.

As it can be observed in Figure 8 the rankings by the optimal discount function are almost identical with the rankings by the exponential gain function. This is one more indication that for the particular test collections with the three grades of relevance the ratio between the gain values for relevant and the gain values for highly relevant documents does not affect the ranking of systems (at least for the ratio values examined in our studies, i.e. ratio values less than 3). What is particularly striking is that even for TREC 12, where the optimal gain function gives the exactly same gain value to both relevant and highly relevant, and thus essentially conflates the two relevance grades in one, the Kendall's $\tau$ between the rankings is 0.94, with the top 6-7 systems ranked in the exact same order by both gain functions. This states that good systems do equally good in retrieving relevant and highly relevant documents, while bad systems do equally bad in retrieving either relevant or highly relevant documents.

The corresponding scatter plots for the linear gain function look very similar to the ones in Figure 8 and for this reason they are not reported here.

## 3.5 Discriminative power

As mentioned before, intuitively, efficiency and stability seem to correlate well with discriminative power, since the variability in a measure that discriminates systems well will most probably be due to actual performance differences between systems. In this section we perform some basic experiments to test whether this hypothesis is correct.

Sakai [12] proposed a methodology to compare evaluation methods in terms of their ability to discriminate between systems based on *Bootstrap Hypothesis Tests*. According to his framework, all pairs of systems are considered and the hypothesis that their mean scores over a set of queries are the same is tested. To test this hypothesis Sakai [12] employs a bootstrap test, creating 1000 bootstrap samples. The achieved significance level (ASL), that is the significance level required to reject the zero hypothesis that two systems have the same mean score, is computed for each pair of systems. Finally, evaluation measures are compared in terms of ASLs. The smaller the ASLs a metric achieves the more discriminative the metric is.

To optimize for discriminative power, one would need to minimize the obtained ASLs while treating gain and discount function as unknowns. This is not a trivial optimization and it seems at least computationally inefficient. However, if stability (efficiency) is well correlated with discriminative power, then the stability-optimal nDCG will also demonstrate high discriminative power.

To test out thesis, we adopted the bootstrap hypothesis testing methodology, and compared 4 variations of nDCG, (a) nDCG with optimal gain and optimal discount, (b) nDCG with linear gain and log discount, (c) nDCG with exponential gain and log discount, and (d) nDCG with linear gain and linear discount. We followed the experimental setup in Sakai [12] and used only the top 30 systems from each data set (TREC 9, 10 and 12), since "near-zero" runs are unlikely to be useful for discussing the discriminative power of the measures. We considered all the remaining pairs of systems and for each one of the pairs we created 1000 bootstrap samples and calculated the achieved significance level (ASL) for all aforementioned nDCG measures. Figure 9 illustrates, for each one of the nDCG measures, the ASLs of systems pairs. The horizontal axis represents all system pairs sorted by ASL. The pairs of systems at the left of a given ASL level, are those that the measure cannot discriminate.

As it can be observed from the plots, when the stability-(efficiency-) optimal gain and discount functions are utilized nDCG outperforms all other variations with respect to discriminative power. The linear/exponential gain and log discount nDCG measures appear to be the next most discriminative ones, while the linear gain and linear discount nDCG appears to be the less discriminative one.

# 4. CONCLUSIONS

Despite the flexibility nDCG offers in the selection of the appropriate gain and discount function, so far this selection has been done rather arbitrarily, based on speculations of the search behavior of an average user and speculations of the correlation of the measure to user satisfaction. Recent work [1] has shown that the most commonly employed gain and discount functions are loosely related to user satisfaction which underlines the need for a more methodological selection of gain and discount function. However, given the infinite number of possible gain and discount functions, the vast differences in user search behavior, the many different possible retrieval tasks, a complete analysis of the different gain and discount functions with respect to the user satisfaction is prohibitively expensive, if at all possible.

In this work, we numerically computed a stability- or efficiency-optimal gain and discount function by treating gain values and discount factors as unknowns and optimizing for a stability/efficiency measure defined based on Generalizability theory. We compared the resulting gain function to both the linear and the exponential functions and the resulting discount function to the log, Zipfian and linear ones.

According to our results, the optimal discount function is less steep than all commonly used discount functions, giving reasonably high weights to lower ranks, while the relative difference between gain values is much smaller than the commonly used ones, giving almost equal weights to both relevant and highly relevant documents. The latter was rather striking, since weighting relevant and highly relevant documents equally did not seem to alter the ranking of systems. Note that this is true for the particular collections and systems under study and it may not reflect the general case.

Finally, we demonstrated that the stability- (efficiency-) optimal nDCG measure outperforms the dominant in the literature nDCG measure with respect to discriminative power as well.

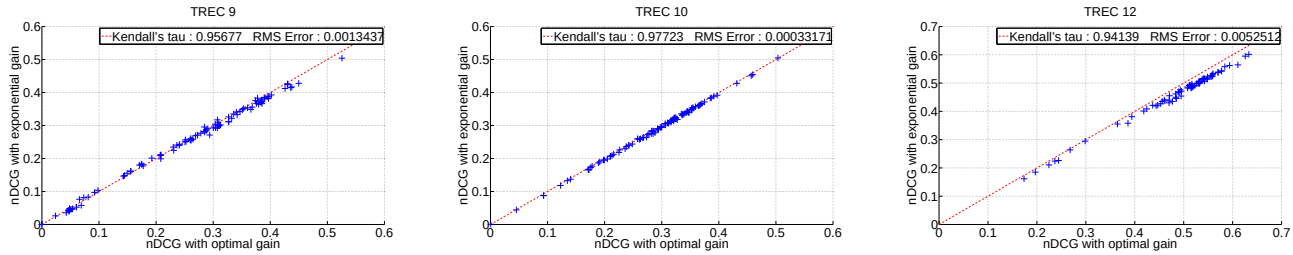


**Figure 8: Scatter plots of the mean nDCG scores for the optimal gain function versus the exponential discount function.**

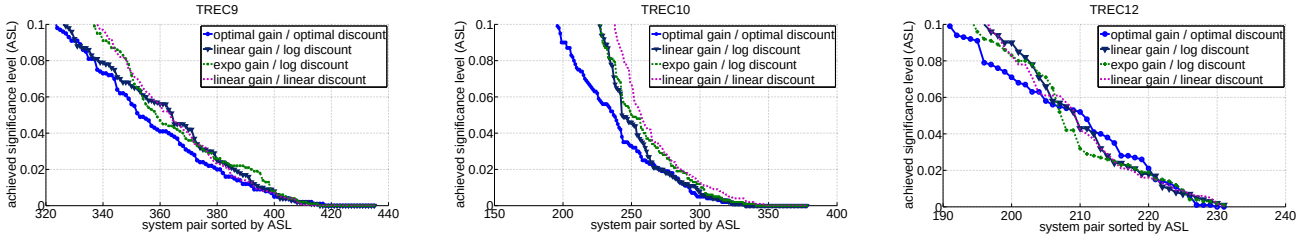


**Figure 9: ASL curves for TREC 9, 10 and 12 with nDCG computed at rank 20.**

# 5. REFERENCES


[1] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between ir effectiveness measures and user satisfaction. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 773–774, New York, NY, USA, 2007. ACM.

[2] J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 27–34, New York, NY, USA, 2005. ACM.

[3] D. Banks, P. Over, and N.-F. Zhang. Blind men and elephants: Six approaches to trec data. *Inf. Retr.*, 1(1-2):7–34, 1999.

[4] D. Bodoff and P. Li. Test theory for assessing ir test collection. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 367–374, 2007.

[5] R. L. Brennan. *Generalizability Theory*. Springer-Verlag, New York, 2001.

[6] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, 2000.

[7] C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 89–96, New York, NY, USA, 2005. ACM.

[8] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation over thousands of queries. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 651–658, New York, NY, USA, 2008. ACM.

[9] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48, New York, NY, USA, 2000. ACM.

[10] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[11] J. Kekäläinen. Binary and graded relevance in ir evaluations: comparison of the effects on ranking of ir systems. *Inf. Process. Manage.*, 41(5):1019–1033, 2005.

[12] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 525–532, New York, NY, USA, 2006. ACM.

[13] T. Sakai. On penalising late arrival of relevant documents in information retrieval evaluation with graded relevance. In *First International Workshop on Evaluating Information Access (EVIA 2007)*, pages 32–43, 2007.

[14] E. M. Voorhees. Evaluation by highly relevant documents. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82, New York, NY, USA, 2001. ACM.