# Expected Reciprocal Rank for Graded Relevance

Olivier Chapelle
Yahoo! Labs
Santa Clara, CA
chap@yahoo-inc.com

Donald Metzler
Yahoo! Labs
Santa Clara, CA
metzler@yahoo-inc.com

Ya Zhang
Yahoo! Labs
Sunnyvale, CA
yazhang@yahoo-inc.com

Pierre Grinspan
Google Inc
San Bruno, CA
pgrinspan@gmail.com

## ABSTRACT

While numerous metrics for information retrieval are available in the case of binary relevance, there is only one commonly used metric for graded relevance, namely the Discounted Cumulative Gain (DCG). A drawback of DCG is its additive nature and the underlying independence assumption: a document in a given position has always the same gain and discount independently of the documents shown above it. Inspired by the "cascade" user model, we present a new editorial metric for graded relevance which overcomes this difficulty and implicitly discounts documents which are shown below very relevant documents. More precisely, this new metric is defined as the expected reciprocal length of time that the user will take to find a relevant document. This can be seen as an extension of the classical reciprocal rank to the graded relevance case and we call this metric Expected Reciprocal Rank (ERR). We conduct an extensive evaluation on the query logs of a commercial search engine and show that ERR correlates better with clicks metrics than other editorial metrics.

## Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

## General Terms

Experimentation, Measurement

## Keywords

evaluation, non-binary relevance, web search, user model

## 1. INTRODUCTION

Evaluation has gained a great deal of attention in the field of information retrieval recently, primarily due to the rapidly changing landscape of information retrieval systems. The validity of the assumptions underlying Cranfield and TREC-style evaluation have been questioned, especially in light of growing test collections, new types of information needs, and the availability of new sources of relevance data, such as click logs and crowdsourcing.

How to properly evaluate web search engines continues to be a challenging, open research problem. Most web search evaluations in the information retrieval literature make use of cumulative gain-based metrics, such as Discounted Cumulative Gain (DCG) [15]. These metrics are popular because they support graded relevance, which is often used when judging the relevance of web documents. Although support for graded relevance is important, there are other important factors that should be considered when evaluating metrics.

One important factor that DCG does not account for is how the user actually interacts with the ranked list. The metric assumes that users will browse to some position in the ranked list according to some probability that only depends on the position. However, in reality, the probability that a user browses to some position in the ranked list depends on many other factors other than the position alone. One serious issue with DCG is the assumption that the usefulness of a document at rank $i$ is independent of the usefulness of the documents at rank less than $i$. Recent research on modeling user click behavior has demonstrated that the position-based browsing assumption that underlies DCG is invalid [12, 8]. Instead, these studies have shown that the likelihood a user examines the document at rank $i$ is dependent on how satisfied the user was with previously observed documents in the ranked list. This type of user model is known as the cascade model.

To make things more concrete, let us consider a simple example. Suppose that we are evaluating two ranked lists, where judgments are on a 5 point scale (*perfect*, *excellent*, *good*, *fair*, and *bad*). Suppose that the first ranked list consists of 20 *good* documents and the second list has 1 *perfect* document followed by 19 *bad* documents. Which ranked list is better? Under most settings, DCG would indicate that the first list is better. However, the single *perfect* document in the second ranked list completely satisfies the user's information need,[1] and thus the user would observe the first document, be satisfied, stop browsing, and never see any of the *bad* documents. On the other hand, with the first ranked

[1]In our guidelines a *perfect* grade is typically only given to the destination page of a navigational query.


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

list, the user would have to expend much more effort to satisfy his information need, because each *good* document only partially satisfies the need. Thus, users would actually prefer the second ranked list over the first, because it satisfies their information need with the least amount of effort.

The rank biased precision (RBP) metric recently proposed by Moffat and Zobel [17] incorporates a simple user model, but does not directly address the issue described here. The metric models the user's *persistence* in finding relevant documents. Less persistent users are only likely to look at a small number of results, while very persistent users will look deep in the ranked list. The primary issue with this simple user model is that real user browsing behavior is unlikely to be determined based entirely on user persistence, but also on the quality of the results in the ranked list, as illustrated in the example above. Indeed, Moffat and Zobel acknowledge this and explain that a more complex user model could be used instead, but leave this direction as future work.

The primary goal of this paper is to devise a metric that infuses RBP with a more accurate user model, thereby providing a better alternative to DCG and RBP. To achieve this, we propose a novel metric called expected reciprocal rank (ERR). The metric supports graded relevance judgments and assumes a cascade browsing model. In this way, the metric quantifies the usefulness of a document at rank $i$ conditioned on the degree of relevance of the items at ranks less than $i$. We argue, and show empirically, that our metric models user satisfaction better than DCG and is therefore a more appropriate metric for evaluating retrieval systems, especially those with graded judgments, such as web search engines.

There are two primary contributions of this work. First, we propose a novel metric based on a cascade-style user browsing model. We will demonstrate that most previous metrics have assumed position-based browsing, which has been shown empirically to be a poor assumption, and that the cascade model captures real user browsing behavior better. Second, our experimental results, which are carried out in the context of a very large test collection from a commercial search engine, show that ERR correlates with a range of click-based metrics much better than other editorial metrics such as DCG.

The remainder of this paper is laid out as follows. First, Section 2 provides some background on various aspects of information retrieval evaluation. Section 3 then describes recent research on user browsing models. Our proposed ERR metric is explained in Section 4. Section 5 explores how ERR is related to several existing retrieval metrics. Our rigorous, comprehensive experimental evaluation is described in Section 6. Finally, in Section 7 we discuss possible extensions of the model and conclude the paper in Section 8.

## 2. IR EVALUATION

Evaluation plays a critical role in the field of information retrieval. Retrieval systems are often evaluated in terms of their *effectiveness* and *efficiency*. Effectiveness evaluations quantify how good the search system is at satisfying users' search needs, while efficiency evaluations measure the speed of the system. Since our focus here is effectiveness, we will provide a brief overview of the various effectiveness measures that have been proposed for information retrieval.

Most, if not all, information retrieval effectiveness measures depend on the ill-defined notion of *relevance*. This is largely due to the prevalence and popularity of Cranfield and TREC-style evaluations [10, 25]. These evaluations are based on a fixed set of queries, a fixed set of documents, and a fixed set of *relevance judgments*. Relevance judgments are collected by asking human editors to assess the relevance of a document to a given query. In this way, relevance judgments capture the notion of *user relevance*. Retrieval metrics are then computed by comparing the output of the retrieval system, typically in the form of ranked lists over the queries in the evaluation set, and the relevance judgments.

Although relevance judgments and metrics can be decoupled, they are closely related. For this reason, whenever a new method for collecting relevance judgments is proposed, it typically comes along with a new retrieval metric. However, the other direction is not as common, as newly proposed metrics do not always require new relevance judgment criteria.

Relevance judgments come in many different flavors. The Cranfield experiments, and many subsequent evaluations, make a certain set of assumptions about the judgments. For example, it is assumed that relevance is *topical* (relevant documents are on the same topic as the query), that judgments are *binary* (relevant / not relevant), *independent* (relevance of document A does not depend on relevance of document B), *stable* (judgments do not change over time), *consistent* (judgments are consistent across editors), and *complete* (there are no missing judgments) [23]. These assumptions underlie most of the classical information retrieval metrics, such as precision, average precision, and recall.

While these assumptions simplify the editorial process to some extent, many are unrealistic and not well aligned with user relevance. For this reason, researchers have looked at various relaxations of the assumptions that have subsequently led to new retrieval metrics. We now highlight several examples. First, Järvelin and Kekäläinen [15] looked at graded relevance judgments and proposed the DCG metric that can exploit such judgments. Second, the TREC Novelty track investigated dependent relevance assessments [14]. The TREC Interactive Track evaluations had editors assign *subtopics* to each query. Editors were then asked to judge each document with respect to the subtopics [18]. This has led to various subtopic retrieval and diversity metrics [27, 9, 1]. Finally, various researchers have showed how the completeness assumption could be relaxed by inferring the relevance of missing judgments in various ways [3, 5, 7], ignoring unjudged documents [21], or intelligently choosing which unjudged documents to obtain judgments for [6].

Evaluation metrics themselves tend to be much simpler and have fewer issues compared to relevance judgments. Evaluation metrics are computed given the output of a retrieval system and the relevance judgments. Most measures make various assumptions about what makes for a "good" ranked list based on the existing editorial judgments and operationalizes that using some easy-to-compute mathematical formulation. For example, the DCG at rank $K$ for a given query is computed as [4]:

$$DCG@K := \sum_{i=1}^{K} \frac{2^{g_i} - 1}{\log(i+1)} \quad (1)$$

where $g_i$ is the relevance grade of the document at rank $i$. The numerator of the metric rewards documents with large relevance grades, while the denominator discounts the gains

at lower ranks. This simple metric operationalizes the notion that systems that rank highly relevant documents high in the ranked list are better than systems which rank highly relevant documents deep in the ranking. This general idea forms the basis for most precision-based metrics, including average precision, and the RBP metric described in the introduction, which is computed as follows:

$$RBP := (1-p) \cdot \sum_{i=1}^{n} g_i \cdot p^{i-1}$$

where $g_i$ indicates the degree of relevance of the document at rank $i$ and $p$ is a parameter that models how persistent a user is while looking through the ranked list. This measure makes similar assumptions to DCG, except the persistence parameter $p$ models some notion of user browsing behavior, which is absent in DCG. We will return to this important fact, and how it relates to our proposed metric, later in this paper.

We now explain where our proposed metric fits into the vast evaluation research landscape. Our metric is similar in spirit to DCG, in that we assume that relevance judgments are graded, independent, and complete. However, as with DCG, it is important to note that our metric can easily be extended to handle incompleteness [7], but for simplicity we assume completeness. Our metric is different from DCG, however, in that it incorporates a user model that acts as a proxy for dependent relevance judgments. DCG only discounts based on the rank of the document, but does not consider any of the documents previously seen by the user. Our metric, however, implicitly discounts documents based on the relevance of previously seen documents. The discounting that we use corresponds to a user browsing model.

## 3. USER MODELS

An accurate user model, which closely reflects users' interactions with the retrieval system, is essential for developing a good relevance metric. Before we go into the details of our proposed metric, it is worthwhile to first review existing user models for retrieval system. In general, there are two main types of user models: position models and cascade models. Both types of models attempt to capture the position bias of search result presentation. While the positional models assume independence among documents in different positions and model the examination probability as a function of the position, cascade models simultaneously model the relevance and examination probability of documents in the entire result set.

### *Position models.*

Position-based models [12, 20] are a popular class of methods for dealing with the presentation bias problem inherent in ranked retrieval systems. Among them, the *examination* model explicitly predicts the probability of examination at various positions. It relies on the assumption that the user clicks on a link if and only if the following two conditions are met: the user examined the URL *and* found it relevant; in addition, the probability of examination depends only on the position. The probability of click on the URL $u$ in position $p$ is thus modeled as [12, eq (3)]:

$$P(C=1|u,p) = a_u b_p,$$

where $a_u$ is the *attractiveness* of the URL $u$ – or the propensity of users to click on that URL, independent of the position – and $b_p$ is the probability of examination in position $p$, which depends only on that position.

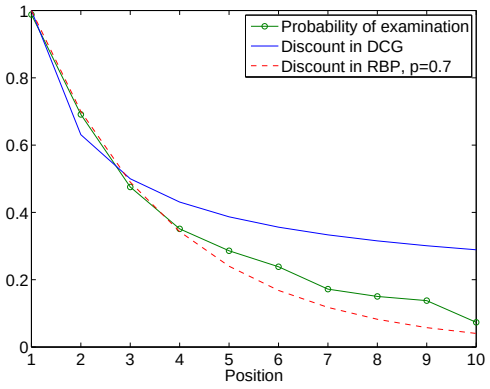


**Figure 1: Comparison of the discounting function in DCG and RBP with the probability of examination at a given position (estimated from click logs).**

Both the DCG metrics and RBP metrics adopt the position model as their underlying user browsing model and apply a position-based discounting function to progressively reduce the contribution of a document as its rank increases. Figure 1 compares the discounting functions of DCG and RBP with the probability of examination $b_p$, which was estimated from the click logs of a commercial search engine using the position model described above. As the graph shows, RBP with $p = 0.7$ closely approximates the estimated probability of examination. However, as we show next, assuming that the probability of examination only depends on position has some serious fundamental drawbacks.

Take for instance a relevant document in position 3; if both documents in position 1 and 2 are very relevant, it is likely that this document will be examined less and hence have very few clicks. On the other hand, if the two top documents are non-relevant, it is more likely to be examined and receive many clicks. A click model depending only on the position will not be able to model these two cases – the same document at the same position has a different click through rate (CTR) when the documents ranked above are different. A real example of this phenomenon, taken from the click logs of a commercial search engine, is shown in Figure 2. In this example, we observe that the CTR of the URL `www.myspace.com` in position 1 is 9 times larger than the CTR of the same URL when it is shown in position 2. On average, however, the CTR of position 1 is roughly twice the CTR of position 2. The difference is so large here because the URL shown in position 1 is an excellent match, and hence the user rarely even browses to position 2. Position models, assuming the independence among URLs, fail to explain such a drastic CTR difference.

### *Cascade Models.*

The examples above showed that in order to accurately model clicks and probabilities of examination, position is not sufficient and the relevance of documents above the document of interest has to be considered. Cascade models differ from the position models in that they consider the depen-

| Ranking 1 | | Ranking 2 | |
|---|---|---|---|
| URL | CTR | URL | CTR |
| uk.myspace.com | 0.97 | www.myspace.com | 0.97 |
| www.myspace.com | 0.11 | | |

**Figure 2: Illustration of the problem with position-based models. The query is myspace in the UK market. See text for discussion.**

dency among URLs on a search results page. In its generic form, the cascade model assumes that the user views search results from top to bottom and at each position, the user has a certain probability of being satisfied. Let $R_i$ be this probability at position $i$.[2] Once the user is satisfied with a document, he/she terminates the search and documents below this result are not examined regardless of their position. It is of course natural to expect $R_i$ to be an increasing function of the relevance grade, and indeed in what follows we will assimilate it to the often loosely-defined notion of "relevance". This generic version of the cascade model is summarized in Algorithm 1.

**Algorithm 1** The cascade user model

**Require:** $R_1, \ldots, R_{10}$ the *relevance* of the 10 URLs on the result page.

1: $i = 1$
2: User examines position $i$.
3: **if** random(0,1) $\leq R_i$ **then**
4: User is satisfied with the document in position $i$ and stops.
5: **else**
6: $i \leftarrow i + 1$; go to 2
7: **end if**

Two instantiations of this model have been presented in [12, 8]. In the former, $R_i$ is the same as the attractiveness defined above for position-based models: it measures a probability of click which can be interpreted as the relevance of the snippet. In that model, it is assumed that the user is always satisfied after clicking. It can however be the case that the snippet looks attractive, but that the user does not find any relevant information on the corresponding landing page. This is the reason why an extended cascade model has been proposed in [8, Section 5], in which the user might not be satisfied after clicking. More precisely, there is a probability, depending on the landing page, that the user will go back to the search result list after clicking. The $R_i$ in Algorithm 1 have now to be understood as the relevance probability of the landing page.

In both models a document satisfies the user with probability $R_i$. The values $R_i$ can be estimated by maximum likelihood on the click logs. Alternatively, as we will do in the next section, the $R_i$ values can be set as a function of the editorial grade of the URL. For a given set of $R_i$, the likelihood of a session for which the user is satisfied and stops at position $r$ is:

$$\prod_{i=1}^{r-1}(1 - R_i)R_r, \tag{2}$$

[2]The probability is in fact a function of the $i$-th document $d(i)$. However, for simplicity we shorten $R_{d(i)}$ to $R_i$.

which is simply the probability the the user is not satisfied with the first $r-1$ results and is satisfied with the $r$-th one.

# 4. PROPOSED METRIC

We now introduce our proposed metric based on the cascade model described in the previous section. A key step is the definition of the probability that a user finds a document relevant as a function of the editorial grade of that document. Let $g_i$ be the grade of the $i$-th document, then:

$$R_i := \mathcal{R}(g_i), \tag{3}$$

where $\mathcal{R}$ is a mapping from relevance grades to probability of relevance. $\mathcal{R}$ can be chosen in different ways; in accordance with the gain function for DCG used in [4], we might take it to be:

$$\mathcal{R}(g) := \frac{2^g - 1}{2^{g_{\max}}}, \quad g \in \{0, \ldots, g_{\max}\}. \tag{4}$$

When the document is non-relevant ($g = 0$), the probability that the user finds it relevant is 0, while when the document is extremely relevant ($g = 4$ if a 5 point scale is used), then the probability of relevance is near 1.

We first define the metric in a more general way by considering a *utility* function $\varphi$ of the position. This function typically satisfies $\varphi(1) = 1$ and $\varphi(r) \to 0$ as $r$ goes to $+\infty$.

Definition 1 (Cascade based metric). *Given a utility function $\varphi$, a cascade based metric is the expectation of $\varphi(r)$, where $r$ is the rank where the user finds the document he was looking for. The underlying user model is the cascade model* (2)*, where the $R_i$ are given by* (3)*.*

In the rest of this paper we will be considering the special case $\varphi(r) = 1/r$, but there is nothing particular about that choice and, for instance, we could have instead picked $\varphi(r) = \frac{1}{\log_2(r+1)}$ as in the discount function of DCG.

Definition 2 (Expected Reciprocal Rank). *The Expected Reciprocal Rank is a cascade based metric with $\varphi(r) = 1/r$.*

It may not seem straightforward to compute ERR from the previous definition because there is an expectation. However it can easily be computed as follows:

$$ERR := \sum_{r=1}^{n} \frac{1}{r} P(\text{user stops at position } r),$$

where $n$ is the number of documents in the ranking. The probability that the user stops at position $r$ is given by the definition of the cascade model (2). Plugging that value into the above equation, we finally obtain:

$$ERR := \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1}(1 - R_i)R_r. \tag{5}$$

A naïve computation using the above requires $O(n^2)$ operations. But as shown in Algorithm 2, ERR can easily be computed in $O(n)$ time.

Compared to position-based metrics such as DCG and RBP for which the discount depends only the position, the discount in ERR depends on the relevance of documents

**Algorithm 2** Algorithm to compute the ERR metric (5) in linear time.

**Require:** Relevance grades $g_i, 1 \leq i \leq n$, and mapping function $\mathcal{R}$ such as the one defined in (4).
  $p \leftarrow 1$, $ERR \leftarrow 0$.
  **for** $r = 1$ to $n$ **do**
    $R \leftarrow \mathcal{R}(g_r)$
    $ERR \leftarrow ERR + p \cdot R/r$
    $p \leftarrow p \cdot (1 - R)$
  **end for**
  **return** $ERR$

shown above it. The "effective" discount in ERR of document at position $r$ is indeed:

$$\frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i).$$

Thus the more relevant the previous documents are, the more discounted the other documents are. This diminishing return property is desirable because it reflects real user behavior.

Figure 3 summarizes our discussion up until this point. The figure shows the connection between user models and metrics. As the figure shows, most traditional measures, such as DCG and RBP assume a position-based user browsing model. As we have discussed, these models have been shown to be poor approximations of actual user behavior. The cascade-based user model, which is a more accurate user model, forms the basis for our proposed ERR metric.

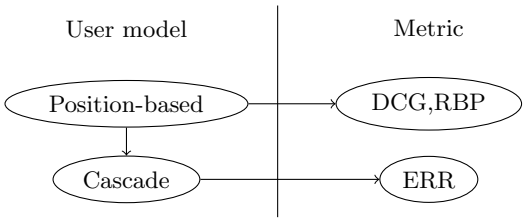


**Figure 3: Links between user models and metrics: DCG and RBP are instances of metrics that can be motivated using a position-based model. But the cascade user model is a more accurate model and the ERR metric derived from it correlates better with user satisfaction.**

## 5. RELATION TO OTHER METRICS

As we discussed in Section 2, our metric is similar to DCG to a certain extent. The metric also shares commonalities with several other metrics, which we will describe in this section.

First, our metric is related to the expected search length (ESL) metric, which was proposed by Cooper in 1968 [11]. The metric, which is defined over a *weak ordering* of documents, quantifies the amount of user effort necessary to find $K$ relevant documents. The measure computes the expected number of non-relevant documents that the user will see, by sequentially browsing the search result list, before finding the $K$th relevant document. In the simplest case, the weak ordering is taken as the ranked output of the system, in which the ESL can be computed by simply counting the number of non-relevant documents that occur before the $K$th relevant document. The measure has been shown to be useful for measuring the effectiveness of web search engines [24]. Our metric differs from ESL in that we explicitly support graded judgments and also take into account a user browsing model that is absent from the ESL model. One of the primary problems with ESL is that it requires knowing the appropriate value of $K$ for each query. Rather than assuming the user wants to find $K$ relevant documents, our metric measures the (inverse) expected user effort required to be satisfied.

Second, ERR is closely related to Moffat and Zobel's RBP metric [17]. Our metric can be thought of as an extension and generalization of RBP that makes use of the cascade model as a user browsing model. We note that Moffat and Zobel discuss the possibility of incorporating a user model into RBP by making $p$ dependent on the previously seen documents, the authors left this direction open as future work. The combination of the cascade model and RBP is natural and provides a number of benefits, including no need to set $p$ *a priori* and the possibility of seemlessly combining human judgments and clicks in a single unified framework, as will be discussed in Section 7.3.

Third, suppose that all of the $R_i$ values are either 0 or 1, which corresponds to the binary relevance setting. In this scenario it is easy to see that:

$$ERR := \frac{1}{\min\{r : R_r = 1\}}$$

which is exactly the reciprocal rank (RR) metric [26]. Thus, under binary relevance, ERR simplifies to RR.

Fourth, ERR can be seen as a special case of *Normalized Cumulative Utility* (NCU) [22], which is defined as

$$NCU := \sum_{r=1}^{n} p(r) NU(r),$$

where $p(r)$ is the probability that the user stops at position $r$ and $NU(r)$ is a *utility*, defined as a combination of benefit and effort for the user to have examined all the documents from position 1 to $r$. In the case of ERR, $p(r)$ is given by the cascade model (2), while $NU(r) = 1/r$. But we could have considered other choices for the utility function $NU$, such as precision or a normalized cumulative gain as discussed in [22]. The important common point between ERR and NCU is the separation of the stopping point probability from utility as discussed in [22, section 3.1].

Finally, one can recover the case of additive metrics such as DCG in the limit where the $R_i$ are infinitesimally small. In this case, $\prod_{i=1}^{r-1} R_i \approx 1$ and equation (5) is approximatively equal to:

$$\sum_{r=1}^{n} \frac{R_r}{r}.$$

It course does not make sense from a user model point of view to consider infinitesimally small $R_i$, but the point here is that when the $R_i$ are far away from 1, ERR turns out to be more similar to DCG. This happens in particular for difficult queries where there are only marginally relevant documents. This behavior has also been empirically observed as we shall see at the end of section 6.2.

# 6. EVALUATION

The evaluation of new metrics is challenging because there is no ground truth to compare with. Because of that, most papers that propose new metrics do not have direct evaluations. For instance in [16, 17] it is shown that the newly introduced metrics correlate well with other standard metrics. However, that does not imply that these metrics are "better" in the sense of user satisfaction.

We attempt to narrow this gap in this paper by considering click metrics. Even though clicks constitute indirect and noisy user feedback, they still contain some valuable information about user preferences. In fact, it has been shown that the quality of a retrieval system can be rather well estimated with clickthrough data [19]. In this evaluation, we will compute correlations between various click metrics and editorial metrics.

## 6.1 Data collection

We collected clickthrough data from a major commercial search engine in two different markets. The first step involves constructing *session* data. A session can be defined in various ways, but for this evaluation, it is defined as follows. A session always has a unique user and unique query. It starts when a user issues a query and ends with 60 minutes idle time on the user side. For each session, we get the query, list of URLs from the result sets, and a list of clicked URLs. Simple normalization is applied to queries and URLs. We restrict ourselves to either the top 5 or top 10 URLs of the first page along with the clicks on these URLs. In particular, for the top 5, the clicks on URLs after position 5 are ignored. We then intersect this data with editorial judgments and keep only the sessions for which we have editorial judgements for all the URLs. This is the reason why we consider sessions with depth 5: the intersection with editorial judgments is larger than with depth 10 and a larger number of sessions are retained. The editorial judgments are on a 5 grades scale from the set:

$$\{perfect, excellent, good, fair, bad\}.$$

Because of variations in the search engine, the result set for a given query may vary. We call a *configuration* a query with a given set of ordered results. The statistics of the data we collected in this way are summarized in Table 1.

## 6.2 Correlation with click metrics

Using the datasets presented in the previous section, we compute a weighted correlation over the configurations between an editorial and a click metric. More precisely, suppose that there are $N$ configurations (remember that a configuration is a query and an ordered set of results). For the $i$-th configuration, let $x_i$ be the value of some editorial metric, $y_i$ the value of the click metric, and $n_i$ the number of times this configuration is present in the dataset. Then, the weighted correlation is computed as follows:

$$C(x,y,n) = \frac{\sum_{i=1}^{N} n_i(x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=1}^{N} n_i(x_i - m_x)^2}\sqrt{\sum_{i=1}^{N} n_i(y_i - m_y)^2}},$$

where $m_x$ and $m_y$ are the weighted averages:

$$m_x = \frac{1}{\sum n_i}\sum_{i=1}^{N} n_i x_i, \qquad m_y = \frac{1}{\sum n_i}\sum_{i=1}^{N} n_i y_i.$$

The reason for introducing weights in the correlation computation is to reproduce the click logs distribution and give more weight to frequent queries.

The five editorial metrics that were compared in this evaluation are:

**DCG** Discounted Cumulative Gain (1), truncated at 5 or 10 depending on the depth of the dataset.

**NDCG** Normalized (with respect to the ideal DCG) DCG.

**AP** Average Precision, where the grades *perfect* and *excellent* are mapped to relevant and the grades *good*, *fair* and *bad* are mapped to non-relevant.

**RR** Reciprocal Rank. The grades are mapped to binary relevance as for AP.

**ERR** The metric proposed in this paper.

Regarding AP and RR, it may seem surprising that we mapped the *good* grade to non-relevant. However, as we will see later, this results in a higher correlation than mapping it to relevant.

Let us now discuss the click metrics. All of them are defined as averages over all the sessions belonging to a given configuration. Leveraging the conclusions from a previous study [19, table 2], we experimented with the most discriminative metrics according to the findings of that paper, namely:

**QCTR** Number of clicks in a session.

**UCTR** Binary variable indicating whether there was a click or not in the session. This is simply the opposite of abandonment.

**Max, Mean, Min RR** Respectively maximum, mean and minimum reciprocal ranks of the clicks. Zero if no clicks.

Of these 4 metrics, we found that QCTR was slightly negatively correlated with editorial metrics and we decided not to include it in the following tables. The problem with QCTR is that there are two conflicting interpretations: on the one hand, more relevant results should lead to higher number of clicks; but on the other hand, worse results can also lead to higher number of clicks because the user is struggling to find the information he is searching for. This is particularly true for navigational queries if the expected URL is not at the top of the ranking.

There are two other click metrics that we considered, which have not been published previously, but that we believe are good measurements of user satisfaction:

**SS** Search Success is similar to UCTR except that clicks on documents judged *fair* and *bad* are ignored. The idea is that the user might click on an attractive snippet, but the landing page is spam; we do not want to account for such a click.

**PLC** Precision at Lowest Click is defined as the number of clicks divided by the position of the lowest click. Assuming that the user scans the results from top to bottom, the position of the lowest click can be seen as a proxy for the number of results the user examined. If a click is interpreted as an indication of relevance, this ratio can be seen as the precision of the examined results.

**Table 1: Statistics of the datasets used in the evaluation after intersecting the query logs with editorial judgements. Coverage represents the number of sessions relative to the overall traffic.**

| | Market 1 | | Market 2 | |
|---|---|---|---|---|
| | Depth 5 | Depth 10 | Depth 5 | Depth 10 |
| Num queries | 16,356 | 5,291 | 3,476 | 543 |
| Num configurations | 69,846 | 59,025 | 33,572 | 26,136 |
| Num sessions | 72,752,548 | 42,442,782 | 186,668,363 | 35,516,572 |
| Coverage | 5.8% | 3.3% | 7.1% | 1.4% |

**Table 3: Correlations (market 1, depth 5) with AP and RR for two different mappings from graded to binary relevance: in the first one (G→NR), a *good* document is considered non-relevant, while in the second one (G→R), it is considered relevant.**

| | AP | | RR | |
|---|---|---|---|---|
| | G→NR | G→R | G→NR | G→R |
| UCTR | 0.225 | 0.004 | 0.238 | 0.111 |
| max RR | 0.348 | 0.033 | 0.366 | 0.172 |
| mean RR | 0.376 | 0.044 | 0.395 | 0.187 |
| PLC | 0.364 | 0.039 | 0.382 | 0.181 |
| SS | 0.402 | 0.335 | 0.415 | 0.454 |

The correlations between the editorial and click metrics described above are listed in Table 2. It is remarkable that for all click metrics and all datasets, ERR correlates better than any other editorial metric. This is an indication that ERR captures user satisfaction better than other metrics. It is worth noting that the correlation of NDCG is overall a bit inferior to DCG. A similar observation was made in [2]. As for the click metrics, the highest correlation seems to be with Min RR, which is the reciprocal rank of the lowest click. This can be explained by the fact that, in general, users are satisfied by the last clicked document, and if clicks are done from top to bottom, this also corresponds to the lowest clicked document.

For AP and RR as editorial metrics, the graded relevance judgements had to be converted to binary relevance. As indicated earlier, *perfect* and *excellent* were converted to relevant while *good*, *fair* and *bad* were converted to non-relevant. We have also tried to consider *good* documents as relevant, but as shown in Table 3, this resulted in much lower correlations.

### *Correlation for different classes of queries.*

In order to have a finer analysis of why ERR correlates better with click metrics, we now compute correlations conditioned on several properties of the query. Note that because of the *amalgamation paradox* [13] (similar to Simpson's paradox), these conditional correlations can be lower than the aggregated correlations reported in Table 2.

The queries have been split on 3 different axes: query length, query frequency, navigational vs non-navigational. For each split, we compute the correlation between the click and editorial metrics as was done in the previous section. For the sake of brevity, we restrict ourselves to market 1 with depth 5 and consider only mean reciprocal rank of the clicks as the click metric. This metric showed the largest overall correlation with editorial metrics in Table 2.

**Table 4: Correlation with mean RR as a function of the query length.**

| | 1 | 2 | 3 | 4+ |
|---|---|---|---|---|
| DCG | 0.218 | 0.180 | 0.261 | 0.265 |
| ERR | 0.525 | 0.372 | 0.415 | 0.361 |

**Table 5: Correlation with mean RR on navigational queries (35% of queries are in that category) and non-navigational queries.**

| | Navigational | Non-navigational |
|---|---|---|
| DCG | -0.005 | 0.104 |
| ERR | 0.222 | 0.226 |

The correlations as a function of the number of query terms and of the query type (navigational or not) are shown in tables 4 and 5 respectively. As for query frequency, we sorted queries by increasing frequency and put them in 10 different bins of equal size. The correlation for each decile is plotted in Figure 4. The same conclusion emerges from these three studies: the difference between DCG and ERR is not so large for difficult and tail queries, but it is more pronounced for easy and head queries. This behavior is not unexpected. As noted earlier, one drawback of DCG is that it does not enough discount documents following a *perfect* result (and to a lesser extent *excellent*. However, ERR does that implicitly by assuming that most users will not examine results after a *perfect*, and *perfect* results are more frequent for easy / head / navigational queries. We had also noted at the end of Section 5 that ERR tends to behave as an additive metric like DCG for difficult queries. The results presented here confirm that argument.

## 6.3 Comparison of two ranking functions

There is a potential pitfall in the evaluation from the previous section: the correlation coefficient implicitly compares metrics across configurations *and queries* which, as suggested for example by [25], has little significance; the only important aspect of these metrics' behavior is with respect to changes in the search results configuration for a given query.

We thus devise a mechanism to measure the correlations between *differences* of click and editorial metrics. From Table 1, it appears that there are on average 10 different configurations per query. We *emulate* two ranking systems by assigning, at random, two of these configurations to two virtual ranking systems for each query. Ideally, the sign of

**Table 2: Correlation between click and editorial metrics. Top row: market 1; bottom row: market 2. Left column: depth 10; right column: depth 5.**

| | DCG | NDCG | AP | RR | ERR |
|---|---|---|---|---|---|
| UCTR | 0.028 | -0.009 | 0.016 | 0.156 | **0.214** |
| max RR | 0.125 | 0.046 | 0.109 | 0.304 | **0.393** |
| mean RR | 0.147 | 0.063 | 0.129 | 0.333 | **0.428** |
| min RR | 0.162 | 0.076 | 0.142 | 0.352 | **0.451** |
| PLC | 0.138 | 0.057 | 0.121 | 0.321 | **0.413** |
| SS | 0.236 | 0.156 | 0.165 | 0.329 | **0.384** |

| | DCG | NDCG | AP | RR | ERR |
|---|---|---|---|---|---|
| UCTR | 0.374 | 0.268 | 0.435 | 0.592 | **0.664** |
| max RR | 0.389 | 0.348 | 0.473 | 0.626 | **0.700** |
| mean RR | 0.410 | 0.347 | 0.479 | 0.642 | **0.710** |
| min RR | 0.423 | 0.346 | 0.479 | 0.649 | **0.712** |
| PLC | 0.399 | 0.347 | 0.480 | 0.635 | **0.708** |
| SS | 0.461 | 0.309 | 0.249 | 0.467 | **0.501** |

| | DCG | NDCG | AP | RR | ERR |
|---|---|---|---|---|---|
| UCTR | 0.131 | 0.108 | 0.225 | 0.238 | **0.307** |
| max RR | 0.197 | 0.199 | 0.348 | 0.366 | **0.449** |
| mean RR | 0.220 | 0.217 | 0.376 | 0.395 | **0.480** |
| min RR | 0.235 | 0.230 | 0.394 | 0.414 | **0.500** |
| PLC | 0.210 | 0.209 | 0.364 | 0.382 | **0.466** |
| SS | 0.357 | 0.361 | 0.402 | 0.415 | **0.482** |

| | DCG | NDCG | AP | RR | ERR |
|---|---|---|---|---|---|
| UCTR | 0.298 | 0.215 | 0.335 | 0.356 | **0.435** |
| max RR | 0.319 | 0.253 | 0.369 | 0.407 | **0.493** |
| mean RR | 0.347 | 0.268 | 0.395 | 0.434 | **0.515** |
| min RR | 0.362 | 0.275 | 0.408 | 0.447 | **0.524** |
| PLC | 0.334 | 0.261 | 0.386 | 0.424 | **0.507** |
| SS | 0.427 | 0.378 | 0.373 | 0.402 | **0.465** |

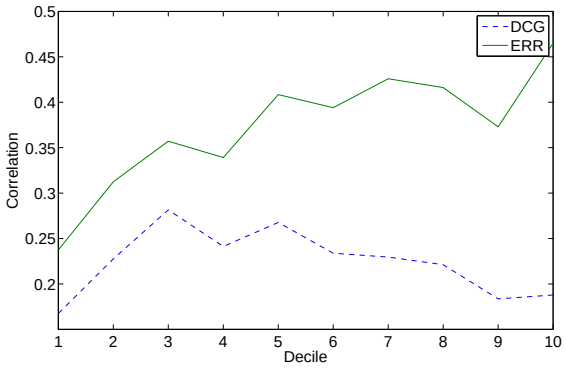


**Figure 4: Correlation with mean RR as a function of the query decile (1 = rarest; 10 = most frequent).**

difference between the editorial metric for these two systems should be the same as the sign of the difference of the respective click metric. To measure this, we repeat this procedure 1000 times and compute the correlation of the differences. The methodology is summarized in Algorithm 3 and the correlations are reported in Table 6.

In this evaluation ERR shows again the largest correlation with all click metrics reflecting that it is suitable to predict accurately, via the clicks, the difference in user satisfaction between two ranking functions. Therefore, ERR is not only effective for evaluating individual systems, but also quite effective at identifying differences between systems, as well.

## 7. EXTENSIONS

This section discusses a number of possible extensions of ERR that may be useful in a variety of evaluation scenarios.

### 7.1 Learning the parameters of $\mathcal{R}$

If $L$ is the number of relevance levels, then there are $L$ adjustable parameters in our metric corresponding to the values $\mathcal{R}(g)$ for each relevance level. Instead of fixing these values as in equation (4), we could optimize them by either maximizing the correlation with clicks or the agreement rate with side-by-side editorial tests. The latter was done for

**Algorithm 3** Methodology used to simulate two ranking functions and see whether the *difference* in the editorial metric correlates with the difference in click metric.

**for** $i = 1$ to 1000 **do**
  **for** $q = 1$ to $Q$ **do**
    $C_q :=$ {set of configurations for query $q$}
    **if** $|C_q| \geq 2$ **then**
      $c_1, c_2 \leftarrow$ two random elements of $C_q$.
      Assign $c_1$ to engine A and $c_2$ to engine B.
    **end if**
  **end for**
  Compute the average editorial metric (over the queries) for engines $A$ and $B$.
  Let $\Delta E_i$ be the difference between these two numbers.
  Similarly, let $\Delta C_i$ be the difference between the average click metric.
**end for**
Compute the correlation between $\Delta E$ and $\Delta C$.

**Table 6: Correlation between the differences of click and editorial metrics for two virtual ranking functions. See Algorithm 3 for details. The dataset is from market 1 with depth 10 (1st table) and 5 (2nd table).**

| | DCG | NDCG | AP | RR | ERR |
|---|---|---|---|---|---|
| UCTR | 0.069 | 0.061 | 0.141 | 0.163 | **0.195** |
| max RR | 0.129 | 0.130 | 0.243 | 0.278 | **0.340** |
| mean RR | 0.160 | 0.156 | 0.287 | 0.321 | **0.391** |
| PLC | 0.143 | 0.144 | 0.267 | 0.300 | **0.367** |
| SS | 0.349 | 0.319 | 0.331 | 0.361 | **0.408** |

| | DCG | NDCG | AP | RR | ERR |
|---|---|---|---|---|---|
| UCTR | 0.113 | 0.063 | 0.164 | 0.169 | **0.215** |
| max RR | 0.166 | 0.145 | 0.258 | 0.267 | **0.331** |
| mean RR | 0.187 | 0.154 | 0.284 | 0.295 | **0.361** |
| PLC | 0.180 | 0.155 | 0.273 | 0.284 | **0.350** |
| SS | 0.348 | 0.299 | 0.316 | 0.321 | **0.379** |

DCG in [28]. In addition to learning to the mapping function $\mathcal{R}$, it is also possible to learn the utility function $\varphi$ (see definition 1), that is, instead of using $\varphi(r) = 1/r$, learn the values $\varphi(1), \ldots, \varphi(K)$, where $K$ is the number of positions.

## 7.2 Extended cascade model

The original cascade model of [12] has later been extended in [8] to include an *abandonment* probability: if the user is not satisfied at a given position, he will examine the next url with probability $\gamma$, but has a probability $1 - \gamma$ of abandoning. In that model, the probability of the user stopping at position $r$ is:

$$\gamma^{r-1} \prod_{i=1}^{r-1} (1 - R_i) R_r,$$

which is the same as (2) but multiplied by $\gamma^{r-1}$. With the possibility of the user giving up, it can make sense to define a simpler utility function: 0 if the user abandoned, 1 otherwise; that is $\varphi(r) = 1$ in definition 1. The resulting metric is then defined as:

$$\sum_{r=1}^{n} \gamma^{r-1} \prod_{i=1}^{r-1} (1 - R_i) R_r,$$

which is very similar to the ERR formulation (5), the only difference being that the $1/r$ decay is replaced by a geometric decay $\gamma^{r-1}$.

## 7.3 Link with clicks

The mapping function from relevance grade to probability of relevance is currently chosen to match the gain function for DCG. An alternative way to define the mapping function is to learn it directly from click logs. For instance $\mathcal{R}(g)$ can be the average relevance estimated from click logs of all the URLs having grade $g$.

Our proposed metric can also be easily extended to accommodate relevance judgments in the form of combined editorial data and click data. So far most previous metrics have been based entirely on one type of relevance judgment and do not easily extend to use more than one. Since our metric is based on a click model, clicks can be combined with relevance judgments seemlessly within the metric. When there is a document with missing editorial judgement, we can use its predicted probability of relevance by fitting the cascade model with click logs. This could help the missing judgment issue. On the other hand, we can also largely rely on clicks for evaluation and only actively collect editorial judgment when we could not get a confident prediction for the probability of relevance from click logs. The latter one would be a more cost effective way of evaluating search engines.

## 7.4 Diversity

Research on metrics that incorporate the notion of *diversity* has recently gained interest [1, 9]. The measure presented in this paper and the underlying cascade model can easily be extended to handle this notion.

Let $P(t|q)$ be the distribution of topics[3] for a given query $q$. Each document is now judged with respect to the possible topics, where $g_i^t$ denotes the grade of the document in position $i$ for topic $t$. The associated probability of relevance is then $R_i^t := \mathcal{R}(g_i^t)$. As in the standard cascade model, a user interested by topic $t$ will stop at rank $r$ with probability:

$$\prod_{i=1}^{r-1} (1 - R_i^t) R_r^t.$$

Marginalizing over the topics, the probability that a user stops at rank $r$ is:

$$\sum_t P(t|q) \prod_{i=1}^{r-1} (1 - R_i^t) R_r^t.$$

And the diversity extension of ERR can be written as:

$$\sum_{r=1}^{n} \frac{1}{r} \sum_t P(t|q) \prod_{i=1}^{r-1} (1 - R_i^t) R_r^t. \quad (6)$$

Interestingly a similar equation to (6) has been derived in [1], but for the purpose of finding a diverse set of results, not for evaluation (their evaluation is based on an expected DCG which, in our opinion, does not give enough weight to minor intents). To be precise, the objective function of [1] is the probability that user finds a relevant result and not the expected reciprocal rank. But both are similar. Also [1] notes in conclusion that in future work, it would be better to optimize for "the expected rank at which the user will find useful information" instead of the probability that he will find something: that is what (6) achieves.

# 8. CONCLUSIONS

In this paper, we proposed a novel evaluation metric for information retrieval called expected reciprocal rank, or ERR. The metric, which was inspired by the cascade user browsing model, measures the (inverse) expected effort required for a user to satisfy their information need. The metric differs from average precision, rank-biased precision, and DCG in that it heavily discounts the contributions of documents that appear after highly relevant documents. This intuition, borrowed from the cascade model, assumes that a user is more likely to stop browsing if they have already seen one or more highly relevant documents. ERR supports graded relevance judgments and simplifies to reciprocal rank in the case of binary relevance judgments.

A rigorous set of empirical evaluations were carried out on a data set from a commercial search engine. The results showed that ERR consistently correlates better with a wide range of click-based metrics compared to DCG and other editorial metrics. The difference in correlation was particularly pronounced for navigational, short, and head queries, where ERR was much more highly correlated than DCG. Our experimental results suggest that ERR reflects real user browsing behavior better and quantifies user satisfaction more accurately than DCG.

Finally, we proposed several possible extensions to ERR that make the metric even more robust and attractive. These extensions include a method for automatically estimating the metric parameters, the ability to use both human editorial judgments and click data with the metric, and a simple way to incorporate the notion of diversity into the metric.

Thus, we argue that ERR should replace DCG as the *de facto* evaluation measure for web search engines, since it is inspired from an accurate user browsing model, correlates much better with a wide range of click-based metrics, and has a number of highly practical and useful extensions.

[3]Instead of topics, [1] refers to classes and [9] to nuggets.

## 9. REFERENCES


[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proc. 2nd Intl. Conf. on Web search and Web Ddata Mining*, pages 5–14, New York, NY, USA, 2009. ACM.

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

[3] J. A. Aslam and E. Yilmaz. Inferring document relevance from incomplete information. In *Proc. 16th Intl. Conf. on Information and Knowledge Management*, pages 633–642, New York, NY, USA, 2007. ACM.

[4] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the International Conference on Machine Learning*, 2005.

[5] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proc. 30th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 63–70, New York, NY, USA, 2007. ACM.

[6] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *Proc. 29th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 268–275, New York, NY, USA, 2006. ACM.

[7] B. Carterette and R. Jones. Evaluating search engines by modeling the relationship between relevance and clicks. In *Proc. 21st Proc. of Advances in Neural Information Processing Systems*, 2007.

[8] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proc. 18th Intl. Conf. on World Wide Web*, pages 1–10, New York, NY, USA, 2009. ACM.

[9] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proc. 31st Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 659–666, New York, NY, USA, 2008. ACM.

[10] C. W. Cleverdon. *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Cranfield College of Aeronautics, 1962.

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

[12] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proc. 1st Intl. Conf. on Web search and Web Ddata Mining*, pages 87–94, New York, NY, USA, 2008. ACM.

[13] I. Good and Y. Mittal. The amalgamation and geometry of two-by-two contingency table. *The Annals of Statistics*, pages 694–711, 1987.

[14] D. Harman. Overview of the TREC 2002 novelty track. In *Proc. 11th Text REtrieval Conference*, 2002.

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

[16] J. Kekäläinen. Binary and graded relevance in IR evaluations – comparison of the effects on ranking of IR systems. *Information Processing and Management*, 41(5):1019–1033, 2005.

[17] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):1–27, 2008.

[18] P. Over. TREC-7 interactive track report. In *Proc. 7th Text REtrieval Conference*, 1998.

[19] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Proc 17th Intl. Conf. on Information and Knowledge Management*. ACM New York, NY, USA, 2008.

[20] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 521–530. ACM, 2007.

[21] T. Sakai. Alternatives to bpref. In *Proc. 30th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 71–78, New York, NY, USA, 2007. ACM.

[22] T. Sakai and S. Robertson. Modelling a user population for designing information retrieval metrics. In *Proceedings of the Second Workshop on Evaluating Information Access (EVIA)*, 2008.

[23] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. *J. Am. Soc. Inf. Sci. Technol.*, 58(13):2126–2144, 2007.

[24] M.-C. Tang and Y. Sun. Evaluation of web-based search engines using user-effort measures. *Library and Information Science Research Electronic Journal*, 13(2), 2003.

[25] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Proc. 2nd CLEF Workshop on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, London, UK, 2002. Springer-Verlag.

[26] E. M. Voorhees and D. M. Tice. The TREC-8 question answering track evaluation. In *In Text Retrieval Conference TREC-8*, pages 83–105, 1999.

[27] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proc. 26th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 10–17, New York, NY, USA, 2003. ACM.

[28] K. Zhou, H. Zha, G. Xue, and Y. Yu. Learning the gain values and discount factors of DCG. In P. N. Bennett, B. Carterette, O. Chapelle, and T. Joachims, editors, *SIGIR Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments*, pages 7–14, 2008.