Extending Average Precision to Graded Relevance Judgments

Stephen E. Robertson* ser@microsoft.com
Evangelos Kanoulas†* e.kanoulas@sheff.ac.uk
Emine Yilmaz* eminey@microsoft.com

*Microsoft Research
7 JJ Thomson Avenue
Cambridge CB3 0FB, UK

†Department of Information Studies
University of Sheffield
Sheffield S1 4DP, UK

ABSTRACT

Evaluation metrics play a critical role both in the context of comparative evaluation of the performance of retrieval systems and in the context of learning-to-rank (LTR) as objective functions to be optimized. Many different evaluation metrics have been proposed in the IR literature, with average precision (AP) being the dominant one due a number of desirable properties it possesses. However, most of these measures, including average precision, do not incorporate graded relevance.

In this work, we propose a new measure of retrieval effectiveness, the Graded Average Precision (GAP). GAP generalizes average precision to the case of multi-graded relevance and inherits all the desirable characteristics of AP: it has a nice probabilistic interpretation, it approximates the area under a graded precision-recall curve and it can be justified in terms of a simple but moderately plausible user model. We then evaluate GAP in terms of its informativeness and discriminative power. Finally, we show that GAP can reliably be used as an objective metric in learning to rank by illustrating that optimizing for GAP using SoftRank and LambdaRank leads to better performing ranking functions than the ones constructed by algorithms tuned to optimize for AP or NDCG even when using AP or NDCG as the test metrics.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]

General Terms: Experimentation, Measurement, Performance

Keywords: information retrieval, effectiveness metrics, average precision, graded relevance, learning to rank

1. INTRODUCTION

Evaluation metrics play a critical role both in the context of comparative evaluation of the performance of retrieval systems and in the context of learning-to-rank (LTR) as objective functions to be optimized. Many different evaluation metrics have been proposed and studied in the literature. Even though different metrics evaluate different aspects of retrieval effectiveness, only a few of them are widely used, with average precision (AP) being perhaps the most commonly used such metric. AP has been the dominant system-oriented evaluation metric in IR for a number of reasons:

- It has a natural top-heavy bias.
- It has a nice probabilistic interpretation [25].
- It has an underlying theoretical basis as it corresponds to the area under the precision recall curve.
- It can be justified in terms of a simple but moderately plausible user model [16].
- It appears to be highly informative; it predicts other metrics well [2].
- It results in good performance ranking functions when used as objective in learning-to-rank [27, 24].

The main criticism to average precision is that it is based on the assumption that retrieved documents can be considered as either relevant or non-relevant to a user's information need. Thus, documents of different relevance grades are treated as equally important with relevance conflated into two categories. This assumption is clearly not true: by nature, some documents tend to be more relevant than others and intuitively the more relevant a document is the more important it is for a user. Further, when AP is used as an objective metric to be optimized in learning to rank, the training algorithm is also missing this valuable information.

For these reasons, a number of evaluation metrics that utilize multi-graded relevance judgments has appeared in the literature (e.g. [15, 8, 9, 19, 17]), with nDCG [8, 9] being the most popular among them, especially in the context of learning-to-rank as most learning to rank algorithms are designed to optimize for nDCG [6, 5, 22, 24].

In the framework used to define nDCG, a relevance score is mapped to each relevance grade, e.g. 3 for highly relevant documents, 2 for fairly relevant documents and so on. The relevance score of each document is viewed as the *gain* returned to a user when examining the document (utility of the document). To account for the 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

*We gratefully acknowledge the support provided by the European Commission grants FP7-ICT-248347 and FP7-PEOPLE-2009-IIF-254562.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

patience of a user to step down the ranked list of documents. The discounted gain values are then summed progressively from rank 1 to k. This discounted cumulative gain at rank k is finally normalized in a 0 to 1 range to enable averaging the values of the metric over a number of queries, resulting in the *normalized Discounted Cumulative Gain*, nDCG.

The nDCG metric is thus a functional of a gain and a discount function and thus it can accommodate different user search behavior patterns on different retrieval task scenarios. As it has been illustrated by a number of correlation studies different gain and discount functions lead to radically different rankings of retrieval systems [23, 12, 11].

Despite the great flexibility nDCG offers, defining gain and discount functions in a meaningful way is a difficult task. Given the infinite number of possible discount and gain functions, the vast differences in users search behavior, the many different possible retrieval tasks and the difficulty in measuring user satisfaction, a complete and rigorous analysis of the relationship between different gain and discount functions and user satisfaction under different retrieval scenarios is prohibitively expensive, if at all possible.

For this reason, in the past, the selection of the gain and discount functions has been done rather arbitrarily, based on speculations of the search behavior of an average user and speculations of the correlation of the metric to user satisfaction. For instance, Burges et al. [5], 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 a highly relevant document is very much more valuable than one of a slightly lower grade. Further, the logarithmic discount function ($1/log(r+1)$) dominated the literature compared to the linear 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 linear discount.

Despite the reasonable assumptions behind the choice of the gain and discount function that dominates nowadays the literature, recent work [1] demonstrated 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 strongly questions the validity of the aforementioned assumptions but mostly underlines the difficulty in specifying gain and discount functions in a meaningful manner.

Due to the above difficulties associated with the current multigraded evaluation metrics, even when multigraded relevance judgments are available, average precision is still reported (together with the multigraded metrics) by converting the relevance judgments to binary [4, 3]. Thus, despite the invalid assumption of binary relevance, average precision remains one of the most popular metrics used by IR researchers (e.g. in TREC [3]).Furthermore, even though AP is wasting valuable information in the context of learning-to-rank, since it ignores the swaps between documents of different positive relevance grades, it has been successfully used as an objective metric [27]. Therefore, we believe that a direct extension of the metric to the multigraded case in a systematic manner is needed and it will become a valuable tool for the community both in the context of evaluation and in the context of LTR.

In this paper, we generalize average precision to the multigraded relevance case in a systematic manner, proposing a new metric, the *graded average precision* (GAP). The GAP metric is a direct extension of AP and thus it inherits all the desirable properties that average precision has:

- It has the same natural top-heavy bias average precision has.
- It has a nice probabilistic interpretation.
- It has an underlying theoretical basis as it corresponds to the area under the "graded" precision-recall curve.
- It can be justified in terms of a simple but moderately plausible user model similarly to AP
- It appears to be highly informative.
- When used as an objective function in learning-to-rank it results in good performance retrieval systems (it outperforms both AP and nDCG).

The incorporation of multi-graded relevance in average precision becomes possible via a simple probabilistic user model which naturally dictates to what extend documents of different relevance grades account for the effectiveness score. This user model corresponds to one of the approaches briefly discussed in Sakai and Robertson [20]. This model offers an alternative way of thinking about graded relevance compared to the notion of utility employed by nDCG and other multi-graded metrics.

Sakai [19] for instance has previously introduced a multigraded measure (the Q-measure) which has been shown to behave similarly to AP for ranks above R (where R is the number of relevant documents in the collection). Nevertheless, the incorporation of graded relevance by the Q-measure follows the same model with nDCG. GAP on the other hand is based on the well-trusted notions of precision and recall as is AP.

In what follows, we first describe the user model on which GAP is based and define the new metric. We then describe some desirable properties GAP possesses. In particular, we describe a probabilistic interpretation of GAP, generalize precision-recall curves for the multigraded relevance case and show that GAP is an approximation to the area under the graded precision-recall curves. Further, we evaluate GAP in terms of informativeness [2] and discriminative power [18]. Finally, we extend two popular LTR algorithms, SoftRank [22] and LambdaRank [6], to optimize for GAP and test the performance of the resulting ranking functions over different collections.

2. GRADED AVERAGE PRECISION (GAP)

2.1 User Model

We start from a rudimentary user model, as follows: assume that the user actually has a binary view of relevance, determined by thresholding the relevance scale $\{0..c\}$. We describe this model probabilistically – we have a probability g_i that the user sets the threshold at grade i, in other words regards grades $i, ..., c$ as relevant and the others as non-relevant. We consider this probability to be defined over the space of users. These should be exclusive and exhaustive probabilities: $\sum_{j=1}^{c} g_j = 1$.

2.2 Definition of GAP

Now, we want some form of expected average precision, the expectation being over this afore-defined probabilistic event space. Simple interpretation of this (just calculate

average precision separately for each grade and take a probabilistically weighted combination) has problems; for instance, in the case of an ideal ranked list, when there are no documents in some grades, the effectiveness score returned is less than the optimal value of 1. So, instead, we extend the non-interpolated form of AP; that is, we step down the ranked list, looking at each relevant document in turn (the "pivot" document) and compute the expected precision at this rank. With an appropriate normalization at the end, this defines the graded average precision (GAP).

In particular, suppose we have a ranked list of documents, and document d_n at rank n has relevance $i_n \in \{0..c\}$. If $i_n > 0$, d_n, as pivot document, will contribute a precision value to the average precision calculations for each grade j, $0 < j \leq i_n$, since for any threshold set at grades less than or equal to i_n, d_n is considered relevant. The binary precision value for each grade j is, $\frac{1}{n}(|d_m : m \leq n, i_m \geq j|)$, while the expected precision at rank n over the aforementioned probabilistic user space can be computed as,

$$E[PC_n] = \sum_{j=1}^{i_n} \left(\frac{1}{n} |d_m : m \leq n, i_m \geq j| \right) \cdot g_j$$

Let $I(i, j)$ be an indicator variable equal to 1 if grade i is larger than or equal to grade j and 0 otherwise. Then, the expected precision at rank n can also be written as,

$$\begin{aligned} E[PC_n] &= \sum_{j=1}^{i_n} \left(\frac{1}{n} |d_m : m \leq n, i_m \geq j| \right) \cdot g_j \\ &= \frac{1}{n} \sum_{j=1}^{i_n} g_j \sum_{m=1}^{n} I(i_m, j) \\ &= \frac{1}{n} \sum_{m=1}^{n} \sum_{j=1}^{min(i_n, i_m)} g_j \quad \text{if } i_m > 0 \end{aligned}$$

By observing the new form of calculation of $E[PC_n]$, we can compute the contribution of each document ranked at $m \leq n$ to this weighted sum for those grades $j \leq i_m$. Thus we define a contribution function:

$$\delta_{m,n} = \begin{cases} \sum_{j=1}^{min(i_m, i_n)} g_j & \text{if } i_m > 0 \\ 0 & \text{otherwise} \end{cases}$$

Now the contribution from the pivot document can be defined as, $E[PC_n] = \frac{1}{n} \sum_{m=1}^{n} \delta_{m,n}$.

The maximum possible $E[PC_n]$ depends on the relevance grade i_n, it is the probability that this document is regarded as relevant by the user, $\sum_{j=1}^{i_n} g_j$. We must take account of this when normalizing the sum of $E[PC_n]$'s. Suppose we have R_i total documents in grade i (for this query); then the maximum possible value of cumulated $E[PC_n]$'s is, $\sum_{i=1}^{c} R_i \sum_{j=1}^{i} g_j$, which corresponds to the expected number of documents considered relevant in the collection, with the expectation taken over the space of users, as above.

The graded average precision (GAP) is then defined as:

$$GAP = \frac{\sum_{n=1}^{\infty} \frac{1}{n} \sum_{m=1}^{n} \delta_{m,n}}{\sum_{i=1}^{c} R_i \sum_{j=1}^{i} g_j}$$

Remark on thresholding probabilities: The user model that GAP is based on dictates the contribution of different relevance grades to the GAP calculation by considering the probability of a user thresholding the relevance scale at a certain relevance grade (the g values). This allows a better understanding and an easier mechanism to determine the relative value of different relevance grades to an average user than the underlying model for the current multi-graded evaluation metrics. For instance, given the relevance grades of documents, click through data can be utilized to conclude relative preferences of users among documents of different relevance grades [10, 14]. Assuming that the user only clicks on the documents he finds relevant, the g values correspond to the probability that a user clicks on a document of a particular relevance grade, given all the documents clicked by the user. In this paper, given that our goal is to develop a good system-oriented metric, we propose an alternative way of setting the g values by considering which $g = \{g_i\}$ makes the metric most informative (see Section 4.1).

3. PROPERTIES OF GAP

In this section we describe some of the properties of GAP that make the metric understandable and desirable to use.

First, it is easy to see that GAP generalizes average precision – it reverts to average precision in the case of binary relevance. With respect to the model described in Section 2.1, binary relevance means that all users find documents with some relevance grade $t > 0$ relevant and the rest non-relevant (i.e., $g_j = 1$ if $j = t$, for some relevance grade $t > 0$ and 0 otherwise).

Furthermore, GAP behaves in the expected way under document swaps. That is, if a document is swapped with another document of smaller relevance grade that appears lower in the list, the value of GAP decreases and vice-versa. As a corollary to this property, GAP acquires its maximum value when documents are returned in non-increasing relevance grade order.

In the following sections, we describe a probabilistic interpretation of GAP and show that GAP is an approximation to the area under a graded precision-recall curve.

3.1 Probabilistic interpretation

In this section we define GAP as the expected outcome of a random experiment, which is a generalization of the random experiment whose expected outcome is average precision [25], for the case of graded relevance. This offers an intuition behind the new measure.

3.1.1 Probabilistic interpretation of AP

Yilmaz and Aslam [25] have shown that AP corresponds to the expected outcome of the following random experiment:

1. Select a relevant document at random. Let the rank of this document be n.
2. Select a document at or above rank n, at random. Let the rank of that document be m.
3. Output 1 if the document at rank m, d_m, is relevant.

In expectation, steps (2) and (3) effectively compute the precision at a relevant document. Then step (1), in combination with steps (2) and (3), effectively computes the average of these precisions. Hence, average precision corresponds to the probability that a document retrieved above a randomly picked relevant document is also relevant.

3.1.2 Probabilistic interpretation of GAP

Consider the case where graded relevance judgments are available. We claim that GAP corresponds to the expected outcome of the following random experiment:

1. Select a document that is considered relevant by a user (according to the afore-defined user model), at random. Let the rank of this document be n.
2. Select a document at or above rank n, at random. Let the rank of that document be m.
3. Output 1 if the document at rank m, d_m, is also considered relevant by the user.

Hence, GAP can be seen as the probability that a document retrieved above a randomly picked "relevant" document is also "relevant", where relevance is defined according to the user model previously described.

We compute the expectation of the above random experiment to show that it corresponds to GAP. In expectation, step (3) corresponds to the conditional probability of document d_m being considered as relevant given that document d_n is also considered as relevant. To calculate this probability, let's consider all possible cases of the relative ordering of the relevant grades for documents d_n and d_m.

- $(i_n \leq i_m)$: Since the relevance grade of d_n is smaller than or equal to the one for d_m, if d_n is considered relevant then d_m will also be considered as relevant.

$$Pr(d_m = \text{rel}|d_n = \text{rel}) = \\ = 1 = \frac{\sum_{j=1}^{i_n} g_j}{\sum_{j=1}^{i_n} g_j} = \frac{\sum_{j=1}^{min(i_n,i_m)} g_j}{\sum_{j=1}^{i_n} g_j}$$

since $min(i_n, i_m) = i_n$.

- $(i_n > i_m)$: By applying the Bayes' Theorem,

$$Pr(d_m = \text{rel}|d_n = \text{rel}) = \\ = \frac{Pr(d_n = \text{rel}|d_m = \text{rel}) \cdot Pr(d_m = \text{rel})}{Pr(d_n = \text{rel})} \\ = \frac{1 \cdot \sum_{j=1}^{i_m} g_j}{\sum_{j=1}^{i_n} g_j} = \frac{\sum_{j=1}^{min(i_n,i_m)} g_j}{\sum_{j=1}^{i_n} g_j}$$

since $min(i_n, i_m) = i_m$

In expectation, steps (2) and (3) together, correspond to the value the "pivot" document d_n will contribute to GAP,

$$\frac{1}{n} \cdot \sum_{m=1}^{n} \frac{\sum_{j=1}^{min(i_n,i_m)} g_j}{\sum_{j=1}^{i_n} g_j}$$

In step (1), the probability that a document d_n is considered relevant is $\sum_{j=1}^{i_n} g_j$. Thus, the probability of selecting this document out of all documents that are considered relevant is,

$$p_{d_n} = \frac{\sum_{j=1}^{i_n} g_j}{\sum_{i=1}^{c} R_i \sum_{j=1}^{i_n} g_j}$$

Therefore, step (1) in combination with steps (2) and (3) effectively computes the average of the contributed values, which corresponds to GAP,

$$GAP = \sum_{n=1}^{\infty} \frac{1}{n} \sum_{m=1}^{n} \frac{\sum_{j=1}^{min(i_n,i_m)} g_j}{\sum_{j=1}^{i_n} g_j} \cdot \frac{\sum_{j=1}^{i_n} g_j}{\sum_{i=1}^{c} R_i \cdot \sum_{j=1}^{i_n} g_j} \\ = \frac{\sum_{n=1}^{\infty} \frac{1}{n} \sum_{m=1}^{n} \sum_{j=1}^{min(i_n,i_m)} g_j}{\sum_{i=1}^{c} R_i \sum_{j=1}^{i} g_j}$$

3.2 GAP as the area under the graded precision-recall curves

In this section we first intuitively extend recall and precision to the case of multi-graded relevance, based on the probabilistic model defined in Section 2.1. Then we define the graded precision-recall curve, and finally show that GAP approximates the area under the graded precision-recall curve, as AP approximates the area under the binary precision-recall curve.

Precision-recall curves are constructed by plotting precision against recall each time a relevant document is retrieved. In the binary relevance case, recall is defined as the ratio of relevant documents up to rank n to the total number of relevant documents in the query. In the graded relevance case, a document is considered relevant only with some probability. Therefore, recall at a relevant document at rank n can be defined as the ratio of the expected number of relevant documents up to rank n to the expected total number of relevant documents in the query (under the independence assumption between numerator and denominator).

In particular, according to the user model defined in Section 2.1, documents of relevance grade i_m are considered relevant with probability $\sum_{j=1}^{i_m} g_j$, and thus, the expected number of relevant documents up to rank n is, $\sum_{m=1}^{n} \sum_{j=1}^{i_m} g_j$, while the expected total number of relevant document is, $\sum_{i=1}^{c} R_i \sum_{j=1}^{i} g_j$.

Hence, the graded recall at rank n can be computed as,

$$\text{graded Recall@n} = \frac{\sum_{m=1}^{n} \sum_{j=1}^{i_m} g_j}{\sum_{i=1}^{c} R_i \sum_{j=1}^{i} g_j}$$

The recall step, i.e. the proportion of relevance information acquired when encountering a "relevant" document at rank n to the total amount of relevance, is, $\sum_{j=1}^{i_n} g_j / \sum_{i=1}^{c} R_i \sum_{j=1}^{i} g_j$. This corresponds to the expected outcome of step (1) of the random experiment described in Section 3.1 and expresses the probability of selecting a "relevant" document at rank n out of all possible "relevant" documents.

In the binary case, precision at a relevant document at rank n is defined as the fraction of relevant documents up to that rank. In the multi-graded case, precision at a "relevant" document at rank n can be defined as the expected number of documents at or above that rank that are also considered as "relevant" This quantity corresponds to the expected outcome of steps (2) and (3) of the random experiment in Section 3.1,

$$\text{graded Precision@n} = \frac{1}{n} \cdot \sum_{m=1}^{n} \frac{\sum_{j=1}^{min(i_n,i_m)} g_j}{\sum_{j=1}^{i_n} g_j}$$

Therefore, graded average precision can be alternatively defined as the cumulated product of graded precision values and graded recall step values at documents of positive relevance grade, as average precision can be defined as the cumulated product of precision values and recall step values at relevant documents.

Given the definitions of graded precision and graded recall, one can construct precision-recall curves. Now it is easy to see that GAP is an approximation to the area under the non-interpolated graded precision-recall curve as AP is an approximation to the area under the non-interpolated binary precision-recall curve.

Note that Kekäläinen and Järvelin [13] have also proposed a generalization of precision and recall. The way they generalized the two statistics is radically different than the one we

propose; in their work precision and recall follow the nDCG framework where gain values are assigned to each document.

4. EVALUATION OF GAP

There are two important properties that a system-oriented evaluation metric should have: (1) it should be highly informative [2] – that is it should summarize the quality of a search engine well, and (2) it should be highly discriminative – that is it should identify the significant differences in the performance of the systems. We evaluated GAP in terms of both of these properties. We used nDCG as a baseline for comparison purposes. Given that our goal is to propose a good system-oriented metric that can be used as an objective function to optimize for in LTR, in what follows we mostly focus on the informativeness of the metric since it has been shown to correlate well with the effectiveness of the trained ranking function [26].

In particular, when a ranking function is optimized for an objective evaluation metric, the evaluation metric used during training acts as a bottleneck that summarizes the available training data. At each training epoch, given the relevance of the documents in the training set and the ranked list of documents retrieved by the ranking function for that epoch, the only information the learning algorithm has access to is the value of the evaluation metric. Thus, the ranking function will change on the basis of the change in the value of the metric. Since more informative metrics better summarize the relevance of the documents in the ranked list and thus better capture any change in the ranking of documents, the informativeness of a metric is intuitively correlated with the ability of the LTR algorithm to "learn" well.

4.1 Informativeness

To assess the informativeness of the evaluation metrics we use the *Maximum Entropy Method* (MEM) as proposed in Aslam et al. [2].

Similar to Aslam et al. we make the assumption that the quality of a list of documents retrieved in response to a given query is strictly a function of the relevance of the documents within that list (as well as the total number of relevant documents for the given query). Then, the question that naturally arises is how well does a metric capture the relevance of the output list and consequently the effectiveness of a retrieval system? In other words, given the value of a metric, for a given system on a given query, how accurately can one predict the relevance of documents retrieved?

Suppose that you were given a list of length N corresponding to output of a retrieval system for a given query, and suppose that you were asked to predict the probability of seeing a relevant document at some rank. Since there are no constraints, all possible lists of length N are equally likely, and hence the probability of seeing a relevant document at any rank is $1/2$. Suppose now that you are also given the information that the expected number of relevant documents over all lists of length N is R. The most natural answer would be a R/N uniform probability for each rank. Finally, suppose that you are given the additional constraint that the expected value of a metric is v. Under the assumption that our distribution over lists is a product distribution, i.e. $p(r_1, r_2, ..., r_N) = p(r_1) \cdot p(r_2) \cdot ... \cdot p(r_N)$ (Aslam et al. call this *probability-at-rank distribution*), we can solve the problem by using MEM. That is, we find the most random probability-at-rank distribution (by maximizing the entropy of p) that satisfies the following constraints: (a) the expected value of the metric over the probability-at-rank distribution is v, and (b) the expected number of relevant documents in each grade ξ is R_ξ).

To apply the maximum entropy method we derive the expected GAP and nDCG over the probability-at-rank distribution. The derivations are omitted due to space limitations. The maximum entropy formulations are shown in Figure 1. Both of them are constraint optimization problems and numerical methods were used to determine their solutions.

The result of the above optimization is a maximum entropy probability-at-rank distribution (over all relevance grades). Using this probability-at-rank distribution, we can infer the maximum entropy precision-recall curve. If a metric is very informative then the maximum entropy precision-recall curve should approximate well the actual precision-recall curve.

We then test the performance of GAP and nDCG using data from TRECs 9 and 10 Web Tracks (ad-hoc task) and TREC 12 Robust Track (only the topics 601-650 that have multi-graded judgments). Using the setup described above, we first infer the probability-at-rank distributions given the value of each metric and then calculate the maximum entropy precision-recall curves when only highly relevant documents are considered as relevant and when both relevant and highly relevant documents are considered as relevant (the graded PR-curves described in Section 3.2 are not used due to their bias towards GAP). As in Aslam et al. [2], for any query, we choose those systems that retrieved at least 5 relevant and 5 highly relevant documents to have a sufficient number of points on the precision-recall curves. We use different values for g_1 and g_2 to investigate their effect on the informativeness of GAP.

The mean RMS error between the inferred and the actual precision-recall curves, calculated at the points where recall changes, is illustrated in Figure 2. The x-axis corresponds to different pairs of threshold probabilities, g_1 and g_2. The blue solid line corresponds to the RMS error between the actual and the inferred precision-recall curves subject to GAP, while the red dashed line indicates the RMS error of the inferred precision-recall curves subject to nDCG.

As it can be observed (1) the choice of g_1 and g_2 appears to affect the informativeness of GAP; when g_1 is high GAP appears to summarize well the sequence of all relevant documents independently of their grade, while when g_2 is high GAP appears to summarize well the sequence of all highly relevant documents, (2) choosing g_1 and g_2 to be relatively balanced (around 0.5) seems to be the best compromise between summarizing well the sequence of all relevant documents independent of their grade and highly relevant documents only, and (3) with g_1 and g_2 to relatively balanced GAP appears to be more informative than nDCG in most of the cases[1]. Finally, note that when the thresholding probability $g_1 = 1$ (the right-most point for GAP curve in all plots), GAP reduces to average precision since relevant and highly relevant documents are conflated in a sin-

[1]Different gain (linear vs. exponential) and discount (linear vs. log) functions used in the definition of nDCG were tested. The ones that utilized the log discount function appeared to be the most informative, while the effect of the gain function on informativeness was limited. The nDCG metric used here utilizes an exponential gain and a log discount function.

Maximize: $H(\vec{p}) = \sum_{n=1}^{N} H(p_n)$

Subject to:

1. $\sum_{n=1}^{N}\sum_{\xi=0}^{c} \frac{Pr(i_n=\xi)}{n} \cdot \left(\sum_{j=1}^{\xi} g_j + \sum_{m=1}^{n-1}\sum_{\zeta=0}^{c}\left(\left(\sum_{j=1}^{min(\zeta,\xi)} g_j\right) Pr(i_m=\zeta)\right)\right) / \left(\sum_{i=1}^{c} R_i \sum_{j=1}^{i} g_i\right) = gap$
2. $\sum_{n=1}^{N} Pr(i_n=\xi) = R_\xi \quad \forall \xi : 1 \le \xi \le c$
3. $\sum_{\xi=0}^{c} Pr(i_n=\xi) = 1 \quad \forall n : 1 \le n \le N$

Maximize: $H(\vec{p}) = \sum_{n=1}^{N} H(p_n)$

Subject to:

1. $\sum_{n=1}^{N}\sum_{\xi=0}^{c} \frac{(e^{g(\xi)}-1)\cdot Pr(i_n=\xi)}{lg(n+1)} / (\text{optDCG}) = ndcg$
2. $\sum_{n=1}^{N} Pr(i_n=\xi) = R_\xi \quad \forall \xi : 1 \le \xi \le c$
3. $\sum_{\xi=0}^{c} Pr(i_n=\xi) = 1 \quad \forall n : 1 \le n \le N$

Figure 1: Maximum entropy setup for GAP and nDCG, respectively.

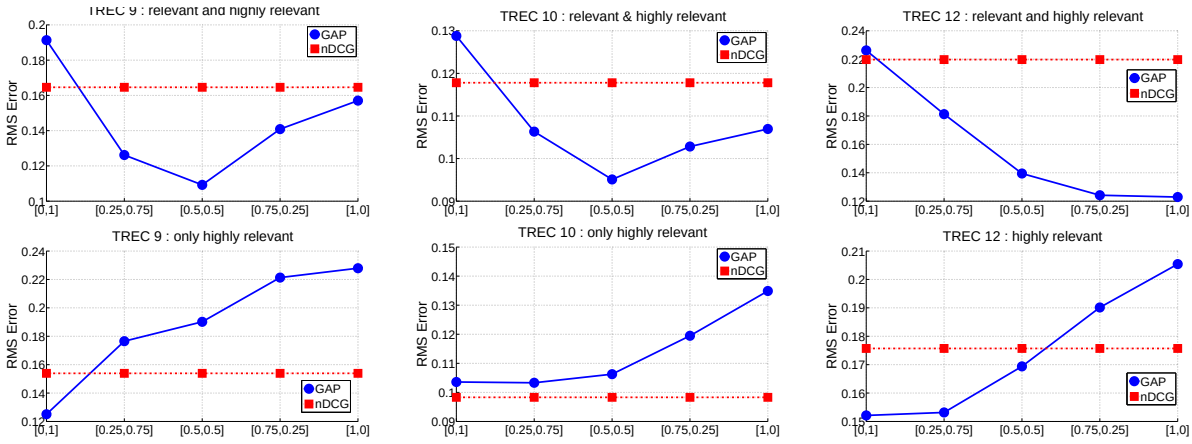

Figure 2: Mean RMS error between inferred and actual PR curves when only highly relevant documents are considered as relevant and when both relevant and highly relevant documents are considered as relevant.

gle grade. Therefore, one can compare the informativeness of GAP with the informativeness of AP by comparing the right-most point on the GAP curve with any other point on the same curve. For instance one can compare GAP with equal thresholding probabilities ($g_1 = g_2 = 0.5$) with AP by comparing the point on the blue line that corresponds to the [0.5,0.5] on the x-axis with the point on the blue line that corresponds to the [1,0] on the x-axis. This way we can test whether graded relevance add any value in the informativeness of the metric on the top of binary relevance. What is striking about Figure 2 is that in TREC 9 and 10 GAP (with $g_1 = g_2 = 0.5$) appears more informative than AP when relevant and highly relevant documents are combined (top row plots). That is, the ability to capture the sequence of relevance regardless the relevance grade is benefited by differentiating between relevant and highly relevant documents.

4.2 Discriminative Power

A number of researchers have proposed the evaluation of effectiveness metrics based on their discriminative power.That is, given a fixed set of queries, which evaluation metric can better identify significant differences in the performance of systems? By utilizing the framework proposed by Sakai [18], based on the *Bootstrap Hypothesis Testing* and using data from TREC 9, 10 and 12, we observed that the GAP metric appeared to outperform nDCG over TREC 12 data while the opposite was true for TREC 9 and 10. When limiting our experiments to the best performing systems (top 15 by both metrics), GAP consistently outperformed nDCG in all three data sets. The results for TREC 9 are illustrated in Figure 3. Due to space limitations we omit the figures from TREC 10 and 12. In the figure the more towards the origin of the axes the curve is the more discriminative the metric is. The inner plot corresponds to the test over the best performing systems.

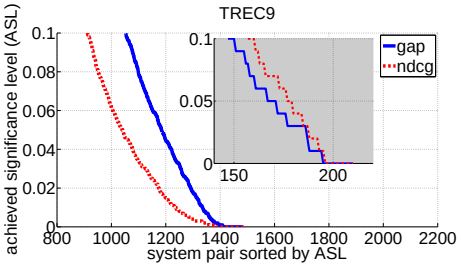

Figure 3: Discriminative power based on bootstrap hypothesis tests for TREC 9.

5. GAP FOR LEARNING TO RANK

Finally, we employed GAP as an objective function to optimize for in the context of LTR. For comparison pur-

		Test Metric		
		nDCG	AP	PC(10)
SoftRank	Opt nDCG	0.6162	0.6084	0.5329
	Opt GAP	**0.6290**	**0.6276**	**0.5478**
	Opt AP	0.6129	0.6195	0.5421
LambdaRank	Opt nDCG	0.6301	0.6158	0.5355
	Opt GAP	**0.6363**	**0.6287**	**0.5388**
	Opt AP	0.6296	0.6217	0.5360

Table 1: Test set performance for different metrics when SoftRank and LambdaRank are trained for nDCG, GAP, and AP as the objective over 5K Web Queries from a commercial search engine.

		Test Metric		
		nDCG	AP	PC(10)
SoftRank	Opt nDCG	0.4665	0.4452	0.4986
	Opt GAP	**0.4747**	**0.4478**	**0.5001**
	Opt AP	0.4601	0.4448	0.4900
LambdaRank	Opt nDCG	0.4585	0.4397	0.5005
	Opt GAP	**0.4665**	**0.4432**	**0.5042**
	Opt AP	0.4528	0.4408	0.4881

Table 2: Test set performance for different metrics when SoftRank and LambdaRank are trained for nDCG, GAP, and AP as the objective over the OSHUMED data set.

poses we also optimized for AP and nDCG. In our experiments we employed two different learning algorithms, (a) SoftRank [22] and (b) LambdaRank [6] over two different data sets, (a) a Web collection with 5K queries and 382 features taken from a commercial search engine, and (b) the OHSUMED collection provided by LETOR [21]. The relevance judgments in the both data set are in a 3 grade scale (non-relevant, relevant and highly relevant). Five-fold cross validation was used in the case of OHSUMED collection.

Since the informativeness of the metric is well correlated with the effectiveness of the constructed ranking function, we select g_1 and g_2 based on the criterion of informativeness. As we observed in Section 4.1, the values of g_i that result in the most informative GAP variation is $g_1 = g_2 = 0.5$. Intuitively, these values of g_i indicate that highly relevant documents are "twice as important as relevant documents.

LTR algorithms: *SoftRank* [22] is a neural network based algorithm that is designed to directly optimize for nDCG, as most other learning to rank algorithms. Since most IR metrics are non-smooth as as they depend on the ranks of documents, the main idea used in SoftRank to overcome the problem of optimizing non-smooth IR metrics is based on defining smooth versions of information retrieval metrics by assuming that the score s_j of each document j is a value generated according to a Gaussian distribution with mean equal to s_j and shared smoothing variance σ_s. Based on this, Taylor et al. [22] define π_{ij} as the probability that document i will be ranked higher than document j. This distribution can then be used to define *smooth* versions of IR metrics as expectations over these rank distributions.

Based on these definitions, we extend SoftRank to optimize for GAP by defining SoftGAP, the *expected* value of Graded Average Precision with respect to these distributions and compute the gradient of SoftGAP.

Given the probabilistic interpretation of GAP defined earlier and the distribution π_{ij}, the probability that document i will be ranked higher than document j, SoftGAP can be computed as follows:

Let PC_n be:

$$PC_n = \frac{\sum_{j=1}^{i_n} g_j + \sum_{m=1}^{N} \pi_{mn} \sum_{j=1}^{min(i_m, i_n)} g_j}{\sum_{m=1, m\neq n}^{N} \pi_{mn} + 1}$$

$$\text{then } SoftGAP = \sum_{n=1}^{N} \frac{PC_n}{\sum_{i=1}^{c} R_i \sum_{j=1}^{i} g_i}$$

Optimizing for an evaluation metric using neural networks and gradient ascent requires computing the gradient of the objective metric with respect to the score of an individual document $s_{\bar{m}}$. To compute the gradients of SoftGAP, we use a similar approach as the one Taylor et al. [22] used to compute the gradients of nDCG. Detailed derivations for the computation of the gradients are omitted due to space limitations.

LambdaRank [6] is another neural network based algorithm that is also designed to optimize for nDCG. In order to overcome the problem of optimizing non-smooth IR metrics, LambdaRank uses the approach of defining the gradient of the target evaluation metric only at the points needed.

Given a pair of documents, the virtual gradients (λ functions) used in LambdaRank are obtained by scaling the RankNet [5] cost with the amount of change in the value of the metric obtained by swapping the two documents [6].

Following the same setup, in order to optimize for GAP, we scale the RankNet cost with the amount of change in the value of GAP metric when two documents are swapped. This way of building gradients in LambdaRank is shown to find the local optima for the target evaluation metrics [7]. Detailed derivations for the computation of the virtual gradients for LambdaRank are also omitted due to space limitations.

Results: Tables 1 and 2 show the results of training and testing using different metrics. In particular the rows of the table correspond to training for nDCG, GAP and AP, respectively. The columns correspond to testing for nDCG at cutoff 10, AP and precision at cutoff 10. As it can be observed in the table training for GAP outperforms both training for nDCG and AP, even if the test metric is nDCG or AP respectively. The differences among the effectiveness of the resulting ranking functions are not large, however, (1) most of them are statistically significant, indicating that the fact that GAP outperforms AP and nDCG is not a results of any random noise in training data, (2) GAP consistently leads to the best performing ranking function over two radically different data sets, and (3) GAP consistently leads to the best performing ranking function over two different LTR algorithms. Thus, even if the differences among the constructed ranking functions are not large, optimizing for GAP can only lead to better ranking functions.

These results strengthen the conclusion drawn from the discussion about the informativeness of the metrics. First, it can be clearly seen that even in the case that we care about a binary measure (AP or PC at 10) the utilization of multi-graded relevance judgments is highly beneficial. Furthermore, these results suggest that even if one cares for nDCG at early ranks, one should still train for GAP as opposed to training for nDCG.

6. CONCLUSIONS

In this work we constructed a new metric of retrieval effectiveness (GAP) in a systematic manner that directly generalizes average precision to the multi-graded relevance case. As such, it inherits all desirable properties of AP: it has a nice probabilistic interpretation and a theoretical foundation; it estimates the area under the non-interpolated grade precision-recall curve. Furthermore, the new metric is highly informative and highly discriminative. Finally, when used as an objective function for learning-to-rank purposes GAP consistently outperforms AP and nDCG over two different data sets and over three different learning algorithms even when the test metric is AP or nDCG itself.

7. 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 G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, August 2005.

[3] P. Bailey, N. Craswell, A. P. de Vries, I. Soboroff, and P. Thomas. Overview of the trec 2008 enterprise track. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008)*, 2008.

[4] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 667–674, New York, NY, USA, 2008. ACM.

[5] 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 Press.

[6] C. J. C. Burges, R. Ragno, and Q. V. Le. Learning to rank with nonsmooth cost functions. In B. Schölkopf, J. C. Platt, T. Hoffman, B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *NIPS*, pages 193–200. MIT Press, 2006.

[7] P. Donmez, K. M. Svore, and C. J. Burges. On the local optimality of lambdarank. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 460–467, New York, NY, USA, 2009. ACM.

[8] 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 Press.

[9] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[10] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA, 2005. ACM.

[11] E. Kanoulas and J. A. Aslam. Empirical justification of the gain and discount function for ndcg. In *To appear in CIKM '09: Proceedings of the 18th ACM international conference on Information and knowledge management*, 2009.

[12] 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.

[13] J. Kekäläinen and K. Järvelin. Using graded relevance assessments in ir evaluation. *J. Am. Soc. Inf. Sci. Technol.*, 53(13):1120–1129, 2002.

[14] T. Minka, J. Winn, J. Guiver, and A. Kannan. Infer.net user guide : Tutorials and examples.

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

[16] S. Robertson. A new interpretation of average precision. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 689–690, New York, NY, USA, 2008. ACM.

[17] T. Sakai. *Ranking the NTCIR Systems Based on Multigrade Relevance*, volume 3411/2005 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, February 2005.

[18] 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.

[19] 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.

[20] T. Sakai and S. Robertson. Modelling a user population for designing information retrieval metrics. In *The Second International Workshop on Evaluating Information Access (EVIA 2008) (NTCIR-7 workshop) Tokyo, December 2008*, 2008.

[21] J. X. Tao Qin, Tie-Yan Liu and H. Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval Journal*, 2010.

[22] M. Taylor, J. Guiver, S. E. Robertson, and T. Minka. Softrank: optimizing non-smooth rank metrics. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 77–86, New York, NY, USA, 2008. ACM.

[23] 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.

[24] J. Xu and H. Li. Adarank: a boosting algorithm for information retrieval. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 391–398, New York, NY, USA, 2007. ACM.

[25] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In P. S. Yu, V. Tsotras, E. Fox, and B. Liu, editors, *Proceedings of the Fifteenth ACM International Conference on Information and Knowledge Management*, pages 102–111. ACM Press, November 2006.

[26] E. Yilmaz and S. Robertson. Deep versus shallow judgments in learning to rank. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 662–663, New York, NY, USA, 2009. ACM.

[27] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2007. ACM Press.