

**Computing Maximized Effectiveness Distance for Recall-Based Metrics**

**Abstract:**

Given an effectiveness metricMð\_Þ, two ordered document rankings X1 and X2 generated by a score-based information retrieval activity, and relevance labels in regard to some subset (possibly empty) of the documents appearing in the two rankings, Tan and Clarke’s Maximized Effectiveness Distance (MED) computes the greatest difference in metric score that can be achieved that is consistent with all provided information, crystallized via a set of relevance assignments to the unlabeled documents such that jMðX1Þ \_MðX2Þj is maximized. The closer the maximized effectiveness distance is to zero, the more similar X1 and X2 can be considered to be from the point of view of the metric Mð\_Þ. Here, we consider issues that arise when Tan and Clarke’s definitions are applied to recall-based metrics, notably normalized discounted cumulative gain (NDCG), and average precision (AP). In particular, we show that MED can be applied to NDCG without requiring an a priori assumption in regard to the total number of relevant documents; we also show that making such an assumption leads to different outcomes for both NDCG and average precision (AP) compared to when no such assumption is made.

**Existign System:**

Recently observed that in some situations measuring the difference between the sets of elements in the two document rankings via RBO may be less useful than computing the extent to which the two rankings can give rise to differing scores when measured via an effectiveness metric Mð\_Þ relative to a set of partial relevance judgments. For example, consider the following two rankings, X1 and X2, in which the notation indicates that the top-ranked document in X1 is A, and that it is non-relevant (that is, has an associated gain of 0); that the second-ranked document in X1 is B and that its relevance is unknown.

**Proposed System:**

Introduce a sequence of three effectiveness metrics: cumulative gain, CG; discounted cumulative gain, DCG; and normalized discounted cumulative gain, NDCG. Originally presented as vector-based computations that span every possible evaluation depth k, all three are usually now evaluated to some fixed depth k in the same way as precision is, with the choice of k influencing the behavior of the metric.

To understand the difference, note that the conversion factor Sk used in SDCG represents the maximum DCG score that can ever be attained on any ranking of k items, and that an SDCG@k score of 1.0 corresponds to the situation in which every one of the first k documents is maximally relevant. In contrast, NDCG@k employs a conversion factor usually referred to as “ideal DCG” that aggregates the known relevance information for the particular topic in question, in order to provide an effectiveness score range in which 1.0 represents “the best that can be done for this ranking”, and may not require that all of the first k documents be relevant.