Reprinted from: *Information Retrieval* (vol.10, 2007, 321–339).
The original article is available from www.springerlink.com.

# On rank-based effectiveness measures and optimization

Stephen Robertson* and Hugo Zaragoza†

July 6, 2007

**Keywords**
effectiveness metrics; ranking functions; optimization

**Abstract**

Many current retrieval models and scoring functions contain free parameters which need to be set – ideally, optimized. The process of optimization normally involves some training corpus of the usual document-query-relevance judgement type, and some choice of measure that is to be optimized. The paper proposes a way to think about the process of exploring the space of parameter values, and how moving around in this space might be expected to affect different measures. One result, concerning local optima, is demonstrated for a range of rank-based evaluation measures.

## 1 Parameters and optimization

This paper addresses some basic features of many of the measures of retrieval effectiveness in current or past use. The objective is to understand aspects of their behaviour, with particular reference to the optimization of parameters of a retrieval model or ranking function. This is a theoretical paper.

Many current models of retrieval, and ranking functions derived from them, contain free tunable parameters. A constantly recurring theme of

*Stephen Robertson / Microsoft Research / 7 JJ Thomson Avenue / Cambridge CB3 0FB, UK / ser@microsoft.com

†Hugo Zaragoza / Yahoo! Research / Ocata 1 / Barcelona 08003, Spain / hugoz@es.yahoo-inc.com

work in this field is the use of experimental data of the sort traditionally used for retrieval system evaluation as training data on which to tune such parameters. Almost all contributions to (for example) the annual TREC test involve some degree of learning/training some aspects of the method on the basis of previous test data. Retrieval methods and ranking functions may have any number of tunable features or parameters, from one to several hundred (typically, web search engines come in at the upper end of this scale).

Before attempting any form of optimization, we have to decide which measure or measures we would like to optimize: the measure of retrieval effectiveness. In the present paper we consider only the problem of optimizing a single measure. For a training set of documents, queries and relevance judgements, and any particular set of parameter values for our ranking function, we can evaluate our chosen effectiveness measure. Thus we regard the measure as a function on the space of all possible combinations of parameter values; we seek an optimum in this space. In principle, if the measure has some nice topological properties in this space (specifically continuity and convexity), there exist optimization methods that tune the parameters and find the maximum relevance setting in reasonable time. However, as we shall see, most such measures (including all the usual IR effectiveness metrics) have no such nice properties.

With one or a handful of parameters it is often possible to explore the parameter space exhaustively (at least to some degree of granularity). That is, we may define a grid of points over the parameter space, and specifically evaluate the measure at every point on the grid. With (say) 10 continuous parameters, this starts to become infeasible; one might instead explore parameters singly or in smaller sets, covering all in a heuristically-specified sequence of separate optimizations, possibly iterating the sequence. With (say) 100, it becomes almost imperative to consider other methods. The field of machine learning offers some such methods, often based on some form of gradient descent (or ascent) of the objective function, the function that we want to optimize. Essentially such methods follow a *trajectory* of points through parameter space – at each point a decision is taken as to where the next point should be.

But here is the rub: not one of the measure of retrieval effectiveness commonly used is suitable for gradient ascent. Measures based on ranked output are discontinuous, non-differentiable functions of the parameters, and cannot be used (at least in any simple straightforward way) for this purpose.

The primary purpose of this paper is to explore and further understand the characteristics of traditional rank-based effectiveness measures when these are regarded as functions in the parameter space. The basic tool is an analysis of the behaviour of the measures as we move along trajectories in parame-

ter space. It turns out that all rank-based measures share some interesting properties in this regard. In part at least, this behaviour is probably understood at an intuitive level by people working in the field. However, we believe that the formal analysis does yield some surprising results which will help to improve our understanding. The results apply to parameter spaces of any dimensionality.

The paper does not attempt to provide solutions to the problems raised. Some practical consequences are discussed, but the main aim is understanding.

In Section 2 we develop a characterization of traditional effectiveness measures, which allows us in Section 3 to generalize about a wide range of rank-based measures. In this section we develop the central idea of the paper, which concerns the behaviour of measures of effectiveness as we move around the parameter space; the theory is explored more fully in the appendix. In Section 4 we consider some alternative approaches, score-based measures and non-gradient-descent methods. We present some conclusions in the final section.

# 2 Measures of effectiveness

The measures proposed for or used in retrieval tests may be characterized in any number of different ways. The present characterization is primarily intended to emphasize the similarities rather than the differences, and to identify a significant subset of the totality of measures about which we can generalize.

## 2.1 Set-based measures

The original measures of Recall and Precision were defined in terms of a set of retrieved documents. The output of the system was assumed to be a yes-no judgement on each document, thus a single undifferentiated set of documents is retrieved from the collection. The F and FBeta measures commonly used in text categorization are similar, as are the various measures (often forms of utility) used in the various Filtering tasks at past TRECs (see e.g. Robertson & Soboroff 2002). As we have seen in the latter case (Robertson 2002), a retrieval function typically involves both a ranking function and a threshold, both of which have to be optimized. This is a complication not considered further in the present paper.

For convenience, we reiterate the definitions of Recall (proportion of all the relevant documents that are retrieved) and Precision (proportion of all the

retrieved documents that are relevant). F (FBeta) is a (weighted) harmonic mean of Recall and Precision, and Utility is typically defined as a function of the number of relevant documents retrieved discounted by the number of nonrelevant retrieved.

## 2.2 Rank-based measures

Most measures in common use, including some adaptations of the above set-based measures, assume system output in the form of a ranking of documents. This ranking normally derives from a scoring function: in relation to a given query, each document is assigned a real-value score, and the documents are ranked according to this score. Ranking is assumed to start at the highest score (the document which seems most likely to satisfy the information need).

For convenience in later discussion, some of these measures are briefly defined below; it is assumed for the purpose of definition that (a) the ranking is complete, with no ties; (b) all documents are judged for relevance, on either a binary or a discrete ordered scale; (c) we are dealing with a single query/topic only. All these matters are discussed further below.

**Average Precision (AP):** Determined by locating the relevant documents in the ranking, calculating precision at those points, and averaging over all relevant documents for the query (this is 'non-interpolated' AP).

**Precision at $n$ (P@$n$):** Defined as precision calculated at rank $n$, for any integer $n$.

**RPrecision (RPrec):** Defined as precision calculated at rank $R$, where $R$ is the total number of relevant documents for this query.

**Recall (Rec):** In form used in TREC and elsewhere, defined as recall at rank $n$, for some integer $n$ which is usually set to 1000.

**Discounted Cumulative Gain (DCG):** To each relevance grade, a utility or gain figure is assigned. To each rank position, a discount is assigned (usually following a simple formula). DCG is calculated by accumulating the gain, discounted by the discount, as we traverse the ranking. Note that this definition specifies a whole family of measures, depending on the choice of both gain parameters and discount function (Järvelin & Kekäläinen 2000).

**Search Length:** The number of non-relevant documents seen by the time the user reaches the $n$th relevant document, for some integer $n$ (Cooper 1968).

**Correct Pairs:** The number of pairs of documents which are ranked in the correct order, taken from a list of $n$ candidate pairs of document (relevant/non-relevant or more/less relevant). Note: this measure itself is not used in evaluation, but is the basis for **bpref**, discussed below.

**Success at $n$ (S@$n$):** Defined as whether or not a relevant document has been retrieved (irrespective of how many) by rank $n$, for some integer $n$.

**Reciprocal Rank:** Defined as the reciprocal of the rank of the first relevant document.

All of the above measures depend to some degree on the actual ranking of the documents. All of them also place very strong emphasis on the early part of the ranking – at least in normal retrieval circumstances, where both the total number of relevant documents and the chosen value of $n$ in any of the above is orders of magnitude smaller than the collection size (but see below for discussion of Correct Pairs and bpref).

### Multiple queries

All the above measures are typically averaged over a set of queries. In the cases of measures which are naturally between zero and one, this is straightforward (below we refer to Mean Average Precision, MAP, and Mean Reciprocal Rank, MRR). For those with arbitrary scales, it is usual to normalize them per-query, so as to give equal weight to each query. Thus:

**Normalized Discounted Cumulative Gain (NDCG):** DCG is normalized by dividing by its maximum possible value for the query. This maximum is calculated by notionally ranking the documents in the optimum order for the query (all documents of higher relevance before all documents of lower relevance). NDCG can then be averaged over queries.

**Search Length:** Cooper does not propose a method for averaging search length over queries.

**Correct Pairs:** As with DCG, this must be divided by its maximum, which is the number of candidate pairs chosen for the measure (see discussion of bpref below).

All the above methods of averaging start from the principle that each *query* is to be given equal weight (as opposed, for example, to averaging

non-normalized DCG, which would give more weight to queries with many relevant documents). Without departing from this principle, we may seek to reweight different *ranges of the measure* concerned. A current approach to emphasizing poor performance is to take a geometric mean of the measure rather than the usual arithmetic mean – or equivalently to apply a log transformation to the measure before averaging. All the arguments in this paper apply to geometric means such as GMAP as well as to arithmetic means such as MAP.

### Unjudged documents

Usually not all documents in the collection have been judged for relevance. We may have the situation that a pool including all the top-ranked documents from all the searches being evaluated has been judged. Or it may be that such a pool has been judged, but we are evaluating a new search, which might retrieve some unjudged documents. In this case it might be a reasonable assumption that most of the top-ranked have been judged, and furthermore that those that have not been judged are most likely to be non-relevant. For most of the above measures, the assumption that all unjudged documents are non-relevant provides a reasonable approximation in such circumstances, because of the built-in bias to the early part of the ranking. The exception is Correct Pairs, which has no such bias and would be unworkable if we counted all known-relevant / unjudged pairs as well as all known-relevant / known-nonrelevant pairs. The simple choice for Correct Pairs is to consider only judged documents (but see the discussion of bpref that follows).

When the relevance judgements are thought to be seriously incomplete, these arguments look more suspect. The bpref measure was introduced by Buckley & Voorhees (2000) to deal with this situation. It involves the Correct Pairs measure (normalized by its maximum per-query and then averaged over queries). However, the main issue concerns the choice of pairs. Out of various methods investigated by Buckley and Voorhees, they choose the following: take all R known relevant documents for the query, and the top R ranked known non-relevant documents, making $R^2$ candidate pairs (this value $R^2$ is now the maximum number of correct pairs). This heuristic gives bpref the charateristic that the other measures have, of being strongly biased towards the early part of the ranking. It also gives some degree of stability and correlation with MAP, according to their experiments.

Note that such heuristics may depend on the judging regime. In the case of TREC corpora (used by Buckley and Voorhees), judgements have usually been made on deep pools of documents taken from the output of a variety of different search systems for each query. This means that there is

almost always a large number and variety of judged non-relevant documents – certainly many times more than judged relevant. Even when they sample to simulate missing judgements, this remains the case. However, if judgements are based on shallow pools from one or a small number of systems, different methods may be called for.

### Truncation and ties

The above discussion assumes that there are no ties and that the ranking for each query is complete (the entire collection is ranked). Most practical ranking methods involve far-from-complete rankings, which we might regard as having a single bucket at the end containing a very large number of ties. In addition, there may be ties earlier in the ranking.

Measures like P@$n$ depend only on the top $n$ ranks; $n$ is usually chosen so that the tail-end bin is not an issue. Measures like MAP are usually truncated at some $n$; the precision at the rank of any relevant document that is not in the top $n$ is assumed to be zero; this is a reasonable approximation. Similar considerations apply to NDCG. Correct Pairs in the form of bpref is similarly not affected by the tail-end bin.

Ties earlier in the ranking have to be dealt with; the full specification of any measure has to include a suitable method. Note however that many modern ranking algorithms use such a variety of features that ties (apart from the large end bucket) are exceedingly unlikely. Note also that (in a somewhat earlier era) Cooper (1968) devoted most of his Search Length paper to dealing with ties. The name he gave to the measure is Expected Search Length; the 'expected' refers solely to the tie situation.

We may deal with (i.e. break) ties in a way that depends on the measure (e.g. pessimistic), or in an arbitrary way, either random or determinate. In effect the Cooper method was random. In the determinate category, we may have some arbitrary unique document id, and break ties by means of a secondary ranking by id. For the purposes of the present paper, we will assume some arbitrary deterministic tie-breaking, determined by the documents only: that is, given any pair of documents which may obtain equal scores for one or more queries, we apply a predetermined ordering of the two documents, which is independent of the query.

We will encounter the problem of ties in a slightly different way below.

### Redundancy

The above measures make no reference to possible duplication or redundancy. The effect of redundancy is that a document that might on its own be judged

relevant becomes less useful because a document seen earlier by the user has made it more or less redundant. This effect was partially simulated in the filtering task at TREC by means of a non-linear utility measure, and possibly also by Cooper's Search Length. But the issue is attacked much more directly by various authors, including in recent work on INEX (Kazai, Lalmas & de Vries 2004). Note that this is not at all the same as the discount in DCG: the rank-based discount is applied irrespective of how many *relevant* documents have been seen earlier.

**Score-based measures**

A few measures have been based on the actual scores given to documents, rather than the resulting ranks. These are discussed further below.

# 3 Trajectories in parameter space

We now return to the consideration of parameters and parameter optimization. Assume that we have a ranking function with $n$ free continuous parameters, and that we are exploring the resulting $n$-dimensional parameter space. Further assume that the mode of exploration is a smooth trajectory through the space. Note that in practical optimization trajectories are not smooth – they normally involve a series of discrete steps. However, looking in principle at smooth trajectories will allow us to see characteristics which may not be visible in stepwise trajectories.

Now consider the behaviour (as we follow the trajectory) of *any one* of the above rank-based measures – let us say measure M.

## 3.1 Flips

A rank-based measure can only change when two documents switch ranks. Without loss of generality we can look only at pairs of documents that are currently neighbours in the ranking for any particular query. We assume for simplicity of this discussion that we are looking at just one query, although the argument does not depend on this assumption. A switch between two neighbouring documents in a ranking is described as a flip. As we follow the trajectory, our chosen measure M will remain the same until we encounter a flip, at which point it may change. A flip is good or bad or neutral, in the obvious sense that moving a relevant above a nonrelevant (or a more above a less relevant) is good.

More formally, we have relevance judgements $\mathrm{Rel}(q, d)$, which may be binary or multi-valued discrete, and a scoring function $s_{\mathbf{P}}(q, d)$, where $\mathbf{P}$

is the vector of parameters defining the specific scoring/ranking function, and $q$ and $d$ are query and document respectively. A flip occurs between documents $d_1$ and $d_2$ if the trajectory passes a point at which the sign of $s_{\mathbf{P}}(q, d_1) - s_{\mathbf{P}}(q, d_2)$ reverses. A flip is good if e.g. $\text{Rel}(q, d_1) > \text{Rel}(q, d_2)$ and the flip causes $s_{\mathbf{P}}(q, d_1) - s_{\mathbf{P}}(q, d_2)$ to go from negative to positive.

As before, we need to consider also the point at which the two scores are equal. We have already assumed a tie-breaking method, so that even given equal scores, the ranking is fully defined, and the flip actually occurs either upon reaching the flipping point, or upon leaving it. This matter is discussed further in the Appendix.

M may or may not respond to a flip – that is, it may ignore, treat as neutral, some good or bad flips. But if it does respond, it will reward a good flip and penalize a bad one;[1] and it will always be neutral about a neutral one. This statement applies *precisely* to every single one of the above rank-based measures, both for individual queries and for averages across queries. It is at least arguable that a measure which behaved differently would not be a sensible measure of effectiveness. Measures vary in regard to (a) which flips they respond to, and (b) how much they respond, but they never vary in the direction of response.

To make a formal statement of the above:

**Assertion 1** *Flips may be identified unambiguously (modulo the relevance judgements in a particular training set) as good or bad or neutral. Any reasonable rank-based measure of effectiveness, including each of those defined above, satisfies the following: as we move along a trajectory in parameter space, the measure:*

(a) does not change until a flip is encountered for one of the queries in the training set;

(b) does not change if the flip is neutral;

(c) may or may not change if the flip is good;

(d) may or may not change if the flip is bad;

but

(e) will only reward a good flip or penalize a bad one.

[1] All the measures defined above are positive effectiveness measures; in this context, 'M rewards' means 'M is increased by', and 'M penalizes' means 'M is decreased by'. However, obvious reversals occur if we consider a cost function where lower is better rather than an effectiveness measure.

A short summary will illustrate the choice of flips question: MAP and NDCG respond to all good or bad flips (or if truncated, to all such flips down to the truncation point). P@$n$ responds only to flips between ranks $n$ and $n+1$; S@$n$ similarly, but then only if there are no higher-ranked relevant documents. MRR responds only to flips involving the top-ranked relevant document. bpref responds only to flips involving the chosen pairs.

The diagrams at the end of the Appendix represent a two-parameter space; the straight lines represent flip boundaries. That is, a trajectory in this space encounters a flip when it crosses a boundary. The full details of this example are discussed in the Appendix, but the illustration may help the reader to visualize the situation.

## 3.2 Optima

Now suppose we have two measures $M_1$ and $M_2$ which respond to the same flips (such as MAP and NDCG based on binary relevance, or two different NDCG functions, with the same truncation point, or MAP and GMAP). Further, we assume that we seek to optimize effectiveness, in a certain model parameter space, according to one or other of these measures, and on a specified training set of documents, queries and relevance judgements. We have the following result:

**Theorem 1** *On the assumption that a trajectory encounters only one flip at a time, $M_1$ and $M_2$ will share all local optima.*

Suppose we start at a local optimum of $M_1$: that is, at a point in the parameter space such that, if we follow any trajectory away from this point, $M_1$ remains constant or gets worse. This means that, whatever trajectory we take, the first flip we encounter and to which $M_1$ responds must be a bad flip. Since $M_2$ responds in the same direction to the same flips as $M_1$, we must also have started at a local optimum of $M_2$. □

The mathematical detail of this result is discussed further in the Appendix. The assumption sounds a bit strong, but actually further discussion in the Appendix will specify it more precisely, and relax it somewhat. The relaxed assumption appears to be at least reasonable. The resulting version of the theorem is quite robust.

## 3.3 Example 1

We here show an example of the way in which flips control the optima. The example is a very simple one, consisting of the following:

- A single query.
- A collection of 100 non-relevant documents and two relevant a,b. We assume that a is highly relevant and b less so.
- A ranking function in the form of a linear combination of two features, controlled by a single parameter (the weight of feature 2). The trajectory in this case is simply a scan along this parameter.
- Effectiveness measures AP and NDCG. Since there is a single query involved, no averaging is required. For NDCG, the gain for b is fixed at 1, while that for a is varied, giving three variant NDCGs, denoted NDCG($g_{\mathsf{a}}$) where $g_{\mathsf{a}} = 1, 3, 10$.

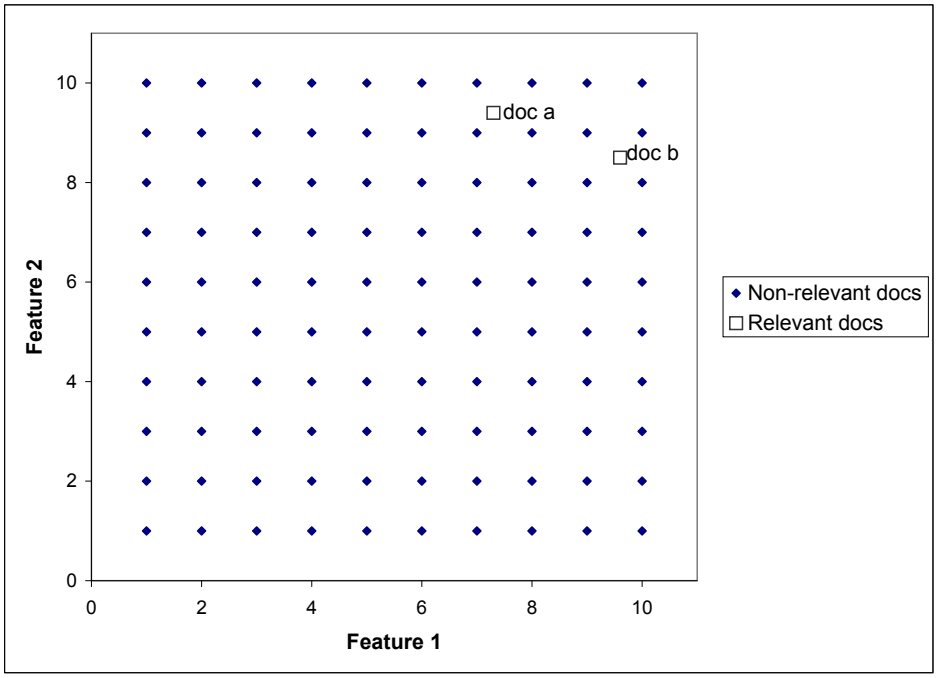


Figure 1: Example 1: one query, 100 non-relevant and two relevant documents. Two features. Figure shows the distribution of documents by feature.

The distribution of feature values is indicated in Figure 1; Figure 2 shows the four effectiveness measures as functions of the single parameter. Note the following:

- The measures agree almost entirely on which flips to respond to. The only disagreement is about a flip involving the two relevant documents

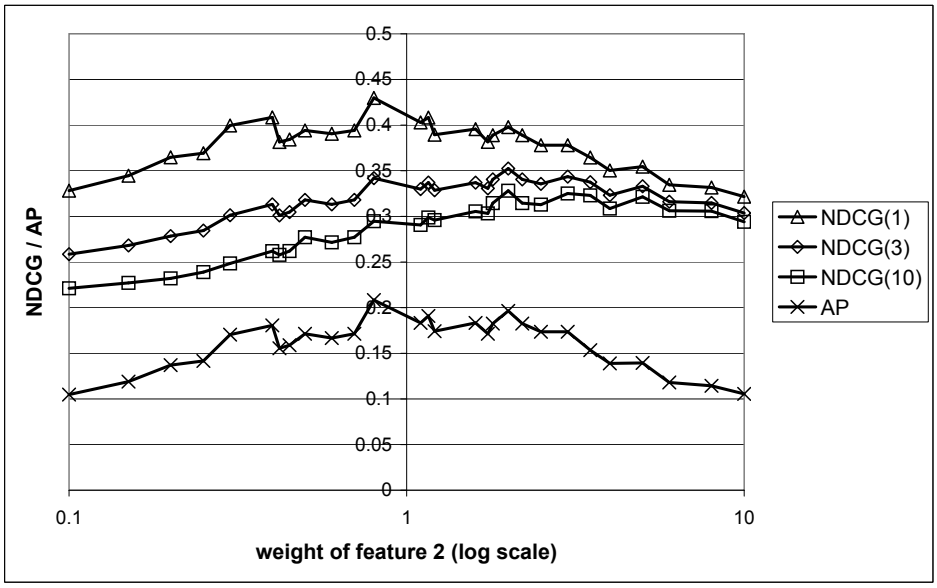


Figure 2: Example 1: effectiveness as a function of the weight of feature 2.

a,b (which occurs once on this trajectory). For AP and NDCG(1) this flip is neutral; NDCG(3) and NDCG(10) both prefer a to rank above b. The flip is represented by a small section of the graph (weight 2.5–3) where the former two measures are flat, but the latter two are not.

- With this single exception, the four graphs agree everywhere about local optima. Even in this simple example, there are 7 or 8 local optima.
- Each optimum is actually a small flat region; properly, the graph should be step-like, consisting of a series of horizontal line segments, joined by vertical drops at the exact points where flips occur. However, for simplicity a single point has been chosen in each flat region.
- AP and NDCG(1) agree on placing the global optimum at a weight of approximately 0.8, where a has rank 12 and b is at 4; the other two agree on a global optimum at about 2, where the ranks are 8 and 7 respectively.
- Other variations (for example different effectiveness measures, or NDCG with different discounts as well as gains) would promote to global status other local optima. Indeed, it would be possible to discover or invent a sensible measure to promote *any one* of the 8 local optima to global status.

### 3.4 Discussion of the theorem

Note that it is an empirically observed fact that the global optima of different measures differ. For example, in the TREC Robust Track (Voorhees 2006), it has been found that methods designed to enhance MAP are not necessarily good for GMAP. The above result might therefore seem slightly surprising. At minimum, it means that a global optimum of one of the measures *must* also be a local optimum (at least) of the other (MAP and GMAP at least agree precisely on which flips to respond to).

The reality is likely to be that there are many different local optima, in a given training collection with multiple queries. A measure that responds to many flips will have many optima; a measure that responds to fewer flips will have fewer optima, but larger flat areas (large areas of the parameter space that it cannot distinguish).

However, the shape of the measure surface in parameter space will depend a great deal on the sample size (size of the training set). The larger the training set, the more local optima there will be, but probably the less significant each optimum will be. As we enlarge the training set, we may imagine the surface on which we are trying to find an optimum showing finer and finer grain variations. At the same time, larger training sets appear to smooth things out, if we ignore the very small-scale variations. We may reach a point where a grid-based exploration will simply not see these small variations.

This paper includes no analysis of such large-sample behaviour. There are many interesting questions here. Given a large enough sample, the apparent smoothness might allow an approximate gradient analysis based on observing empirical gradients over finite grid steps. However, the underlying small-scale variation prohibits the use of gradient functions derived by differentiation on the rank-based measures.

## 4 Some alternative approaches

### 4.1 Score-based measures

In the machine learning community, there are well-understood criteria for objective functions which optimizers would like to find. Ideally, the objective function should be a continuous, differentiable, convex function of the parameters. Differentiability allows a gradient-descent type of procedure; convexity ensures that there is exactly one (global) optimum. It is very clear that none of the above rank-based measures comes near to satisfying these requirements.

Could we nevertheless define an effectiveness measure for IR which has these properties? One possible route would be to define the measure in terms of scores rather than ranks (since we would expect the scoring function to be continuous in the parameters at least). There has been a scattering of proposals for such measures, beginning with Swets (1963) in the 1960s.

There is a problem in principle with such measures. If we assume that the only user-visible output of the system is the ranking, then this output is independent of any monotonic transformation of the scores. An effectiveness measure defined on scores will in general be affected by any such transformation, and an optimum parameter set defined by optimizing such a measure will also generally be affected. It might be argued that the system can and should reveal scores as well as ranks, or even (Cooper, Chen & Gey 1994) that scores should be calibrated into realistic probabilities and then shown to users. Nevertheless, the likely outcome is that the user primarily responds to the ranking, and only secondarily to the scores. Thus an optimum that changes with monotonic transformations of the score seems perverse.

Nevertheless, measures based on scores might be candidates for gradient descent optimization, on the grounds that we may be able to discover (if not an optimum) at least a reasonably good parameter set for the rank-based measure that we really want to optimize. An early contribution along these lines is presented by Bartell (1994), and a similar approach is taken more recently in RankNet (Burges, Shaked, Renshaw et al. 2005). Here a measure based on scores, actually score *differences* between pairs of documents. This has the advantage of giving very limited monotonic-transformation independence (independence of zero-shifts), though also the disadvantage associated with bpref, that there has to be very careful choice of pairs. Bartell's (1994) measure is intended to estimate a rank correlation; Burges et al.'s (2005) measure is tested against NDCG.

There are some optimization methods and related techniques now being developed at the interface of machine learning and information retrieval. In the RankNet work, for example, the validation set heuristic is used not just to prevent overfitting the model to the training data, but also to help overcome the fact that the measure used in training is not the one we really care about. Some further developments are explored in Burges (2005). Other methods have been proposed in the general area of ranking optimization (for example in Freund, Iyer, Schapire & Singer 2003, Herbrich, Graepel & Obermayer 2000).

Such methods may give reasonable practical solutions. However, they leave one wondering if another method would have found something closer to a real global optimum. They render it difficult to reason about or have confidence in the result. There is a well-established principle in optimization

that one should optimize the objective function that one is really interested in. We seem to be being forced away from this principle.

## 4.2 Optimization without trajectories

So far we have discussed some characteristics of IR performance measures as we move parameters in some trajectory across the parameter space in an optimization procedure, and how this may affect optimization of parameters with gradient descent methods. There exist however some optimization methods that do not create a trajectory in parameter space and do not use gradient information. We briefly discuss the most important of these methods here and argue that they do not provide an effective alternative for parameter optimization in todays high-dimensional IR ranking functions.

Non-gradient methods treat the function as a black box and require only that it can be evaluated at any given point of its domain. In one dimension, *Brent's method* (Press, Teukolsky, Vettering & Flannery 2002) is guaranteed to find a local minimum relatively fast; this is done by bracketing the minimum and using a parabolic approximation of the remaining function to move quickly towards the minimum point. In higher dimensions, *direction-set methods* (Press et al. 2002) (e.g. *Powell's method* ) can be used; these methods call Brent's method iteratively in multiple dimensions to find the minimum. However, all of these methods require very large numbers of function evaluations, which in our case is extremely costly since each evaluation requires ranking the collection with respect to every query, sorting and evaluating the performance measure. Furthermore the storage requirements of these methods grow quadratic with the number of dimensions, which can be a problem when many hundreds of dimensions are used. In practice, such methods are only useful in IR for ranking functions of 20 parameters or less. Furthermore, all these methods rely on bracketing the minimum and will converge on a local minima. To have a chance of finding the global minima, the minimization procedure needs to be run multiple times from random starting points.

The Genetic Algorithms approach (Mitchell 1996) is an alternative to bracketing methods; here a population of possible solutions (leading to low values of the function) is perturbed in different ways to find, with high probability, the regions of the function domain with lowest values. Because there is a non-zero probability of visiting any domain region, there is a probabilistic guarantee that we will find the global (as well as the local) minima at some point, although this may require an unreasonable amount of time. More interestingly, the algorithm does not get stuck on local minima and iteratively improves the solutions. However it may take a very long time to

find reasonably good solutions and some heuristic needs to be designed to determine convergence. For these reasons, genetic algorithms are only used when other minimization methods (such as gradient-based or direction-set methods) cannot be.

# 5 Discussion and conclusions

Optimization over a substantial parameter space has become a significant issue for many current approaches to search. The measures that we use in evaluating search systems are central to optimization, and the nature and characteristics of the measures may make for difficulties in the choice or design of optimization methods.

In this paper, we have shown how a careful investigation of the way a measure might vary across a parameter space yields insights into the problem. In particular, it is clear that for a given training set, not only do the different IR effectiveness measures have different global optima, but they also probably have very many local optima. If we consider any one of these measures as a function in parameter space, it looks like very lumpy function, making many abrupt steps up or down as we move around parameter space. Many different measures share the same lumps, even if they differ substantially on the global optimum. We have expressed this result in a way which is independent of the dimensionality of the parameter space, the performance measure and the ranking method used, and in particular applies to ranking algorithms with many parameters.

We regard this insight as valuable in its own right, as part of our general understanding of the rank-based measures and of the optimization problem. But the insight also leads to various suggestions concerning implications and future work:

- There is scope for further investigation of smooth cost functions which nevertheless approximate or predict the rank-based functions.
- There is scope for the better understanding of the statistical behaviour of the rank-based functions under discrete steps involving many flips.
- When choosing an optimization method, one should seek a method that avoids getting stuck in local optima.

The first suggestion is already the subject of some work, mostly using smooth cost functions based on scores. However, as noted in the previous section, it is hard to see how such a function can avoid the problem concerning monotonic transformations of the scores. The second looks more difficult to

get a handle on, but goes to the heart of the problem, in the sense that we are not interested in the training set *per se*, but in generalizing from it. As indicated in Section 4.2 there are optimization methods which target the third item. It is not obvious that existing methods are particularly suited to this task, but at least there is scope for developing them.

# 6 Acknowledgements

Thanks to Chris Burges, Michael Taylor and Nick Craswell for many discussions on optimization issues.

# References


Bartell, B. (1994), Optimizing ranking functions: A connectionist approach to adaptive information retrieval, Technical report, University of California, San Diego. PhD Thesis.

Belkin, N. J., Ingwersen, P. & Leong, M.-K., eds (2000), *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York.

Buckley, C. & Voorhees, E. (2000), Evaluating evaluation measure stability, *in* Belkin, Ingwersen & Leong (2000), pp. 33–40.

Burges, C. J. C. (2005), Ranking as learning structured outputs, *in* S. Agarwal et al., eds, 'Proceedings of the NIPS 2005 workshop on Learning to Rank'. `http://web.mit.edu/shivani/www/Ranking-NIPS-05/`.

Burges, C. J. C., Shaked, T., Renshaw, E. et al. (2005), Learning to rank using gradient descent, *in* 'Proceedings of the 22nd International Conference on Machine Learning', Bonn.

Cooper, W. S. (1968), 'Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems', *American Documentation* **19**, 30–41.

Cooper, W. S., Chen, A. & Gey, F. C. (1994), Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression, *in* D. K. Harman, ed., 'The Second Text REtrieval Conference (TREC–2)', NIST Special Publication 500-215, Gaithersburg, MD: NIST, pp. 57–66. `http://trec.nist.gov/pubs/trec2/t2_proceedings.html`.

Freund, Y., Iyer, R., Schapire, R. & Singer, Y. (2003), 'An efficient boosting algorithm for combining preferences', *Journal of Machine Learning Research* **4**, 933–969.

Herbrich, R., Graepel, T. & Obermayer, K. (2000), Large margin rank boundaries for ordinal regression, *in* 'Advances in Large Margin Classifiers', MIT Press, pp. 115–132.

Järvelin, K. & Kekäläinen, J. (2000), IR evaluation methods for retrieving highly relevant documents, *in* Belkin et al. (2000), pp. 41–48.

Kazai, G., Lalmas, M. & de Vries, A. P. (2004), The overlap problem in content-oriented XML retrieval evaluation, *in* K. Järvelin, J. Allan, P. Bruza & M. Sanderson, eds, 'SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval', ACM Press, New York, pp. 72–79.

Mitchell, M. (1996), *An introduction to genetic algorithms*, Cambridge, MA: MIT Press.

Press, W. H., Teukolsky, S. A., Vettering, W. T. & Flannery, B. P. (2002), *Numerical Recipes in C++. The Art of Scientific Computing*, second edn, Cambridge University Press.

Robertson, S. E. (2002), 'Threshold setting and performance optimization in adaptive filtering', *Information Retrieval* **5**, 239–256.

Robertson, S. & Soboroff, I. (2002), The TREC 2001 filtering track report, *in* E. M. Voorhees & D. K. Harman, eds, 'The Tenth Text REtrieval Conference, TREC 2001', NIST Special Publication 500-250, Gaithersburg, MD: NIST, pp. 26–37. `http://trec.nist.gov/pubs/trec10/t10_proceedings.html`.

Swets, J. A. (1963), 'Information retrieval systems', *Science* **141**(3577), 245–250.

Voorhees, E. M. (2006), Overview of the TREC 2005 robust retrieval track, *in* E. M. Voorhees & L. P. Buckland, eds, 'The Fourteenth Text REtrieval Conference, TREC 2005', Gaithersburg, MD: NIST. `http://trec.nist.gov/pubs/trec14/t14_proceedings.html`.

# A Appendix: Refinement of the theorem

## A.1 Flip boundaries

We consider a flip between two documents $d_1$ and $d_2$ in the ranking for a given query $q$. In other words, the pairs $(q, d_1)$ and $(q, d_2)$ flip: their respective scores reverse their order.

In an $n$-dimensional parameter space, the flip boundary (the locus of points at which this flip occurs) is in general a hypersurface: that is, a manifold of dimension $n-1$, which divides the space into two parts. In one part, the scores satisfy $s_{\mathbf{P}}(q, d_1) > s_{\mathbf{P}}(q, d_2)$, and in the other, $s_{\mathbf{P}}(q, d_1) < s_{\mathbf{P}}(q, d_2)$. Exactly on the boundary, the two document scores are tied (but

see further below). Thus the flip boundary is also the locus of points at which the two documents score the same. We will call this flip boundary $\mathrm{FB}_q(d_1, d_2)$.

In general again, two such hypersurfaces intersect in a manifold of dimension $n-2$, and with a third in a manifold of dimension $n-3$, and so on. If a trajectory passes through one of these manifolds of dimension $n-m$, at the intersection of $m$ flip boundaries, then the $m$ flips may occur simultaneously.

This might suggest that the simple assumption that trajectories encounter only one flip at a time is bad. However, it also gives us the means to refine the argument. We first identify what the optimum of a rank-based measure looks like in the space, and revisit the ties problem. Next, we consider scoring functions that are linear in the parameters. In this (not realistic) case, the result can be made quite strong. We then consider other things that might happen in a non-linear case.

## A.2 Optima of rank-based measures

As we have seen, a rank-based measure can only change its value at a flip boundary. Therefore, in any region without internal flip boundaries, such a measure must be constant. There may also be boundaries for flips to which the measure is neutral. Thus a measure will be constant over a region of parameter space bounded by some set of significant flip boundaries; the region may also be crossed by other (non-significant = neutral for this measure) flip boundaries. If we are talking about the value of a measure for a single query, the flip boundaries relate to that query only; however, if we take an average over queries, the value may change as a result of flips relating to any of the queries. Thus the flip boundaries for different queries are superimposed on the parameter space.

A local optimum of a rank-based measure (in a training set) is a value which cannot be improved in the immediate locality. That is, if we have a region in which the measure is constant, bounded by various flip boundaries where the measure changes, then this region is an optimum for the measure if and only if every such change is bad for the measure. This definition applies to a single-query measure or to an average over queries. Clearly, in general, the more queries are included, the more significant flip boundaries there are, and therefore the smaller is the optimal region.

In general it is not the whole of a flip boundary which bounds an optimum, but a limited region of this boundary, determined by its intersections with other significant flip boundaries. Note also that the significance of a boundary may change at its intersection with other flip boundaries: a measure may in some circumstances respond to a specific flip between two documents, but

after other flips have occurred may no longer respond to the first.

A global optimum of a rank-based measure is one that cannot be improved anywhere. Note that because the set of values taken by a rank-based measure on a training set is finite, it is quite possible that two unconnected regions of the parameter space will share the same globally optimum value of the measure.

## A.3 Ties

Again, we need to consider ties. In the present discussion, any point exactly on a flip boundary represents a position (parameter setting) where the two documents have equal scores. Following the previous discussion on the subject, we assume that there is some arbitrary but predetermined way of breaking ties, so that at this parameter setting the documents are still assigned different ranks. This ranking of these two documents will match one or other side of the flip boundary.

Thus the boundary regions which bound the optimum of a measure will each either be inside the optimum region or outside it. A point on a trajectory that is inside one of the hypersurfaces will have a well-defined value of the measure (optimum or not). A trajectory can even remain inside a hypersurface without affecting the arguments in this appendix.

## A.4 Linear models

We suppose that the scoring function is linear in the parameters.[2] Now flip boundaries are hyperplanes (linear subspaces of dimension $n-1$), and any two hyperplanes normally intersect in a linear subspace of dimension $n-2$ etc. For readers who like visual imagery, imagine a two- or three-dimensional parameter space. In 2-space, for 'hyperplane' think of *straight line* and for 'linear subspace of dimension $n-2$' think of *point*. In 3-space, think of *plane* and *line* respectively. A toy 2-space example is represented in the diagrams in section A.7.

In this linear case, there is only one kind of thing that can go wrong, indicated by the assumption in the theorem below. Note that part (b) of the

[2]Parameters may be the weights of features which are to be combined linearly. However, given $n$ features, we would not normally have $n$ independent weights, but $n-1$, since (a) we would certainly want to exclude the possibility of setting all weights to zero, and (b) the ranking would be unaffected by a constant (non-zero) multiplier for all weights. So we consider here an $n+1$-dimensional feature space with $n$ independent linear parameters (for example we might fix the weight of one feature to unity). An alternative would be to fix the parameter space to the surface of the unit hypersphere in $n+1$-dimensional feature space; the theorem could be established just as strongly in this model.

assumption is somewhat complex, and is discussed further below. We assume as before two measures $M_1$ and $M_2$ which agree on which flips to respond to.

**Theorem 2** *On the assumption that (a) no two flip boundary hyperplanes coincide, and that (b) no two multi-flip subspaces coincide except where they logically must, $M_1$ and $M_2$ will share all local optima.*

Suppose as before that we start at a local optimum of $M_1$. This means, as indicated, that we are in a region of the space bounded by a number of flip-boundary hyperplanes, each of which represents a bad flip from the point of view of $M_1$, and therefore also from the point of view of $M_2$. A trajectory may cross just one of these boundaries, or may pass through an intersection of two or more of these boundaries. But since *all* these boundaries are bad, crossing two or more of them simultaneously is bad for both measures just as crossing one of them is. Even in the case where one boundary becomes neutral after the other has been crossed, it will still be true that crossing both simultaneously is bad for both measures. □

The assumption as stated here is less strong than the version stated in the text of the paper. As indicated in the proof, a trajectory may pass through a point at which two or more flips happen simultaneously. Nevertheless, the assumption needs further discussion, and in particular, part (b) requires some explanation and analysis.

## A.5 Explanation of the assumption

Part (a) of the assumption is clear enough: we assume that we have no two distinct flips that always occur at *exactly the same parameter values everywhere*. If two flip boundaries in the linear case do not coincide, then their intersection must of necessity be either a subspace of dimension $n-2$, or null; and furthermore, if they do intersect, then each hyperplane is divided in two by the intersection, with one part lying each side of the other hyperplane.

We can nevertheless imagine cases which would violate this assumption. If (from the point of view of one query) two documents look identical, in the sense that they share exactly all the features which affect the score for that query, then they will always get the same score as each other and their flips with any third document will always coincide. This could occur (for example) with a simple BM25 scoring function, if the two documents are of the same length and have the same $tf$s for each of the query terms. A slightly more complex example could cause two flip boundaries for different queries to coincide. However, modern scoring functions based on many features render these possibilities less and less probable. Just for example, web search engines

will typically distinguish two documents whose textual content is absolutely identical, because of any number of extrinsic differences (PageRank, incoming links / anchor text, url depth etc.).

Part (b) of the assumption requires more explanation. In one type of case, the multi-flip subspaces must coincide. If we consider the flip boundary of documents $d_1$ and $d_2$ for a given query, and the flip boundary of $d_2$ and $d_3$ for the same query, it is clear that the intersection defines a coincidence not only of these two flips, but also of $d_1$ / $d_3$ flips. Thus if we suppose that these two flip boundaries form part of the bounding set for the optimum of measure $M_1$, then a trajectory away from the optimum may, by passing through the intersection of these planes, cause a flip of $d_1$ and $d_3$ also, despite the fact that the flip boundary of $d_1$ and $d_3$ was not part of the bounding set.

However, this particular problem is resolvable. Note that the relevance of documents $d_1$ and $d_2$ must differ, ditto $d_2$ and $d_3$. We identify three distinct cases:

1. The shared document $d_2$ is the less relevant of the two in both cases;

2. the shared document is the more relevant of the two in both cases;

3. the shared document lies strictly between the other two in relevance.

Now we consider the boundary $\mathrm{FB}_q(d_1, d_3)$. We have assumed that $\mathrm{FB}_q(d_1, d_2)$ and $\mathrm{FB}_q(d_2, d_3)$ bound the region of optimum $M_1$. In the first case, $d_1$ and $d_3$ both score higher than $d_2$ in this region; in the second, they both score lower. It follows that in both of the first two cases, $\mathrm{FB}_q(d_1, d_3)$ must pass through the region of optimum $M_1$. Therefore $M_1$ must treat the $d_1$ / $d_3$ flip as neutral. $M_2$ must do the same, and the theorem holds.

In the third case, $\mathrm{FB}_q(d_1, d_3)$ passes through the intersection of $\mathrm{FB}_q(d_1, d_2)$ and $\mathrm{FB}_q(d_2, d_3)$, but does not cross the region of optimum $M_1$, because $d_2$ must always rank between the other two documents in this region. If the trajectory (starting from the optimum) passes through this 3-way intersection, we may get the $d_1$ / $d_3$ flip at the same time as the other two. However, this flip is necessarily bad; it must be treated as either bad or neutral for both measures. The theorem holds.

The assumption allows for this possibility (of a shared document and therefore perforce coincident flips), but otherwise excludes coincident multi-flip subspaces. This has similar status to the exclusion of coincident single-flip boundaries: one could imagine violations, but they seem unlikely.

## A.6 The general case

Generally, the theorem as expressed for the linear case applies to the non-linear case also (replace 'hyperplane' by *hypersurface* and 'subspace' by *manifold*). We can think of a few more possible violations: for example, two hypersurfaces may touch tangentially without crossing, a situation that is not possible in the linear case and which could cause the theorem to fail. However, we can again argue that these examples are unlikely to occur in practice.

A mathematician's response to all such issues might be the following. We are worried about finite numbers of discrete events 'accidentally' coinciding in a continuous space. A general solution is to introduce a small random element which does not make these occurrences impossible, but gives them zero probability. Thus if we add a small jitter to our scoring function, so that the score of every document-query pair has a small random addition, then the theorem becomes true 'almost certainly' or 'almost everywhere', meaning that the violations have probability zero. The jitter can be as small as we care to make it, and therefore not affect anything else. This suggestion is in addition to the method for resolving ties which was given in section 2.2; given a jitter which is associated with each document-query pair but does not vary with the search algorithm parameters, there will still be flip boundaries in parameter space within which ties occur.

Thus despite various possible violations of the assumptions, the basic idea of the theorem seems to be fairly robust. However, as discussed in the main text, the effects may only be visible at an extremely fine granularity.

## A.7 Example 2

Figure 3 represents the parameter space for a trivial example. We assume just one query, four documents (a,b,c,d), two of which (a,b) are relevant to the query, and two parameters in a linear model. The diagram shows the flip boundaries as straight lines. Each line is labelled with the two documents whose flip boundary it marks; the ranking of the four documents is shown in each region; within each region (without crossing a line) the ranking does not change. Note also the constraints on the boundaries; following the discussion in section A.5, the line representing the (a,b) flip *must* go through the intersection of the lines representing the (a,c) and (b,c) flips etc.

In the case of possibly significant boundaries (bold lines), the good side is shown by an arrow. Neutral boundaries are thin lines. Figures 4 and 5 show the optimal regions for some of the rank-based measures. The following notes interpret the diagrams.

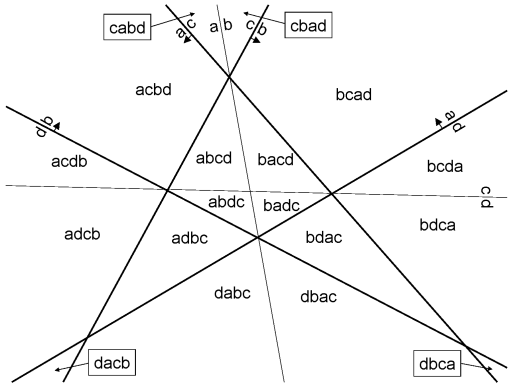


Figure 3: Parameter space for example 2: one query, two relevant documents (a,b), two non-relevant documents (c,d), and two parameters. The lines are flip boundaries, marked to indicate which documents flip. Each region is labelled with the document ranking that obtains in that region.

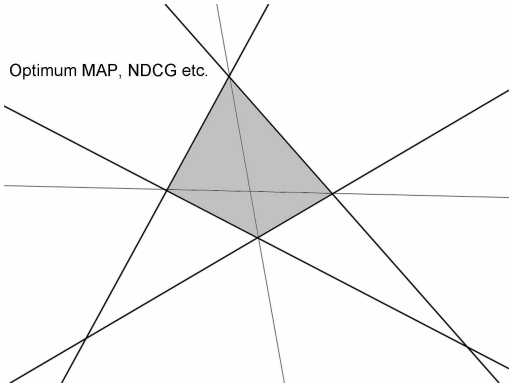


Figure 4: Region of optimum MAP, NDCG and some other measures, for example 2

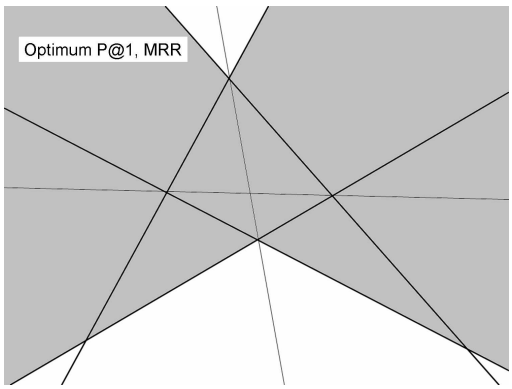


Figure 5: Region of optimum P@1, MRR and some other measures, for example 2

1. In this example, the perfect ranking (both relevant before both non-relevant) is possible, and is represented by the central quadrilateral.

2. Thus all measures will agree that this region is optimal; but some measures will ignore some boundaries, and thus extend the optimal region.

3. If we redefine the example so that c,d are relevant and a,b are not, then the perfect ranking is not possible. Nevertheless, there will still be much agreement between measures; the regions in which the relevant documents are first and third will be optimal by all measures.

4. Again, different measures may extend this optimum region, but in different ways. A measure like P@2 will be as happy with the first relevant in second position; MRR would prefer to drop the second relevant down the ranking.

5. Note that there are two disconnected regions (bottom left and bottom right) where the relevant documents are first and third; these will be distinct optima for a measure such as MAP. MAP will also have another local optimum in the region at the top where the relevant documents are first and fourth.

6. If we further redefine the example so that c is highly relevant and d less so, then the measures based on binary relevance are undefined. But here different NDCGs (with different gain or discount functions) may disagree with each other. We can achieve d at rank 1 and c at rank 3,

but if we want `c` at rank 1 the best we can do with `d` puts it at rank 4. Some NDCGs will prefer the former, some the latter.