Evaluation Measures for Preference Judgments

Ben Carterette
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
140 Governors Drive
Amherst, MA 01003
carteret@cs.umass.edu

Paul N. Bennett
Microsoft Research
One Microsoft Way
Redmond, WA 98052
pauben@microsoft.com

ABSTRACT

There has been recent interest in collecting user or assessor preferences, rather than absolute judgments of relevance, for the evaluation or learning of ranking algorithms. Since measures like precision, recall, and DCG are defined over absolute judgments, evaluation over preferences will require new evaluation measures that explicitly model them. We describe a class of such measures and compare absolute and preference measures over a large TREC collection.

Categories and Subject Descriptors: H.3.4 Information Storage and Retrieval; Systems and Software: Performance Evaluation
General Terms: Performance, Measurement
Keywords: evaluation, preference judgments

1. INTRODUCTION

Information retrieval systems are typically evaluated by calculating some evaluation measure (such as precision, recall, average precision, or DCG) over a set of *relevance judgments* made by human assessors. Relevance judgments have traditionally been made on an absolute scale, with each document judged independently of the others.

Recent work has suggested the use of implicit or explicit *preference judgments*: given two documents, an assessor only expresses a preference for one over the other. Preference judgments are interesting for several reasons:

1. Assessors make preference judgments faster than absolute judgments on a graded scale [1].
2. When preferences are transitive, they can be mapped to a measure of individual document utility that can be understood as an absolute relevance judgment [2].
3. Preferences more naturally reflect objective functions in *pairwise* learning-to-rank algorithms.

Preferences have some disadvantages, most notably the lack of defined evaluation measures for preference judgments and the polynomial increase in the number of preferences needed in a test collection. In this work we address the former by defining a suite of evaluation measures calculated over preference judgments. These measures have two additional desirable properties that go some way to addressing the latter problem: they require no assumption of preference for an unjudged pair of documents, and they remain stable when the set of preferences is dramatically reduced.

SIGIR'08, July 20–24, 2008, Singapore
ACM 978-1-60558-164-4/08/07.

2. PREFERENCE PRECISION AND RECALL

Retrieval systems are typically evaluated by some combination of *precision*, the proportion of retrieved documents that are relevant, and *recall*, the proportion of relevant documents that were retrieved. When "retrieved" is defined in terms of whether a document is ranked before some cutoff k, precision and recall can be calculated at any rank k.

To generalize to preference judgments, we define a few new terms. We will say a pair of documents (i, j) is *ordered* by the system if one or both of i, j appears above rank k. A pair is *unordered* if neither i nor j are above k. A pair is *correctly ordered* if the system's ordering matches assessor preferences, and *incorrectly ordered* otherwise.

We then define precision of preferences (*ppref*) as the ratio of correctly ordered pairs to ordered pairs. For example, at rank $k = 5$ a system has effectively specified an ordering of five documents, and for each of these, orderings in relation to the remaining $n - 5$ documents (where n is the total corpus size). This yields $5(5-1)/2 + 5(n-5) = 5n - 15$ ordered pairs, and more generally $k(2n - k - 1)/2$ ordered pairs.

Likewise, recall of preferences (*rpref*) is defined as the ratio of correctly ordered pairs to the total number of preferences made by assessors. For the example above, rpref would be the proportion of the full set of preferences that are correctly ordered among the $5n - 15$ ordered pairs.

Ties. Assessors may specify no preference between two documents, seeing them as equally relevant, or not specify anything about a pair at all. These pairs may be ordered by a system, but it is not immediately clear how they should be treated for calculation of ppref and rpref. The solution we adopt is to simply not count them as either ordered or unordered, excluding them from both numerator and denominator of ppref and rpref.

Transitivity. Assessors tend to make highly but not completely transitive preference judgments for documents [1, 2]. When preferences are not 100% transitive, it is impossible for the system to correctly order all documents. We will simply allow intransitivity to put an upper bound on ppref and rpref. Since intransitivity is rare, this should not be a serious problem; it can be seen as analogous to a machine learning problem with non-zero Bayes error.

Summary measures. Like traditional precision and recall, ppref and rpref can be plotted against each other for increasing k to create a precision-recall curve. ppref can be interpolated to create smooth curves, or averaged over ranks at which rpref increases, producing *average* precision of preferences (*APpref*).

Weighted preferences. Strictly speaking, precision and

	prec10	rec10	DCG10	NDCG10	MAP
ppref10	**0.968**	0.882	0.970	0.973	0.827
rpref10	0.892	**0.999**	0.905	0.917	0.880
wppref10	0.960	0.873	**0.971**	0.972	0.816
nwppref10	0.971	0.908	0.996	**0.998**	0.893
MAPpref	0.851	0.830	0.833	0.845	**0.984**

Table 1: Pearson correlations between pairwise and absolute measures averaged over 50 topics.

recall can only be calculated for binary relevance. *Discounted cumulative gain* (DCG) is a precision-like measure that supports graded (non-binary) relevance and discounting by rank. We can incorporate this idea into ppref and rpref as well, for when preferences have gradations ("strongly prefer", "slightly prefer", etc) and to discount pairs by rank. We define a weight w_{ij} for each pair of ranks (i, j). By analogy to a commonly-used formulation of DCG, we set

$$w_{ij} = \frac{2^{|pref_{ij}|} - 1}{\log_2(\min\{i, j\} + 1)}$$

where $pref_{ij}$ is the degree of preference between the documents at ranks i and j. *Weighted* precision of preferences (*wppref*) is the sum of the weights over ranks $j > i$ for which the documents are correctly ordered divided by the total weight of all ordered pairs. Normalized wppref (*nwppref*), like normalized DCG (NDCG), is wppref divided by the best possible wppref at the same rank.

3. EXPERIMENTS

There is not yet a large collection of preference judgments to experiment with, so we inferred preferences from absolute judgments in a TREC collection. The ad hoc portion of the Terabyte track at TREC 2005 used three levels of judgments: nonrelevant, relevant, and highly relevant ($rel = 0, 1$, and 2, respectively). Of 45,291 total judgments over TREC topics 751–800, 17% were labeled relevant and 5.8% highly relevant. On average, there were 906 judgments per query.

From these judgments we can extract preferences $i \succ j \Leftrightarrow rel_i > rel_j$, and for weighted measures, some of these can be considered "strong" preferences. On average, the 906 judgments per query became 142,435 preferences ($pref_{ij} = 1$ or 2), of which 34,823 were "strong" preferences ($pref_{ij} = 2$).

We obtained the ranked results of 58 retrieval systems over these topics. We averaged each evaluation measure over the set of topics; Table 1 shows the linear correlation between our preference-based measures and absolute-based measures. Each preference measure correlates highly with its analogous absolute measure (in bold), as well as with other absolute measures. When a system does well by a preference measure, it is likely to do well by an absolute measure as well.

Figure 1 shows traditional precision-recall curves for two systems (left), with preference precision-recall curves for comparison (right). While similar, the preference curve captures distinctions between levels of relevance and suggests that the upper curve is not only doing a better job of ranking relevant documents, but also of ranking highly relevant documents.

Stability. The "stability" of a measure is related to how consistently it is able to identify differences between systems over a sample of queries or topics. We can use analysis of variance (ANOVA) to measure stability. In an ANOVA, we compute the variance in a measure due to the systems being evaluated (MST) and compare it to the total residual error over all systems and topics (MSE). The greater this ratio

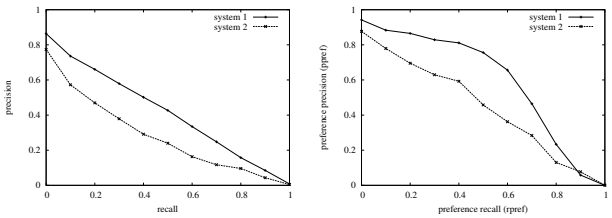

Figure 1: Precision-recall curves (left) and preference precision-recall curves (right) for two systems.

measure	F	measure	F	$F_{-99.4\%}$	$\tau_{-99.4\%}$
prec10	15.109	ppref10	11.686	10.815	0.900
rec10	5.570	rpref10	5.937	5.000	0.860
DCG10	14.273	wppref10	11.900	10.985	0.891
NDCG10	14.257	nwppref10	14.492	12.782	0.900
AP	38.136	APpref	45.981	43.149	0.976

Table 2: Analysis of variance in measures.

(the F-statistic) is, the greater the ability of a measure to distinguish between systems; in other words, fewer queries or topics are needed for evaluation.

The F for each measure is shown in Table 2. Larger F means more of the variance is due to the systems. The table shows that on average, preference measures have similar discriminative capabilities as their absolute counterparts.

Could the stability of preferences be a result of simply having many more judgments? We reduced the set by removing preferences on random document pairs i, j. Even after removing 99.4% of untied preferences—resulting in about 900 preferences per query on average—preference measures remain highly stable, as $F_{-99.4\%}$ in Table 2 shows. This is corroborated by the high rank correlation between a ranking of systems over a small set of preferences and the ranking over all preferences ($\tau_{-99.4\%}$ in Table 2).

4. CONCLUSION

We have presented a suite of evaluation measures that explicitly make use of preference judgments. We have shown that they correlate highly to absolute measures, are roughly as stable on average, and remain stable even with many preferences missing. While hampered here by lack of a large test collection of preferences, in general they can model much finer gradations of relevance; a clear future direction is to collect or simulate actual preferences. Based on the stability experiments, it seems that a very large collection is not necessary to study preference judgments. Furthermore, our weighted measures can be generalized with any gain or discount functions, and seem to allow for evaluation over a mix of absolute and preference judgments.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by Microsoft Live Labs. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

5. REFERENCES

[1] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: Preference judgments for relevance. In *Proceedings of ECIR*, pages 16–27, 2008.

[2] M. E. Rorvig. The simple scalability of documents. *JASIS*, 41(8):590–598, 1990.