A Test Collection of Preference Judgments

Ben Carterette
CIIR, UMass 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

Olivier Chapelle
Yahoo! Research
2821 Mission College Blvd
Santa Clara, CA 95054
chap@yahoo-inc.com

ABSTRACT

We describe an initial release of a set of binary preference judgments over a subset of the LETOR data. These judgments are meant to serve as a starting point for research into questions of evaluation and learning over non-binary, multi-item assessments.

1. INTRODUCTION

Information retrieval test collections traditionally contain binary judgments of the relevance of documents to queries. Recently there has been interest in generalizing to non-binary judgments: graded scales, aspect relevance, and preferences of the form "document A is preferred to document B". Preferences in particular are the foundation of several learning algorithms, such as the ranking SVM [6] and RankNet [2].

Though test collections exist for graded relevance and aspect relevance, most research on preferences to date has used preferences inferred from binary or graded judgments. We have undertaken to construct a test collection of true preference judgments. The initial release, described in more detail below, contains preference judgments for documents judged for the Topic Distillation task of the TREC 2003 Web track [5].

2. PREFERENCE TEST COLLECTION

The construction of this test collection is partially motivated by the use of preferences in algorithms for learning to rank. We decided to assemble preference judgments over a subset of the standard corpus used in learning-to-rank research: the LETOR (LEarning TO Rank) dataset [7].

The LETOR data consists of features and relevance judgments for about 1,000 documents judged for the TREC 2003 and 2004 Web tracks as well as features and relevance judgments for the medical abstracts in the OHSUMED corpus. This initial release includes preferences for the TREC 2003 queries only.

Topics. The topics are the 50 Topic Distillation topics from the TREC 2003 Web track. Each topic consists of a short title query and a longer description of the information need.

Corpus. The corpus is web pages in the .gov domain, crawled for the TREC 2003 Web track. LETOR provides features for the top 1,000 of these documents by an implementation of BM25 along with all documents judged relevant for the Topic Distillation task of the Web track. For this release, we have further restricted the corpus to all documents judged relevant for TREC, rounding out to 50 documents total by including the top documents ranked by LETOR's BM25 feature.

Judgments. Assessors were shown two documents and asked which they prefer for a given query. They could also judge a single document not relevant, effectively indicating that they would prefer every other (relevant) document to that one. They were not allowed to say two documents were equally relevant (except in the case of duplicates). Figure 1 shows a screenshot of the interface.

Based on our previous work [4], we assumed assessors would be transitive in their preferences and used that assumption to select pairs of documents to show. Along with the ability to judge documents nonrelevant, this reduced the total amount of effort needed to construct the collection. Generally one of the two documents was held fixed until it had been compared to all other documents.

The judgments were vetted and obvious errors corrected by one of the authors of this paper.

Agreement with TREC judgments. Our assessors did not always agree with the TREC judgments on whether a document was relevant or not. There are several explanations for this:

1. *Task.* The TREC judgments are for the Topic Distillation task, which is more specific than the traditional ad hoc task. Topic distillation focuses on retrieving home pages that act as gateways to the information and this is reflected in the judgments. Our preference judgments are closer to the ad hoc type, so that many pages that would have been judged nonrelevant for distillation are relevant (though perhaps not greatly preferred) for an ad hoc task by virtue of containing information about the topic.

2. *Noise in TREC judgments.* In the process of vetting our assessors' judgments, we found many documents that had been judged relevant for the distillation task but were tenuously related to the topic at best. Bernstein & Zobel reported a similar phenomenon in judgments for the GOV2 corpus [1].

3. *Noise in our judgments.* Judging preferences requires $O(n \log n)$ judgments, which for large n can become quite tiring. This may have contributed to noise in our preferences and disagreement with the TREC judgments.

SIGIR 2008 Workshop: Beyond Binary Relevance: Preferences, Diversity and Set-Level Judgments, July 24th, 2008, Singapore.
ACM 978-1-60558-164-4/08/07

Figure 1: A screenshot of the preference interface. The query and description are shown at the top. The two pages are shown in inline windows. The parent window can be resized to provide more space for the inline windows. Query terms are highlighted in the web pages. The progress indicator in the upper right lets the assessor know how close they are to completing the query.

3. DATA FORMAT

We have released two different versions of the preference judgments. The first gives the DOCNOs in the .gov corpus. The second uses the docids in the LETOR dataset and is intended for joining with those features.

Each line of both sets has four fields: the query number, two document IDs, and the preference judgment. A judgment of -1 indicates the the first document ID was preferred to the second; 1 indicates the opposite. A 0 means the two documents were judged to be duplicates. If either of the docids is NA, then the other document was judged to be nonrelevant. A 2 or -2 in the judgment column indicates this explicitly.

For the purposes of distribution, the preferences for each topic were reduced to the 50 necessary to reconstruct all 1,225 preferences (assuming transitivity). An example is shown in Figure 2.

```
1 G00-00-1006224 G00-10-3849661 -1
1 G00-10-3849661 G12-90-0628070 -1
...
1 G37-09-0021242 G04-99-1871403 -1
1 G04-99-1871403 G34-06-2520482 -1
1 G34-06-2520482 NA -2
```

```
1 96 5044 -1
1 5044 322933 -1
...
1 848686 136972 -1
1 136972 783579 -1
1 783579 NA -2
```

Figure 2: Example preference judgments. The top judgments use the .gov DOCNOs; the assessor preferred document G00-00-1006224 to G00-10-3849661 and G00-10-3849661 to G12-90-0628070 (and hence preferred G00-00-1006224 to G12-90-0628070). Document G34-06-2520482 was judged nonrelevant. The bottom judgments are for the same documents, but using their corresponding LETOR docids.

4. EVALUATION MEASURES

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.

We have proposed a generalization of precision and recall to preference judgments [3]. First 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.

Note that ppref and rpref at rank $k = \infty$ are proportional to Kendall's τ rank correlation, which is a function of the number of incorrectly ordered pairs, i.e., the number of mis-

classified pairs.

Ties. There are two situations that can be considered "ties": for a given pair of documents, an assessor either judged them to be identical, judged both to be bad, or did not specify anything about them 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.

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 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.

An implementation of these measures is available at `http://ciir.cs.umass.edu/~carteret/preferences.html`.

5. BASELINE RESULTS

Joachims' RankSVM is an adaptation of the support vector machine to learning a ranker [6]. It optimizes a loss function based on preferences to learn a partial ordering of items. Its success has been demonstrated on the LETOR data [7], which consists of binary relevance judgments on the GOV2 documents referenced above.

We trained and tested a RankSVM using the preferences directly, with all $O(n^2)$ preferences inferred from the n that are provided for each query. We joined the preference labels with LETOR features, comprising such information as query term frequency and inverse document frequency, BM25 score, language modeling score, features of the HTML markup, and features of the link graph. Features were normalized to have standard deviation 1.

We tested the standard linear kernel. We compared to two baselines: a random ranking of labeled documents and a SVM classifier trained with binary labels obtained by inference from preferences: documents marked "bad" were considered nonrelevant and all others relevant.

5.1 Training and Testing

The RankSVM was trained using the partitioning described by Liu et al. [7]: five folds, each consisting of 30 training queries, 10 validation queries, and 10 testing queries. The validation set was used to select RankSVM parameter C, the misclassification cost.

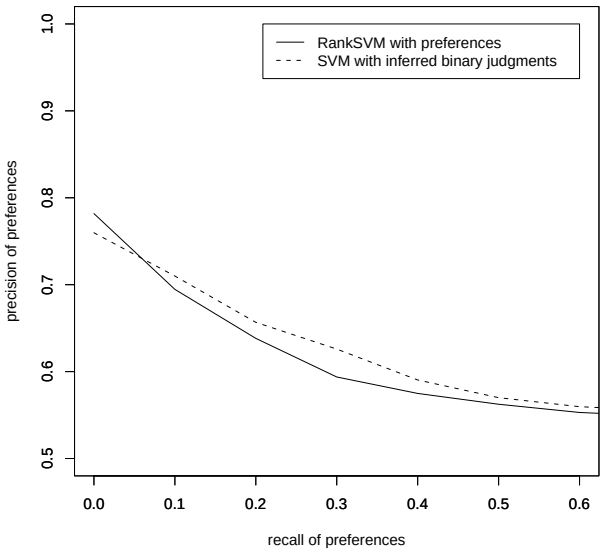

Figure 3: Preference precision-recall curve for the linear-kernel RankSVM trained with preferences and the SVM trained with inferred binary judgments.

5.2 Results

Results for the RankSVM, the binary classifier, and the random ranker are shown in Table 1. Since 50 documents were judged for each query, no more than 50 were ranked by any of the methods; ppref@max and rpref@max therefore refer to the ppref and rpref at the maximum of 50 or the last rank at which any document was ranked. Figure 3 shows the preference precision-recall curve for the RankSVM trained over binary judgments. Note that over half the documents were judged "relevant", i.e. not judged by the assessors to be obviously bad, so ppref cannot be lower than 0.5, while rpref cannot be higher than ppref. Figure 3 also shows the preference precision-recall curve for the binary SVM (note that while the SVM was trained with binary judgments, its results were evaluated with the preference judgments, which are the "truth" in this setting). While the precision of the binary classifier is slightly higher at the top-most ranks, it drops off faster. The two curves hew closely to each other, but the preference curve is clearly superior over the entire ranking.

The difference in performance between preferences and binary labels is small, but preferences provide superior performance for every evaluation measure. Furthermore, preferences are superior at nearly every point in the preference precision-recall curve, and where binary judgments give better performance, it is only a relatively small gain. The random baseline is quite high, likely because so many documents were judged "relevant". Additional research is necessary to understand differences between training over the two different types of data, but these results serve as a useful baseline for future research.

method	ppref@10	ppref@25	ppref@max	rpref@10	rpref@25	rpref@max	APpref
RankSVM-linear (pref)	0.5997	0.5663	0.5549	0.2545	0.4484	0.5549	0.6039
SVM-linear (binary)	0.5835	0.5452	0.5455	0.2368	0.4072	0.5244	0.5987
random	0.4835	0.4959	0.5003	0.1789	0.3663	0.4821	0.5278

Table 1: RankSVM results for preference data along with SVM results for binary classification and results for a random ranker.

6. CITATION

When using this data, please cite:

> B. Carterette, P. N. Bennett, and O. Chapelle. A Test Collection of Preference Judgments. In SIGIR 2008 Workshops: Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments. Edited by Bennett, Carterette, Chapelle, and Joachims.

Acknowledgments

Thanks to Susan Dumais and Microsoft Research, whose generous donation made the preference collection possible, to Thorsten Joachims for helpful feedback, and to our assessors. 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.

7. REFERENCES

[1] Y. Bernstein and J. Zobel. Redundant documents and search effectiveness. In *Proceedings of CIKM*, pages 736–743, 2005.

[2] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of ICML*, pages 89–96, 2005.

[3] B. Carterette and P. N. Bennett. Evaluation measures for preference judgments. In *Proceedings of SIGIR*, 2008. To appear.

[4] 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.

[5] N. Craswell and D. Hawking. Overview of the TREC 2003 Web track. In *Proceedings of TREC*, pages 78–92, 2003.

[6] T. Joachims. Evaluating retrieval performance using clickthrough data. In *Text Mining*, pages 79–96. 2003.

[7] T.-Y. Liu, J. Xu, T. Qin, W. Xiong, and H. Li. LETOR: Benchmark dataset for research on learning to rank for information retrieval. In *Learning to Rank for Information Retrieval workshop in conjunction with SIGIR*, 2007.