Brian Almquist
University of Alberta, Edmonton, Alberta
Padmini Srinivasan
The University of Iowa, Iowa City, Iowa

Refining Ranked Retrieval Results for Legal Discovery Search Through Supervised Rank Aggregation

Abstract: We propose and evaluate a data mining system that uses a set of document features describing each document in the context of partially evaluated ranked results. We find our system to be competitive with existing metasearch ranking strategies for prioritizing the review of evidence for legal relevance.
Résumé : Nous proposons et évaluons un système de fouille de données basé sur une série de descripteurs de documents décrivant chaque document dans un contexte d'évaluation partielle des résultats classés. Nous concluons que notre système est concurrentiel par rapport aux stratégies existantes de classement des métarecherches pour la priorisation de l'examen des preuves en matière de pertinence juridique.

1. Introduction

The goal of this paper is to contribute a machine learning approach for supervised rank aggregation in the context of legal information retrieval. In our system, documents are represented by features extracted from their position in contributing rankings and by the properties of their neighboring documents. Training data derived from relevance feedback is then used to build classifiers for ranking previously unjudged documents.

In litigation, there is a "discovery" stage where a comprehensive set of documents relevant to the case is identified and submitted to the requesting party. The simplest strategy, which is to examine all available documents for selection, presents logistical and legal problems. The job of reviewing the documents must be performed by expensive resources, in this case, lawyers. The cost is exacerbated by the size of enterprise document collections.

In response to these cost and scale challenges, the legal community has developed a system where lawyers construct complex Boolean-style queries to locate all of the pertinent documents. In recent years, the information retrieval community has, through the Text REtrieval Conference (TREC), explored various strategies for ranked retrieval designed to prioritize documents that are predicted to be relevant, a natural alternative to set-based Boolean retrieval.

2. Rank Aggregation

In rank aggregation the goal is to combine ranking decisions made separately by different sources into a single more effective ranking, where the sources represent any of several possible

ranking strategies. In other words, the aim is to use the knowledge of the crowd obtained in the form of rankings to generate a meta-ranking of the items. Supervised ranking allows us to make use of some prior knowledge about the end user's relevance assesments. Our research falls in the 'supervised' rank aggregation framework. In our case, some documents have been judged for relevance and we exploit this information during the merging process. Specifically we train classifiers on the automated rankings of documents combined with associated relevance judgments to create a single merged ranking. The classifiers are built from features derived from the documents position relative to other judged documents in the aggregated rankings.

Most of the methods explored for rank aggregation are unsupervised ones. One of the more popular, and effective of these strategies, the Borda Count, sums the inverse rank for the document across all runs, in effect, the Borda Count strategy (Shaw and Fox, 1993, 243-252). These strategies do not take advantage of the relevance knowledge afforded by preliminary judgments. Instead, these heuristic strategies leverage the "Chorus effect" which assumes that different search algorithms will return common relevant documents and different non-relevant documents (Aslam and Montague, 2001, 276-284). The performance of these methods can be improved using simplified "training" by weighting the input datasets by their scores using evaluation measures.

3. Classifier-Based Rank Aggregation

The input for our problem consists of a set of ranked lists, each containing a minimum of *n* documents, in order of estimated likelihood of relevance. Each of these lists is drawn from the same pool of documents, so it is possible that there is overlap—documents may appear in more than one list. We explore a supervised approach where relevance judgments are available for some of the documents in the ranked set.

Figure 1 illustrates the kind of data we have available. It shows the top 13 documents in each of five runs. Those marked R are relevant, N are non-relevant and ones with a '–' are not yet evaluated. In this example, the 'pool' is made up of documents from five different runs. We use the judged documents from this pool to train our classifier.

Run1	N	R	R	R	R	—	—	N	R	N	N	R	—	...
Run2	N	N	N	N	N	R	N	—	N	N	N	N	—	...
Run3	N	N	N	N	N	—	—	—	N	—	—	N	—	...
Run4	N	N	N	N	N	N	—	—	—	—	N	—	—	...
Run5	N	N	N	N	N	N	—	—	—	—	N	—	—	...

Figure 1. The top 13 ranked documents in five contributing runs. Documents appearing in more than one run share the same shade.

Our system creates a series of features for a document derived from the relevance information available for other documents in the same retrieval run. Additional calculations, including the

various rankings of the document and the frequency of the document's presence in contributing result sets, supplement our features. The features are calculated for each document in the document pool.

Given a set of rankings, we identify the union of the documents from the set. We refer to this combined set of documents as our *document pool*. In the context of legal discovery, it might be assumed that creating a pool from all available contributors would provide access to a greater number of relevant documents, allowing for greater recall.

Using the WEKA framework for machine learning, we build an "SMO" classifier—which uses a sequential minimization optimization algorithm to train a support vector machine—using the evaluated documents (Witten and Frank, 2005). When applied to unevaluated documents, the classifier outputs a non-binary probabilistic prediction of relevance. The unevaluated documents in the pool are then ranked in decreasing likelihood of relevance.

4. Experiments

For our experiments, we work with the datasets generated by the TREC Legal track ad hoc task. We downloaded the results submitted by various participants (systems) during each of three years: 2006 to 2008 (Baron, Lewis, and Oard, 2006, 79-98; Tomlinson, Oard, Baron and Thompson, 2007; Oard, Hedin, Tomlinson and Baron, 2008). To demonstrate potential use outside of the legal discovery domain, we repeated the experiments with data from the TREC Terabyte (Web-based documents and queries) and Genomics (medical literature search) tracks. The documents considered in our experiments included all those that had been submitted by participants, judged or unjudged. We first eliminated submissions for each topic that failed to submit a 1000 documents. We then built our document pools from the remaining runs.

In order to test our approach under the shallow depth scenario, we altered our system to rank evaluated documents. Under a five-fold evaluation, each document in the pool is randomly assigned to one of five test sets. For each test set, the classifier is trained on the remaining judged documents. The classifier is then applied to each document in the test set, and the output ranking score is used to order the test set. Baselines are established by ranking the test set using a Borda Count, and a Borda Count weighted by contributing run precision (BordaFuse) (Ng and Kantor, 1998). The existing relevance judgments are used for assessing the ranked lists compiled by our system and the baseline methods. Each of the ranked test sets is evaluated using Average Precision (P) and AUC measures.

We break our results out into each of the three years of the TREC Legal datasets, and provide summary results over all three years. To judge the effectiveness of our system on other datasets, we also tested it on the non-manual submissions to the 2006 TREC terabyte track.

We constructed the document pools with the top 1000 documents from each of the contributing runs. We assess our system by comparing them against effective heuristic methods for merging result sets. The Borda Count sums up the ranks of a document across all runs, and BordaFuse variant uses the performance measures of contributing runs to give greater weight to more

effective systems when calculating a new ranking score. In each of our test sets, the baseline system ranks the same set of documents as the classifier.

Our system calculates, for each fold, the Average Precision (AP) and the AUC (area under the ROC curve) for each of the ranked results. For each topic, the output of the classifier-based ranker is averaged across all five cross-validation folds. These measures are compared against those of the three baselines over all of the topics under evaluation by the experiment and subjected to two two-sided significance tests (Smucker, Allen and Carterette, 2007, 623-632).

The results of these experiments are presented in Tables 1 and 2. The classifier-based ranker compares favorably to the heuristic methods, with the sole exception where it underperforms the weighted BordaFuse for AUC when tested against the TREC 2007 Legal Track topics. Though the classifier-based ranker generated improvements over the BordaFuse when considering all three Legal datasets—and significant improvements with the Terabyte dataset—it failed to duplicate the result when applied to the medical literature problems.

Table 1. Performance of classifier-based ranker (CBR) against Borda Count (BC), and BordaFuse (BF), measured using average precision. $p<0.0001$ (two-sided Wilcoxon signed rank test) is indicated with †.

Dataset	BC	BF	CBR	Percent Improvement (BF → CBR)
Legal 06	0.2956	0.3739	0.4905	+31.1
Legal 07	0.4961	0.5529	0.5709	+3.3
Legal 08	0.6159	0.6559	0.6576	+0.3
Legal (All)	0.5067	0.5611	0.5906	+5.6
Terabyte 06	0.4978	0.5312	0.6284	+18.1†
Genomics 05	0.3985	0.4565	0.4301	-5.7

Table 2. Performance of classifier-based ranker (CBR) against Borda Count (BC), and BordaFuse (BF), measured using area under the ROC curve (AUC). $p<0.0001$ (two-sided Wilcoxon signed rank test) is indicated with†.

Dataset	BC	BF	CBR	Percent Improvement (BF → CBR)
Legal 06	0.5787	0.7064	0.7476	+5.8
Legal 07	0.7326	0.7868	0.7836	−3.3
Legal 08	0.7232	0.7701	0.7866	+2.1
Legal (All)	0.6989	0.7636	0.7775	+1.8

Terabyte 06	0.7408	0.7721	0.8298	+7.4†
Genomics 05	0.7266	0.7846	0.7686	-2.0

5. Conclusions

We have demonstrated the merits of this technique when applied to smaller data sets. It remains to be seen whether additional improvements can be found when using other machine learning methods. Even more interesting is the improvement found despite the minimally descriptive features. Document profiles in our system reflect nothing about the contents of the documents, nor do they describe the documents with the context of the collection. Additional features that address these gaps may provide information that allows the classifier-based ranker to better distinguish documents at greater depths in the collection.

References

Aslam, J. A. and Montague, M. 2001. Models for metasearch. In SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New Orleans, Louisiana, September 09 - 13, 2001). ACM, New York, NY, USA, 2001, 276-284.

Baron, J. R., Lewis, D. D., and Oard, D. W. 2006. TREC 2006 Legal Track Overview. In Voorhees, E. M. and Buckland, L. P. eds. The Fifteenth Text REtrieval Conference Proceedings (TREC 2006) (Gaithersburg, MD, November 14 - 17, 2006). National Institute of Standards and Technology (NIST), the Defense Advanced Research Projects Agency (DARPA), and the Advanced Research and Development Activity (ARDA), Gaithersburg, MD, 79-98.

Ng, K. and Kantor, P. 1998. An investigation of the preconditions for effective data fusion in information retrieval: a pilot study. In Proceedings of the 61st Annual Meeting of the American Society for Information Science. (Pittsburg, Pennsylvania, 1998).

Oard, D. W., Hedin, B., Tomlinson, S., and Baron, J. R. 2008. Overview of the TREC 2008 Legal Track. In Voorhees, E. M. and Buckland, L. P. eds. The Seventeenth Text REtrieval Conference Proceedings (TREC 2008). (Gaithersburg, MD, November 18 - 21, 2008). National Institute of Standards and Technology (NIST), the Defense Advanced Research Projects Agency (DARPA), and the Advanced Research and Development Activity (ARDA), Gaithersburg, MD.

Shaw, J. A. and Fox, E. A. 1993. Combination of multiple searches. In The Second Text REtrieval Conference. (Gaithersburg, MD, November 1993). 243-252.

Smucker, M. D., Allan, J. and Carterette, B. 2007. A comparison of statistical significance tests for information retrieval evaluation. In CIKM '07: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. (Lisbon, Portugal, November 6-10, 2007). ACM Press, New York, NY, USA, 623-632.

Tomlinson, S., Oard, D. W., Baron, J. R. and Thompson, P. 2007. Overview of the TREC 2007 legal track. In Voorhees, E. M. and Buckland, L. P. eds. The Sixteenth Text REtrieval Conference Proceedings (TREC 2007). (Gaithersburg, MD, November 2007). National Institute of Standards and Technology

(NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA), Gaithersburg, MD.

Witten, I. H. and Frank, E. 2005. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufmann.