Query Expansion for Noisy Legal Documents

Lidan Wang[1,3] and Douglas W. Oard[2,3]
[1]Computer Science Department, [2]College of Information Studies and
[3]Institute for Advanced Computer Studies,
University of Maryland, College Park, MD 20742, USA
lidan@cs.umd.edu,oard@umd.edu

Abstract

The vocabulary of the TREC Legal OCR collection is noisy and huge. Standard techniques for improving retrieval performance such as content-based query expansion are ineffective for such document collection. In our work, we focused on exploiting metadata using blind relevance feedback, iterative improvement from the reference Boolean run, and the effects of using terms from different topic fields for automatic query formulation. This paper describes our methodologies and results.

1 Introduction

The TREC Legal Track is designed to model a real-world challenge known as "e-discovery." In e-discovery, given a "request for production" and its associated complaint, the goal is to return all relevant documents without also returning an unreasonably large number of non-relevant documents. A request for production is a description of documents that are of interest to the requesting party. It is often broadly worded, in order to force the opposing party to produce a comprehensive set of documents that may potentially be useful in court. The two parties might then negotiate a query, which is modeled by a sequence of "Boolean" queries in the Legal Track's Ad Hoc task that lead to an agreed "reference" Boolean query.

Experience from the first two years of the Legal Track suggest that it would be productive to further explore techniques for recall enhancement. In information retrieval, people often find it difficult to anticipate all of the ways in which terms might be used in the collection. Three broad classes of vocabulary enrichment techniques have been used to bridge that gap: interactive relevance feedback [9, 8], fully automatic (i.e., "blind") relevance feedback [4, 6, 5], thesaurus expansion [7, 10]). Noisy OCR results pose challenges for blind relevance feedback and thesaurus expansion; however, the presence of extensive metadata in the Legal Track test collection offers some potential for overcoming these limitations, and hence one goal for the experiments reported in this papers was to begin to explore techniques for leveraging some of that metadata (specifically, the author and recipient fields). The intuition behind our approach is that people who sent or received relevant documents for which OCR text is sufficiently accurate may have also sent or received other relevant documents that cannot be found using OCR text alone. Learning these names using blind relevance feedback techniques and adding them to the query might therefore result in finding additional relevant documents. Preliminary experiments using relevance judgments from 2007 suggested that this technique could be effective.

One of the surprises from the first two years of the TREC Legal Track has been the difficulty of reliably "beating" the reference Boolean run. One possible reason for this is that ranked retrieval techniques weight query terms using techniques such as Inverse Document Frequency (IDF) that are insensitive to the content

in which those terms are used in the query. The reference Boolean run, by contrast, includes query operators such as "AND" and "NOT" that can be selectively applied to more precisely convey the searcher's intent in ways that the system can productively employ. Indeed, "Boolean" is somewhat of a misnomer, since the queries for the so-called reference Boolean run also includes proximity and truncation operators. Precisely replicating the reference Boolean run would be time consuming, however, since underspecified details (e.g., tokenization) differ between systems. We therefore adopted the reference Boolean run as a starting point and sought to make iterative improvements from there. The basic idea was to rank the documents found by the reference Boolean run in a putatively best-first order, to do the same for the documents not found by that run, and then to replace the bottom documents in the Boolean ranking (i.e., those least likely to be relevant) with the top documents in the other ranking (i.e., those most likely to be relevant). Preliminary experiments using relevance judgments from 2007 suggested that this technique could yield small improvements (when averaged across all topics).

As has been observed in previous years, many of the fields in the topic description contain useful query terms, and no one field serves as an optimal selection. In order to provide the strongest possible starting point for our experiments we therefore augmented the original production request (i.e., <RequestText>) with all terms from the Boolean queries proposed by defendant, modified by the plaintiff and agreed as a final query. Preliminary experiments using relevance judgments from 2007 indicated that modest improvements over the use of the <RequestText> field alone could be achieved using this technique.

The remainder of this paper is organized as follows. In the next section, we describe the design of each of our techniques and we explain how they are used together. We then present the results of our preliminary experiments using TREC-2006 and TREC-2007 relevance judgments, including parameter settings that we learned from that data for use in all of our experiments. Section 4 then presents the results for our officially runs using the 2008 relevance judgments. The paper concludes with some remarks on additional experiments and analysis that would be useful, and some implications for future Legal Track design.

2 Method

This section describes our approach of metadata-based query expansion. We also describe how document indexing, query formulation and result sets formulation were done. The process by which specific parameter values were chosen is described in Section 3.

2.1 Indexing

The collection consists of 6,910,192 documents, each of which includes text that was automatically generated from a document image using Optical Character Recognition (OCR) and several fields of metadata (e.g., author, recipient, and organizations). The same document collection have been used in all three years of the track [2, 11]. We indexed the OCR text and each metadata field separately using Indri [1]. No special processing was applied to correct OCR errors; the raw OCR text was indexed. Each term in the OCR text and the title metadata field was stemmed using the Porter stemmer; terms found in the author or recipient field were not stemmed. Indri retrieval model is used for all of our experiments.

2.2 Initial Query Formulation

We extracted terms from the <RequestText>, <FinalQuery>, <ProposalByDefendant> and <RejoinderByPlaintiff> fields to form bag-of-words queries. Functional phrases such as "All documents describing" and "Documents referring to" were automatically removed using a script. If the same word appeared

more than once in the extracted text, it was only counted once. Many Boolean queries contained wildcard characters. For example, "advertis!" was intended to match all words with prefix "advertis". When the <RequestText> field contained a complete word that matched a prefix followed by a wildcard character, we used the full word. Otherwise, we ignored wildcard characters. This would be expected to have little effect on our results because all query terms were subsequently automatically truncated by the Porter stemmer. An exception is that during the metdata-based query expansion phase, the terms found in the author or recipient fields were not stemmed.

2.3 Metadata-Based Query Enrichment

A substantial portion of the document collection consists of memoranda and (printed!) email for which the social network of senders and recipients might be useful. Moreover, some of the same people serve as authors for more formal documents (e.g., reports) that are also contained in the collection. This observation motivated our choice of the social network as a potentially useful source of evidence. The challenge, of course, is query formulation. Although we could, and perhaps should, ask that the names of specific people be included in the negotiated Boolean query, that has not yet been tried in the TREC Legal Track. We therefore adopted a less direct technique of learning potentially useful names—we simply assumed that people who had sent or received documents that we believed were likely to be relevant were more likely than other people to have also sent or received other relevant documents. That formulation motivated our selection of a variant of blind relevance feedback to discover these potentially useful senders and recipients.

The key idea behind blind relevance feedback is that "terms are known by the company that they keep." Specifically, we want to find useful additional terms—in this case the names of senders and recipients—that show up remarkably often in documents on some specific topic. Specifically, we look for some number of tokens that frequently occur in top-ranked documents but are rare in the entire collection. We start by performing an Indri search using the initial query described above to find the top-ranked n documents using the OCR text and the title metadata fields. We then extract terms from the author (<au>) and recipient (<rc>) metadata fields as follows. Let f denote a metadata field ($f \in \{\text{au, rc}\}$), let s denote the string found in f in one document. First, individual terms t are identified in s by tokenizing on punctuation and white space. An IDF weight is then assigned to each term t in s as follows:

$$IDF_{f,t} = \log \frac{N}{DF_{f,t}},$$

where N is the number of documents in the collection, and $DF_{f,t}$ denotes the frequency of term t in the field f in the entire collection. We then sum the IDF weights for each unique term t found in the top m documents to compute an "offer weight" $w(t)$ for each term as follows:

$$w(t) = TF_{f,t} \cdot IDF_{f,t},$$

where $TF_{f,t}$ denotes the number of occurrences of term t in field f in the top m documents. Note that this differs from the usual formulation of TF*IDF weights because the IDF is computed over individual strings, while the TF is computed over a set of strings. Finally, the terms are sorted in decreasing order of offer weight, and the n most highly ranked terms are selected. m and n are free variables that were selected using the process described in Section 3.

A new Indri query is then formed as a weighted combination of the original query and metadata search terms:

$$\#weight(\lambda\, Q_i \;\; (1-\lambda)\, Q_m),$$

where Q_i is the initial query described above and λ is a free parameter selected using the process described in Section 3. The Indri query operator

$$Q_m = \#wsyn(w(t_1)\ t_1\ \ldots\ w(t_n)\ t_n)$$

treats the terms t_i as synonyms, but allows weights $w(t_i)$ to be assigned to each term. A weight on the term denotes its relative importance in the synonym list. For Q_m, Indri operator #od1 is used to perform exact match of t_i in the metadata field f; in other words, only documents whose meta data matches t_{fi} are returned. Stemming is not performed for metadata terms.

2.4 Reference boolean run

Results for the reference Boolean run were provided for each topic, ordered alphabetically by documentID. These documents match the final negotiated Boolean query. The B value, the number of documents matching the final negotiated Boolean query, was also provided (B varies by topic from a low of 170 to a high of 99,374). Since these results are intended to model present practice, one question of interest to us is how to form a set of size B that achieves better results (as measured by F_1 at B) than the reference Boolean run. Tuning Boolean queries to yield result sets that are neither too large nor too small is well known to be difficult [3], but the choice of B as the reference value for this comparison obviated that disadvantage of Boolean queries in the context of the TREC 2007 Legal Track's evaluation framework. As a result, as of 2007, participants in the Ad Hoc task of the Legal Track have yet to reliably "beat Boolean." One possible cause for this is that no research team automatically formulated queries in as nuanced a way as the final negotiated Boolean queries. One potentially productive line of research would be to learn to automatically construct richer queries, perhaps using some combination of blind relevance feedback and rule induction. That's a fairly major undertaking, however, so we have chosen a more modest way of initially exploring the potential of this approach.

Our strategy is to replace p documents in the reference Boolean run results by an equal number of documents chosen from the rest of the collection in a way that we would expect to make them more likely to be relevant. Algorithm 1 presents the details.

Algorithm 1: Forming new ranked list of size B

Input: The reference boolean run R_b; a ranked list of the entire collection R_c
Output: A new ranked list R_{new} of size B
Rank the reference boolean run R_b using R_c. Denote the result by R_{br};
Remove the bottom p documents, denoted by B_p, from R_{br};
Select a set of top p documents from R_c, such that the set contains no overlapping documents with B_p;
Combine these top p documents with R_{br} ;
Optionally, sort the resulting documents by R_c to from a new ranked list R_{new}

The inputs to Algorithm 1 are the reference boolean run R_b and a ranked list for the entire collection R_c created as described above. For efficiency reasons, we actually work with only the part of R_c in which the documents contain at least one term from query Q (because other documents would simply be ordered arbitrarily). We tried two variants of this algorithm, one in which the same value of p was used for every topic and a second in which p was allowed to vary by topic by setting it to a constant fraction of the B value for that topic. These were free variables that were selected using the process described in Section 3.

3 Preliminary Experiments

We conducted some preliminary experiments using the full set of 43 topics from the Ad Hoc task of the TREC-2007 Legal track for which relevance judgments were available.[1] We elected not to use the TREC 2006 legal track relevance judgments because the sampling strategy that year emphasized highly-ranked documents (and because only very limited additional relevance judgments for the 2006 topics were available from the 2007 relevance feedback task).

We initially used these 2007 "training" topics to explore some key alternatives in what was initially quite a broad design space. Preliminary experiments found that retaining wildcard characters could adversely affect retrieval effectiveness when the prefix was common. Retaining the full version of a query term when possible proved to be helpful, perhaps because the full word in <RequestText> is often the most common version of the wildcarded term found in <FinalQuery>, <Proposal ByDefendant> and <RejoinderByPlaintiff>. These initial experiments were not systematic, but they were sufficient to motivate the design of our wildcard character handling.

Figure 1 shows some locally scored results from a few of our initial exploratory runs using the 2007 topics and relevance judgments. Run UMD-CR-C50 was produced by swapping out some low-ranked documents from the 2007 reference Boolean run with highly ranked documents from a ranked retrieval run in which terms from all three Boolean fields were used to form the query; in this case the number of swapped documents was 50. Runs UMD-AURC-P1 and UMD-AU-P5 explored different metadata expansion and swapping strategies: run UMD-AURC-P1 used blind relevance feedback on metadata to add author and recipient names to the ranked query that was used as a basis for swapping (with $p = 0.01 * B$); run UMD-AU-P5 did the same with just author names (in this case with $p = 0.05 * B$). The results of these and other early experiments led us to conclude that metadata-based expansion and our swap-based merging straregy for building up from the reference Boolean run both seemed to be viable strategies. Note that we held out no data for development test, these represent a train-on-test condition that indicate potential, but not actual, retrieval effectiveness on unseen data.

With this as motivation we conducted a more systematic evaluation of parameters for our metadata-based expansion algorithm, trying a range of reasonable values for m and n. Ultimately, we settled on the top 10 terms from the sender and recipient fields in the top 20 documents (i.e, $m = 20$ and $n = 10$). We also found that $\lambda = 0.8$ was a reasonable choice. In our exploration of this parameter space we varied one parameter at a time, selecting the optimal value for each before proceeding to the next parameter. We then conducted a sensitivity analysis around the final values to select the optimal combination.

Our most systematic exploration of the parameter space was for finding the values of p that optimized the value of F_1 at B. In the TREC 2008 Legal Track Ad Hoc task participants were allowed to select any value for K, the threshold at which their system would be evaluated. We selected $K = B$ for each topic because comparability with the reference Boolean query was essential to our goal. We used a bag-of-words query formed from <RequestText>, <FinalQuery>, <ProposalByDefendant> and <RejoinderByPlaintiff> fields with no metadata expansion for the parameter tuning process (although the same parameters were also later used with expanded queries). The algorithm performs a hill-climbing search on a local optimal value of p, the number of lowest-ranked documents in the reference Boolean run for which highly-ranked top-p documents from the remainder of the collection were substituted. We initially tried fixed values of P from $\{10, 20, 30, \ldots, B\}$ in that order, stopping as soon as $F_1@B$ first declined. This

[1] TREC 2007 results were reported on 42 topics, so the results reported here are not strictly comparable with those results. Relevance judgments for one additional TREC 2007 Legal Track Ad Hoc task topic became available after official scores were computed.

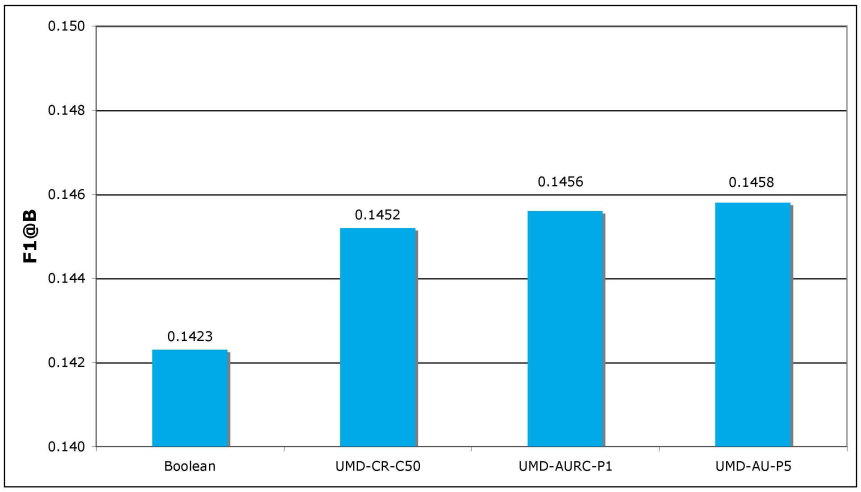

Figure 1: Preliminary 2007 experiments, estimated $F_1@B$.

resulted in an optimal value $p = 40$. This value turns out to be smaller than the minimum B for any 2008 topic, so no special handling is needed for "small" topics. We then repeated the hill climbing process with p be some fraction of B, where $p \in \{0.01 * B, 0.02 * B, \ldots, 1.00 * B\}$ with the same stopping criterion. The optimal fraction turned out to be $0.03 * B$. We used these same parameters for our TREC 2008 Legal Track Ad Hoc task experiments. Both values are relatively small, so the upside potential of this approach is somewhat limited.

4 2008 Experiments

A total of 45 topics were provided as an XML file. We submitted five runs, each containing about $100,000$ rank-ordered documents for each topic. Each run was scored twice, once using all 26 topics for which relevance judgments were available and using only the 24 documents for which highly relevant documents had been identified. In this paper we report only estimated $F_1@B$ because we set $K = B$ and $K_h = B$ (K_h is K value for highly-relevant) for all runs.

4.1 Submitted Runs

1. UMD-STD[2] is our standard condition run, using all terms in the "RequestText" field to form a bag-of-word query using Indri's "#combine" operator. Functional phrases such as "All documents describing" and "Documents referring to" are ignored.

2. UMD-CR-C40 is a ranked list obtained by 1) producing an initial ranked list (R_c) with a bag-of-word query Q from keywords in "RequestText", "FinalQuery", "ProposalByDefendant" and "RejoinderByPlaintiff"; 2) ranking the Boolean run by R_c; 3) replacing the bottom $p = 40$ documents in the ranked Boolean run with the top $p = 40$ documents from R_c (after redundancy elimination w.r.t. the Boolean run); 4) sorting the result by R_c.

[2]We have added dashes that were not present in the submitted run names for improved readability

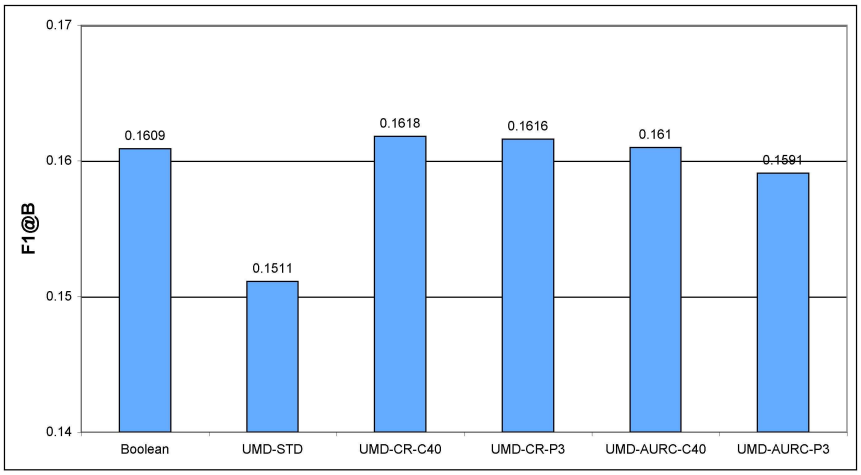

Figure 2: Official results, estimated $F_1@B$, all relevant documents.

3. UMD-CR-P3 is the same with UMD-CR-C40 except p was chosen to be 3% of the B value for each topic.

4. UMD-AURC-C40 is a ranked list obtained by 1) producing an initial ranked list with a bag-of-words query from all terms in "RequestText," "FinalQuery," "ProposalByDefendant" and "RejoinderByPlaintiff;" 2) extracting the top $n = 10$ terms from the author and recipient metadata fields of the top $n = 20$ documents in the initial ranked list (Section 2.3); 3) forming a new query combining the initial query and metadata terms (Section 2.3) denoting the resulting ranked list R_c; 4) ranking the reference Boolean run by R_c, and substituting bottom $p = 40$ documents by the top $p = 40$ documents from R_c (after redundancy elimination w.r.t. Boolean run); 5) sorting the result by R_c.

5. UMD-AURC-P3 is the same as UMD-AURC-C40 except p was chosen to be 3% of the B value for each topic.

4.2 Results

Figure 2 reports our results for the principal evaluation measure F1@B. For comparison, we also show the same result for **Boolean**, the reference Boolean run provided by the organizers. Selectively merging with the Boolean run improves retrieval effectiveness, when the "least likely" Boolean retrieved documents are replaced by the "most likely" documents from the Boolean complement. However, somewhat surprisingly, the runs that employ metadata-based expansion before merging with Boolean (UMD-AURC-P3 and UMD-AURC-C40) do not perform as well as the runs that simply merge with Boolean (UMD-CR-P3 and UMD-CR-C40). One possible reason is that metadata-based expansion is not content-based; it scores documents in part based on how likely their authors or recipients are "relevant". This can bring in some non-relevant documents that would adversely affect retrieval effectiveness.

Figure 3 presents the same conditions using only "highly relevant" judgments. In this case, metadata-based expansion (UMD-AURC-C40) apparently improved retrieval effectiveness, although it is not yet statistically significanct. The results do seem to depend on how the p value is chosen: UMD-AURC-C40 (in which 40 documents from Boolean reference run are swapped) appears to consistently result in slightly bet-

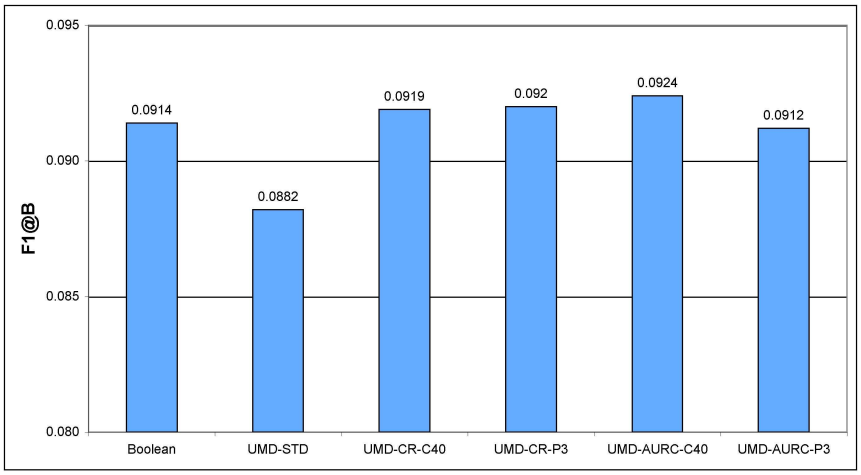

Figure 3: Official results, estimated $F_1@B$, only highly relevant documents.

ter retrieval effectiveness than UMD-AURC-P3 (in which 3% of the Boolean reference run was swapped).

Averages across all topics can mask substantial variation, so Figure 4 compares estimated $F_1@K$ between UMD-CR-C40 and the median of the 41 submitted runs from all participating teams.[3] A total of 7 of the 26 topics show improvements over the median that exceed Sparck-Jones' suggested threshold of 0.1 (absolute) for an improvement that interactive users would find large enough to be "meaningful," while an adverse effect of that magnitude is present for only one of those 26 topics.

Quite a lot of analysis remains to be done, but some implications for future work are already evident. We briefly summarize some of our thoughts on that in those issues in next section.

5 Conclusions and Future Work

The experiment results provide some evidence that we can productively leverage the presence of metadata in the TREC Legal Track test collection. It also seems that we can do at least as well as the reference Boolean run through the simple expedient of building new runs by perturbing the Boolean results rather than starting from scratch without the benefit of the Boolean query. Of course, there are a number of parameters that we could optimize. For example, it may be useful to parameterize our blind relevance feedback process differently, or to perform more than one feedback iteration. But our work is more interesting for the possibilities it opens up for more broadly exploring the design space. For example, we might try different ways of making use of the feedback, perhaps using some form of co-training rather than simply searching a unified index containing terms from multiple fields as we do now. Additional metadata fields might also prove to be useful. For example, the Bates number might provide evidence for the degree of physical proximity of documents in their original storage location, which might be connected to some degree of related meaning. With more than 100 topics now available, the TREC Legal Track test collection is a useful resource for exploring these questions, and we look forward to doing so in our future work.

[3] We show $F_1@K$ rather than $F_1@B$ because K varied for other participating teams.

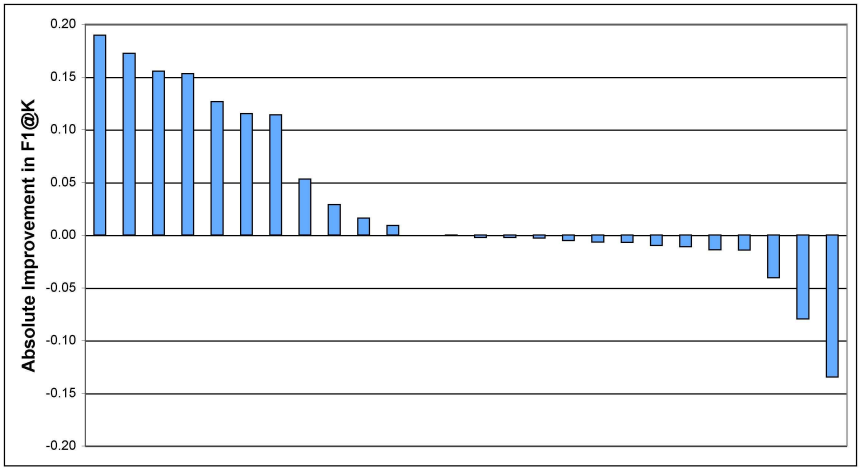

Figure 4: Difference between UMD-CR-C40 and median, estimated $F_1@K$, all relevant documents. Topics are in sorted order.

References

[1] The Lemur Toolkit for Language Modeling and Information Retrieval. http://www.lemurproject.org.

[2] J. Baron, D. Lewis, and D. Oard. TREC 2006 legal track overview. In *Proceedings of the 15th Text Retrieval Conference*, 2006.

[3] D.C. Blair. *Language and representation in information retrieval*. Elsevier Science Publishers, 1990.

[4] W. Croft and D. Harper. Using probabilistic models of information retrieval without relevance information. In *Journal of Documentation*, 1979.

[5] X. Jinxi and W. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.

[6] M. Mitra, A. Singha, and C. Buckley. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1998.

[7] Y. Qiu and H. Frei. Concept based query expansion. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1993.

[8] J. Rocchio. Relevance feedback in information retrieval. In *The SMART retrieval system experiments in automatic document processing*, 1971.

[9] G. Salton(ed). The SMART retrieval system experiments in automatic document processing. 1971.

[10] H. Schutze and J. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. In *Information Processing and Management, v33 n3 p307-18*, 1997.

[11] S. Tomlinson, D. Oard, J. Baron, and P. Thompson. Overview of the TREC 2007 legal track. In *Proceedings of the 16th Text Retrieval Conference*, 2007.