MultiText Legal Experiments at TREC 2007

Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack, Thomas R. Lynam
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

Abstract

For the legal track we used the Wumpus search engine and investigated several methods that have proven successful in other domains, including cover density ranking and Okapi BM25 ranking. In addition to the traditional bag-of-words model we used boolean terms and character 4-grams. Pseudo-relevance feedback was effected using logistic regression on character 4-grams. Some runs specifically excluded documents returned by the boolean query so as to increase the number of such documents in the pool. While our runs were all marked as *manual,* this was only because the process was not fully automated and several tuning parameters were set after we viewed the data; no data-specific tuning was performed in configuring the system for our runs. Our best performing runs used a combination of all of the above-mentioned techniques.

1 Introduction

2 Legal Track

For the legal track, we investigated several primitive approaches that have worked well in other domains, and combinations of approaches (combination itself being an approach that has worked well elsewhere[LBCC04, LC06]).

2.1 Legal Retrieval Methods

The following is a brief description and rationale for each run.

wat1fuse

A fusion of runs wat2nobool, wat3desc, wat4feed, wat6qap, wat7bool, wat8gram. The fusion of runs was done using the CombMNZ[SF94, BKFS95] combination method. CombMNZ is a common method of combining multiple retrieval schemes. It combines and re-scores all documents for each query from a set of retrieval schemes. The fused document score is the sum of the scores for the given document of the schemes multiply by the number of schemes the document appeared.

wat2nobool

One of the goals of the legal track is to compare boolean vs non-boolean information retrieval. To better understand this difference the wat1fusion run excludes all documents matched by the boolean query run. This is done separately for each query due to the possibility of a document being returned for more than one query. The intent of this approach was to explicitly give high rank to relevant documents that were not identified at all by the boolean query.

wat3desc

This run used only the RequestedText field of the topic. The legal track corpus is made up of documents scanned from images on which optical character recognition OCR was performed. This has cause the documents to be what a photographer would describe as "noisy". There are many incorrectly recognized letters and words. N-gram retrieval was use to lessen this problem of "noisy" documents. We know from previous experience that character 4-grams are competitive with bags of words for our IR techniques, and had reason to believe that they might be more robust to the errors introduced by OCR. Furthermore, we know that character 4-grams provide much better performance for spam filtering. Every document in the corpus indexed as 4-gram. The wat3desc queries are the RequestedText field converted to 4-grams and are treated as a bag of words. For example, the phrase

"smoke it"

was considered to have terms

"smok" "moke" "oke " "ke i" "k it"

Source	Run	4-gram Query
Requested Text Field	wat3desc	"Zmem" , "memb" , "embe" , "mber" , "bers" , "ersh" , "rshi" , "ship" , "hipZ" , "ipZa" , "pZan" , "Zand" , "ando" , "ndor" , "dorZ" , "orZp" , "rZpa" , "Zpar" , "part" , "arti" , "rtic" , "tici" , "icip" ,....,"estZ"
Final Query Field	wat9boolgram	"tradeZorganiz" , "Ztra" , "trad" , "rade" , "adeZ" , "deZo" , "eZor" , "Zorg" , "orga" , "rgan" , "gani" , "aniz" , "nizZ" , "tradeZassoc" , "Ztra" , "trad" , "rade" , "adeZ" , "deZa" , "eZas" , "Zass" ,..., "nceZ"
Feedback Method	wat4feed	"dxxe" ,"xckk" ,"irzp" ,"ticu" ,"ztel" ,"geme" ,"oves" ,"sxuu" ,"alty" ,"asua" ,"szys" ,"pzzn" ,"zxkd" ,"tzyf" ,"zflo" ,"mzco" ,"elxw" ,"lxwh" ,"ppar" ,"oned" ,"appa" ,"ofit" ,"xuzc" ,"gnsz" ,"szyf" ,"paym" ,"yxxu",...,"szsa

Table 1: 4-gram queries

The 4-gram bag of word queries are issued against the corpus using the okapi BM25[RWJ+95] document ranking.

wat4feed

We implemented a new pseudo feedback method for the legal track. For feedback, we took the top-scoring 20 documents from each run and assumed them to be relevant. We took the lowest-scoring 20 documents (at the depth returned by the boolean query: the value of the FinalB field) and assumed them to be non-relevant. The documents were parsed into overlapping character 4-grams and logistic regression was used to determine the 4-grams most associated with relevant documents. These 4-grams were used as the query in a BM25 run.

wat5nofeed

A fusion of runs wat2nobool, wat3desc, wat6qap, wat7bool, wat8gram. The fusion of runs was done using the CombMNZ combination method. This run is almost the same as the wat1fusion but wat4feed was not included for combination.

wat6qap

A "relaxed" version of the boolean run. This run was ranked using cover density ranking. The approach that MutliText has used with success over the years for IR and QA[CCKL00, CCL01] that searches for short intervals of text containing important terms from the query. We the run relaxed because the highest-level disjuncts (or conjuncts) from the boolean queries are removed. For example, the query

> ("smoke" or "cigarette") and ("girl" or "boy")

was considered to have two terms:

> ("smoke" or "cigarette") ("girl" or "boy")

wat7bool

This is our "boolean" run. The run is ranked with by proximity-ranked[CC96] boolean queries. The queries were recast in GCL, the MultiText query language, and ranked in inverse proportion to the length of interval of text satisfying the query. This approach should give roughly the same documents as supplied, but ranked so as to improve early relevance.

wat8gram

A fusion of runs wat3desc, wat4feed, wat9boolgram The fusion of runs was done using the CombMNZ combination method.

wat9boolgram

This run was not submitted but was one the of runs combined to create wat8gram. Each boolean query was converted to a bag of words. The bag of words were then converted to 4-grams. The 4-gram bag of word queries are issued against the corpus using the okapi BM25 document ranking. This run is very similar to wat3desc which uses 4-grams from the RequestedText field where this runs user 4-grams from the FinalQuery field.

N-Gram Query Examples

Table 1 shows the 4 grams produced by the different runs for query 93. The FinalQuery field for topic 93 is:

> ("trade organiz!" OR "trade assoc!") AND (member! OR participat!) AND (property OR casualty) AND insurance

The RequestText field is:

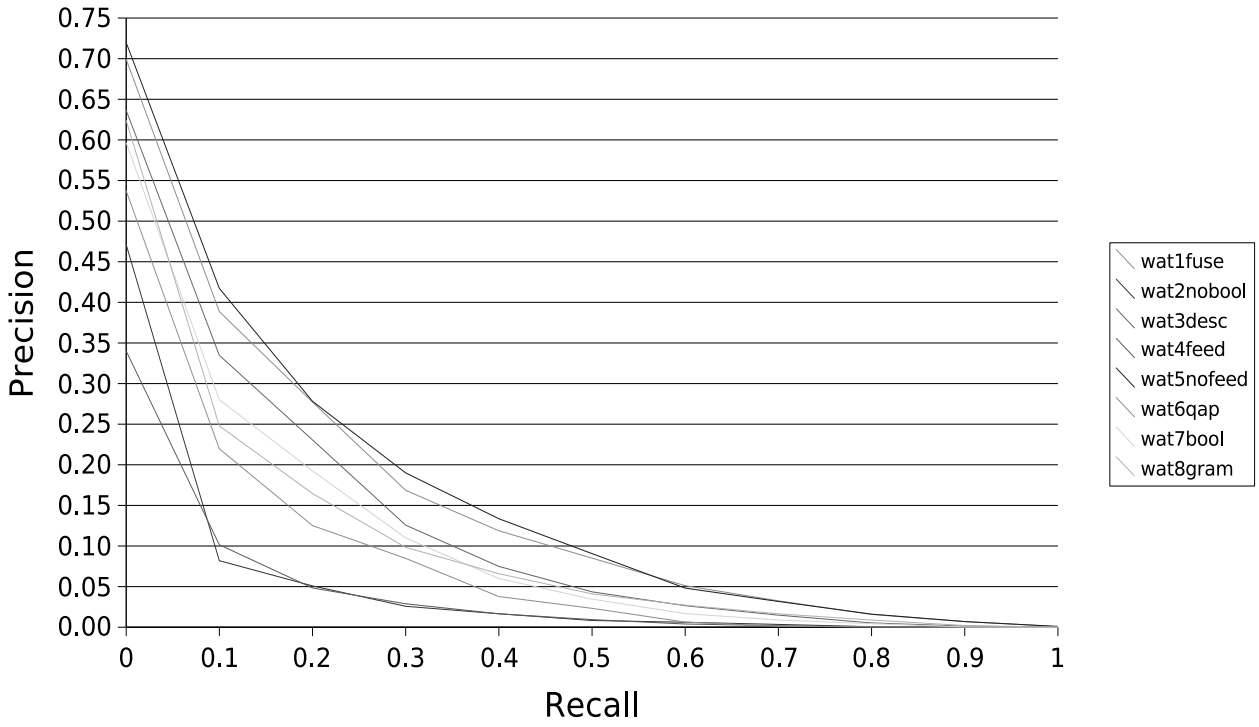

Figure 1: Legal Runs - Precision Recall

> Submit all documents that relate to or discuss membership and/or participation in trade associations or organization related to the propert y and casualty insurance industry. Documents, reports, or publications created or published by such trade associations or organizations are specificall y excluded from this request.

The "Z" in the query defines white space. The Feedback query terms are very interesting as only the most important part of the word is used for feedback. This might appear more true then it really is because the 4-gram are rank by importance and therefore the full word might appear but not be in order.

run	map	bpref	# relevant
wat1fuse	0.1415	0.3834	3666
wat2nobool	0.0383	0.1696	1635
wat3desc	0.1079	0.3273	3365
wat4feed	0.0364	0.2145	1542
wat5nofeed	0.1478	0.3866	3648
wat6qap	0.0692	0.2720	2041
wat7bool	0.0912	0.3046	2133
wat8gram	0.0878	0.3242	3382

Table 2: Legal Track Run Results

2.2 Legal Track Results

Table 2 shows our mean average precision(map), bpref scores and the number of relevant documents returned for our legal track runs. The fusion runs wat1fuse, wat1nofeed outperforms the other runs. Not including the feedback in the fusion has very little affect. Using the RequestText field as the query, wat3desc outperforms the boolean approach wat7bool with respect to map and bpref. It is surprising that using the RequestText retrieves over 50% more relevant documents than the boolean . It is also unexpected that the boolean retrieval retrieves only 2133 of the 4344 relevant documents. This show that methods other than boolean need to be used to achieve full recal. Excluding the documents returned by the boolean run from the fusion does poorly as one might expect but it does find 1635 relevant documents. The Feedback wat4feed and relaxed boolean wat6qap don't seem to work as well as the other methods but they make an important contribution to the fusion run. Figure 1 shows the precision recall graphs for the legal track runs.

References

[BKFS95] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Inf. Process. Manage.*, 31(3):431–448, 1995.

[CC96] C.L.A. Clarke and G.V. Cormack. Interactive substring retrieval (MultiText Experiments for TREC-5). In *5th Text REtrieval Conference*, Gaithersburg, MD, 1996.

[CCKL00] C. L. A. Clarke, G. V. Cormack, D. I. E. Kisman, and T. R. Lynam. Question answering by passage selection. In *9th Text REtrieval Conference*, Gaithersburg, MD, 2000.

[CCL01] Charles L. A. Clarke, Gordon V. Cormack, and Thomas R. Lynam. Exploiting redundancy in question answering. In *SIGIR Conference 2001*, New Oreans, Louisiana, 2001.

[LBCC04] Thomas R. Lynam, Chris Buckley, Charles L. A. Clarke, and Gordon V. Cormack. A multi-system analysis of document and term selection for blind feedback. In *CIKM '04: Thirteenth ACM conference on Information and knowledge management*, pages 261–269, 2004.

[LC06] Thomas R. Lynam and Gordon V. Cormack. On-line spam filter fusion. In *29th ACM SIGIR Conference on Research and Development on Information Retrieval*, Seattle, 2006.

[RWJ+95] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Third Text REtrieval Conference*, Gaithersburg, MD, 1995.

[SF94] Joseph A. Shaw and Edward A. Fox. Combination of multiple searches. In *Text REtrieval Conference*, pages 0–, 1994.