TREC-3 Ad Hoc Retrieval and Routing Experiments using the WIN System

Paul Thompson
Howard Turtle
Bokyung Yang
James Flood
West Publishing Company
Eagan, MN 55123

1 Introduction

The WIN retrieval engine is West's implementation of the inference network retrieval model [Tur90]. The inference net model ranks documents based on the combination of different evidence, e.g., text representations, such as words, phrases, or paragraphs, in a consistent probabilistic framework [TC91]. WIN is based on the same retrieval model as the INQUERY system that has been used in previous TREC competitions [BCC93, Cro93, CCB94]. The two retrieval engines have common roots but have evolved separately – WIN has focused on the retrieval of legal materials from large (>50 gigabyte) collections in a commercial online environment that supports both Boolean and natural language retrieval [Tur94]. For TREC-3 we decided to run an essentially unmodified version of WIN to see how well a state-of-the-art commercial system compares to state-of-the-art research systems.

Some modifications to WIN were required to handle the TREC topics, which bear little resemblance to queries entered by online searchers. In general we used the same query formulation techniques used in the production WIN system with a preprocessor to select text from the topic in order to formulate a query.

WIN was also used for routing experiments. Production versions of WIN do not provide routing or relevance feedback so we were less constrained by existing practice. However, we decided to limit ourselves to routing techniques that generated normal WIN queries. These routing queries could then be run using the standard search engine.

In what follows, we will describe the configuration used for the experiments (Section 2) and the experiments that were conducted (Sections 3 and 4).

2 System Description

The TREC-3 text collection was indexed in essentially the same way for both the ad hoc and routing experiments. Some fields within each document were not indexed; these fields include: CO, DESCRIPT, DOC, DOCID, DOCNO, FILEID, FIRST, G, GV, IN, MS, NS, RE, SECOND. These fields were excluded either because they contained manually indexed terms (which cannot be used under the TREC rules) or because the were considered to be noise. A bounded paragraph algorithm [Cal94] was used to identify paragraph boundaries. Natural paragraphs were used subject to the constraint that a paragraph had to contain a minimum of 50 and a maximum of 200 words.

All of the text not contained in these fields was indexed except for Federal Register documents. Federal Register documents tend to be very long and to contain a great deal of noise. In an attempt to identify text that was a reasonable description of document content we indexed only the "SUMMARY" paragraph if the document contained one, otherise we indexed only the first kilobyte of text in a Federal Register document. Since no Federal Register documents were contained in the routing test collection all text except for the excluded fields was indexed.

3 Ad hoc experiments

The ad hoc experiments used queries that were automatically created from the topic text. The retrieval algorithm used combined document and top paragraph scoring. It was observed that the a priori likelihood of relevance for a document varied from collection to collection. Furthermore each collection's likelihood of relevance given the value of domain field, varied, as well. Some experiments were done in an attempt to exploit these observations.

3.1 Query Processing

A WIN query consists of concepts extracted from natural language text. Rather than extracting concepts from the full topic only the Title field, the Description field, and the first sentence of the Narrative field were used. Each occurrence of a term, or concept, was counted and weighted by field. A term appearing in Title was given a weight of 4, while terms appearing in Description and Narrative were given weights of 2 and 1, respectively.

Normal WIN query processing eliminates introductory clauses and recognizes phrases and other important concepts for special handling. Many of the concepts ordinarily recognized by WIN are specific to the legal domain (e.g., legal citations, West Key Numbers) and were not used in these experiments.

WIN ordinarily makes use of a dictionary of introductory clauses (e.g., "Find cases about ...", "I'm interested in statutes that ...") that don't bear directly on the content of the query. The set of introductory clauses was expanded to include 170 new clauses (e.g., "A relevant document must describe ...") identified in the Description and Narrative fields in the training set. In addition the string "e.g" was added to the set of query stopwords.

WIN also expands some query terms automatically. For example "usa", "us", "u.s", and "united states" were all replaced with the synonym class #syn(ac:us #+1(united states)) that will conflate common variants. Twenty nine new synonym classes were added for automatic expansion.

WIN ordinarily uses a legal dictionary to find phrases in queries. For TREC-3 the dictionary was expanded with phrases extracted from the machine-readable Collins Dictionary. The normal WIN dictionary incorporates information about how a phrase identified in a query is to be matched in document text. For example, query stopwords are generally not

		AP	DOE	FR	WSJ	Ziff
Topics	% of all documents	22.2	30.5	6.2	23.3	17.8
1- 50	% of relevant docments	20.1	0.8	3.1	33.9	42.2
		0.91	0.03	0.50	1.45	2.37
51-100	% of relevant docments	37.2	7.5	3.1	38.0	14.2
		1.68	0.25	0.50	1.63	0.80
101-150	% of relevant docments	41.3	5.8	3.5	39.2	10.2
		1.86	0.19	0.56	1.68	0.57
Total	% of relevant docments	31.4	4.4	3.2	36.7	24.3
		1.41	0.14	0.52	1.58	1.37

Table 1: Collection bias in relevance judgments

considered to be significant, but for some phrases (e.g., "at will") they are used. None of the phrases extracted from the Collins Dictionary used any special recognition features.

3.2 Experiments with different likelihoods of relevance based on collection

In the TREC training set, the likelihood that a document will be judged relevant depends heavily on the collection in which it is found. Table 1 shows the distribution of documents among the five TREC collections and the distribution of relevant documents among the five collections. The AP collection, for example, contains 22.2% of all documents in the TREC collection, but it contains 31.4% of all relevant documents in the TREC collection. Table 1 shows that, for all topics, documents in two of the collections (DOE and Federal Register) are substantially less likely to be judged relevant as would be expected if there were no collection bias whereas documents from the Wall Street Journal, AP, and Ziff collections are much more likely to be judged relevant than expected.

Table 1 also shows that the distribution of relevant documents among the collections varies for different topic sets. For example, Ziff documents are much more likely to be judged relevant than expected for Topics 1-50, but less likely than expected for the remaining two topic sets.

A set of experiments was conducted in which the prior probability of relevance was set to the observed probability of relevance for each of the TREC collections rather than a default probability that was the same for all documents. This essentially biased retrieval in favor of AP, Wall Street Journal, and Ziff documents and against DOE and Federal Register documents. These experiments showed a slight drop in retrieval effectiveness because the priors computed for the entire topic set rarely match the priors computed for individual topics.

A second set of experiments was conducted to determine whether it would be possible to predict the appropriate collection biases based on the characteristics of individual topics. Approaches were tried using both the language contained in the topics and the domain field contained in many of the training topics (note, however, that the test topics do not contain domain fields). None of these approaches significantly improved performance, but the amount of effort devoted to these experiments was limited. We regard this as a promising line of future research.

4 Routing Experiments

The routing experiments used the same techniques as the ad hoc experiments to index the text collection, except that idf values were derived differently. Since the test collection was to be used as a simulation of routing, the TREC guidelines do not allow use of any collection wide statistics, such as idf. Accordingly, the idf values from the CD-1 training set were used instead. Query processing, or profile creation, however, was done in a substantially different manner. No attempt was made to use the observed likelihoods of relevance of different collections, as was done with the ad hoc queries. The routing experiments were based on query expansion. No term reweighting was done.

4.1 Profile Processing

As was the case with the ad hoc queries, only certain portions of the topic text were used for profile creation. These were the Title and Concepts fields. As before, each occurrence of a term, or concept, was counted and also weighted by field. A term appearing in the Title field was given a weight of 2, while a term appearing in the Concepts field received a weight of 1. Any term not appearing in any of the relevant training documents, was removed. Consideration was given to increasing the weight of any term appearing in relevant, but not in irrelevant documents. This had no effect. Only one term met this condition. As a form of normalization the maximum weight that a term could attain was set. This weight was variously set at 5, 6, 7, and 8. This maximum included the contribution provided by the term expansion process, which was always 1 for a selected term, or 0 for a non-selected term (see below). A term might appear multiple times in the Concepts field, thus resulting in a unnormalized term weight that exceeded the maximum.

None of the usual WIN query formulation aids used with ad hoc queries (elimination of introductory clauses, use of replacement strings, and use of a phrase dictionary) were used for profiles. The Title and Concepts fields did not contain any introductory phrases, or clauses. Simple acronyms, such as "RISC" or "MIPS" that were found in the text of relevant training documents during query expansion were identified as acronyms in the profiles, so that they would be treated as instances of the same concept in subsequent processing.

4.2 Query expansion

The focus of the routing experiments was on query expansion. Three different approaches to query expansion were used: "best entire document", "best rntidf top 200 paragraphs", and "best rntidf top paragraph". Ultimately these three approaches were combined in the "best overall" approach. The approaches were themselves based on three methods of term selection: "rdfidf", "rntidf", and "rtfidf" [HC93]. The rdfidf score of a term was calculated by multiplying its idf value by the number of relevant training documents in which the term occurred. The rntidf score for a term was calculated by multiplying its idf value by the summation over all relevant training documents of the ratio of the term's frequency to the frequency of the maximally-occurring term for that particular document. The rtfidf score was simply the multiplication of the term's idf value by the number of occurrences of the term within relevant training documents. The rtfidf score did not perform as well as the other two term selection methods, and so was not used as part of the final runs. For each term selection method, terms selected were those with the highest scores. Terms were only selected from relevant documents. The term scores were only used for term selection, not for term reweighting. A selected term was given a weight of 1 in the expanded query. If the term duplicated a term already represented in the topic profile, then 1 was added to that term's current score.

A baseline run was made using the profile creation process described above, but with no term expansion. Each of the three expansion approaches was then run with terms added by one or both of the remaining term selection methods, i.e., excluding rtfidf. For each approach runs were made with from 5 to 50 terms added, in increments of 5. The "best entire document" approach used both the rdfidf and the rnfidf methods of term selection, with terms selected from any part of the document. The term selection that performed better on the training set was selected on a topic per topic basis. With the "best rntidf top 200 paragraphs", and the "best rntidf top paragraph" approaches, as their names imply, only the rntidf method was used, as it provided better results. For the "best rntidf top 200 paragraphs" approach searches were done for each topic using the baseline, i.e., unexpanded, profile as a query against the training collection. For each topic the top 200 scoring paragraphs from relevant documents were identified, using the WIN paragraph scoring method. Terms were then selected from these paragraphs using the rntidf method, rather than from the entire text of the relevant documents. For the "best rntidf top paragraph" approach a similar procedure was followed, except that instead of using the top 200 paragraphs from any relevant documents, the top scoring paragraph of each relevant document was used as a source for rntifd term selection. For each of the three approaches the maximum weight allowed for a term, i.e., 5, 6, 7, or 8 (see section 4.1), on a topic by topic basis, was the weight that gave the best performance on the training set.

Finally, the method of query expansion used on the officially submitted run, "best overall", was a combination of the three approaches described above. This method was to select the best query expansion provided by any of the three approaches on a topic per topic basis. Rather than simply selecting the best approach per topic in this manner, some consideration was given to trying to combine the results of the different methods [FS94, BKCQ94], but no experiments were carried out.

5 Summary

WIN was able to achieve strong performance on both the ad hoc retrieval and routing tasks without any major modifications being made to its retrieval engine. The ad hoc results show the effectiveness of its basic indexing and retrieval operations. Some techniques that were expected to give improved performance, did not lead to much improvement. In some cases this may be because only limited investigations could be done, e.g., when using the collection-dependent likelihood of relevance. In other cases, such as the failure of phrases, to yield much improvement, the result may indicate the difficulty in effective use of a feature which has given good results on smaller collections on a collection the size of the TREC collection [Har93].

References

[BCC93] John Broglio, James P. Callan, and W. Bruce Croft. INQUERY system overview. In *Proceedings of the TIPSTER Text Program (Phase 1) Workshop*, pages 47–67, Morgan Kaufmann, September 1993. ISBN:1-55860-337-9.

[BKCQ94] N. J. Belkin, P. Kantor, C. Cool, and R. Quatrain. Combining evidence for information retreival. In Donna K. Harman, editor, *The Second Text Retrieval Conference (TREC-2)*, pages 35–44, National Institute of Standards and Technology, March 1994. Proceedings available as NIST Special Publication 500-215.

[Cal94] James P. Callan. Passage-level evidence in document retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual International Conference on Research and Development in Information Retrieval*, pages 212–221, Springer-Verlag, London, July 1994.

[CCB94] W. Bruce Croft, Jamie Callan, and John Broglio. TREC-2 routing and adhoc retrieval evaluation using the INQUERY system. In Donna K. Harman, editor, *The Second Text Retrieval Conference (TREC-2)*, pages 75–84, National Institute of Standards and Technology, March 1994. Proceedings available as NIST Special Publication 500-215.

[Cro93] W. Bruce Croft. The University of Massachusetts TIPSTER project. In Donna K. Harman, editor, *The First Text Retrieval Conference (TREC-1)*, pages 101–105, National Institute of Standards and Technology, March 1993. Proceedings available as NIST Special Publication 500-207.

[FS94] Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. In Donna K. Harman, editor, *The Second Text Retrieval Conference (TREC-2)*, pages 243–252, National Institute of Standards and Technology, March 1994. Proceedings available as NIST Special Publication 500-215.

[Har93] Donna Harman. Document detection summary of results. In *Proceedings of the TIPSTER Text Program (Phase 1) Workshop*, pages 33–46, Morgan Kaufmann, September 1993. ISBN:1-55860-337-9.

[HC93] David Haines and W. Bruce Croft. Relevance feedback and inference networks. In Robert Korfhage, Edie Rasmussen, and Peter Willett, editors, *Proceedings of the Sixteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–11, June 1993.

[TC91] Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.

[Tur90] Howard Turtle. *Inference Networks for Document Retrieval.* PhD thesis, Computer Science Department, University of Massachusetts, Amherst, MA 01003, 1990. Available as COINS Technical Report 90-92.

[Tur94] Howard Turtle. Natural language vs. Boolean query evaluation: a comparison of retrieval performance. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the Seventeenth Annual International Conference on Research and Development in Information Retrieval*, pages 212–221, Springer-Verlag, London, July 1994.