Overview of the TREC 2010 Relevance Feedback Track (Notebook)

Chris Buckley[1], Matthew Lease[2], and Mark D. Smucker[3]

[1]Sabir Research
[2]School of Information, University of Texas at Austin
[3]Department of Management Sciences, University of Waterloo

Abstract

This year the relevance feedback track further examined relevance feedback with a single document relevance feedback task. Seven groups participated in the track. At this time, relevance judging is on-going and no results are available. This notebook version of the track describes the track, presents the current status of the track, and includes participant summaries.

1 Introduction

This is the third year of the TREC relevance feedback (RF) track. The first year concentrated on the RF algorithm itself. All participants were given the same sets of judged documents, and used their own algorithms to retrieve a new set of documents. In the second year, the concentration shifted to finding good sets of documents for feedback. This year the track aimed to examine what makes an individual document good or bad for feedback by focusing on single document relevance feedback as the track's main task.

With a focus on single document RF, the hope was that participating groups would examine the structure of the document and its language and step away from only using techniques that treat documents as unordered bags of words.

The basic user scenario is that a user has submitted a short (title only in TREC parlance) query to a retrieval system that has returned one or more documents to the user that it thinks are relevant/useful. The user then identifies a relevant document and submits positive feedback on this one document. Based upon this document and the original query, the system reconstructs the list of documents to be given the user. This new list is then evaluated for both accuracy (top documents are relevant) and completeness (all relevant documents are retrieved, not just those that have the same aspect as the known document). Like almost all relevance feedback tasks, we assume the user has a need for several or many relevant documents.

The track investigated this scenario by providing groups 100 topics for which we already had some known relevant and non-relevant documents. For each topic, we selected 5 documents to be used for single document RF. Groups were to treat this document as an example of a relevant document and submit 5 result lists for each topic. Groups also submitted baseline runs that utilized no relevance information. In addition, groups had the option to supply their opinion, in the form of a relative ranking, of how well the 5 feedback documents would perform. The aim of this ranking is to allow the track to determine if systems could pick out those documents that they feel will provide good relevance information.

Portions of this overview are taken directly from the track guidelines.

2 Methods and Materials

The track used the English portion (English[1-10]) of the ClueWeb09[1] collection. In 2009, a distinction was made between the Category B subset (CatB is English_1) and the full English portion. Participants could use the standard subset of full English[1-10], CatB, but the track encouraged use of the full dataset.

The topics consisted of the first 100 even topics from the TREC 2009 Million Query track (which includes those queries run in both web and relevance feedback tracks in 2009). By using the even topics, groups could train on the odd numbered topics.

For each topic, we selected 5 documents to be used individually for feedback. The five selected documents were:

[1] http://boston.lti.cs.cmu.edu/Data/clueweb09/

1. A randomly chosen document from among the topic's known relevant documents.
2. The most commonly returned relevant document.
3. The least commonly returned relevant document.
4. The longest relevant document.
5. The shortest relevant document.

The documents were not marked with the criteria used to choose them, and the documents chosen by any one criteria were randomly mixed among the sets of documents given the participants.

The track also provided a second set of 5 documents per topic for groups wishing to have their methods tested on an additional 5 factors. Groups were to only submit runs for this set of documents if they already had done the first set of 5. Groups were to use the same technique on this set as they did on the first set. These five selected documents were:

1. A random relevant document.
2. The most spammy relevant document where spaminess was from the fusion model of web spam provided by Cormack et al. [1]
3. The least spammy relevant document.
4. A random highly relevant document.
5. The most commonly returned non-relevant document.

With both sets of documents, the goal here was to start an investigation into some of the properties of documents and how they affect relevance feedback performance, and determine whether those affects are general or system-dependent.

3 Assessments

We are in the process of getting the top 10 results of all submitted runs assessed. We are using Amazon's Mechanical Turk for judging the relevance of documents. In order to protect the judges from malicious web pages, we produced a combination of screenshots, PDFs, and plain text versions of the web pages for judging. In most cases, the judges are shown an image of the webpage and can also view a PDF version and a plain text version. In some cases, one or more of the page viewing options is not available. In all cases, there is at least a plain text rendering. In the final overview paper, we will provide more details on the assessment process.

4 Evaluation

The primary official evaluation measure will be a recall oriented measure like MAP, statMAP (an approximation to MAP that handles missing judgments better), MAPjudged (MAP, but eliminating all unjudged documents from the top 1000 before the evaluation) or Prec@1000. All recall-oriented measures have problems on ClueWeb09 that are still being investigated. There will be a secondary official measure of Precision at 10 docs. These measures will be calculated using binary relevance judgments — possibly conflicting Mechanical Turk judgments will be coerced into binary judgments.

There will also be unofficial measures reported which will explicitly model the probability of relevance of a document given the Mechanical Turk judging process. However, since these measures have not yet been fully developed or tested, and there are as yet no test collections that participants can train their systems with using these measures, they will not be official measures, but only used as guides for future evaluations.

The official measures were chosen to reflect a task where the user is interested in not just a single relevant document, but all varieties of relevant documents for this topic. Thus a recall oriented measure like MAP is desired. In addition, we would like some measure that gives an indication of just performance at the top of the rankings. Ideally, we would like that measure to have a diversity component as in the Web Track 09, but given the Mechanical Turk environment, we don't want to commit to being able to do that.

A final single score for each participating group will be the average over all 5 runs over all the topics, for each of the measures. In addition, the input documents will be broken down into sets based on the criteria used to select them, one per topic. Thus there might be a set for "long documents" where each topic has one long relevant document in it. There will be a set for the "standard" randomly selected relevant documents, one per topic. Thus, scores will be reported for each of the 5 criteria sets.

The predictive relative effectiveness task for the track will be evaluated for each topic, comparing the predicted ordering of the 5 input documents against the actually recall system performance ordering the 5 documents. Average Kendall tau over all topics will be reported.

The second optional task will be evaluated the same as the required task. There will be results reported for the 5 additional categories of the second task.

5 Participating Groups

Seven groups submitted runs to the track: Beijing University of Posts and Telecommunications, Harbin Institute of Technology, Queensland University of Technology, Sabir Research, University of Amsterdam, University of Delaware, and University of Padova. Six of the seven submitted summaries of their work, which we include verbatim below.

5.1 Harbin Institute of Technology

Indri is used to build index and search. Three kinds of expanded queries are mined from the feedback documents: terms of pseudo RF, terms of explicit RF and term dependencies. The first terms are selected from the pseudo relevance documents in terms of TFIDF. The second terms are chosen from the given feedback document in the same way. Finally, we get the term dependencies in feedback document. The term dependency includes the consecutive ordered term pairs and the proximal unordered term pairs. We choose these term pairs with the following constrains: 1) the term pair contains at least one term of the original query; 2) the co-occurrence statistics of term pair exceed predefined thresholds. Each kind of expanded query is combined with the original query using weight operator. So we have three new queries. Retrieve these three queries and return three retrieval results. A classification model is trained to choose which retrieval result is the best. The model is logistic regression and features are defined on the basis of the original query, feedback documents, etc. Because all selected topics are even-numbered topics, we train the model on 100 odd-numbered topics from MQ track. We aggregate the best retrieval result per topic as our submission.

5.2 Queensland University of Technology

Sequential closed patterns in data mining have capacity to improve the performance of pattern-based information retrieval. In this track, we tested an innovative pattern mining approach, Relevance Feature Discovery (RFD), for using both positive and negative user feedback. RFD discovers both positive and negative patterns as higher level features in order to accurately evaluate low-level features (terms). This evaluation is completed based on the terms' specificity and their distributions in the higher level features, where a term's specificity describes the extent of the term to which the topic focuses on what users want. Based on the specificity of terms, in this research, low-level terms are classified into three categories: positive specific terms, general terms, and negative specific terms.

The detailed process is as follows: (i) Given a topic, 15000 relevant documents were extracted using Rocchio and Cosine similarity via content search; (ii) Using the high frequent terms extracted from user feedback, the 15000 documents were re-ranked again using Rocchio; (iii) The top 10 documents were selected as the positive feedback and the bottom 10 as negative. These documents then fed into the RFD model which calculated weights for all 15000 documents; (iv) Re-rank the 15000 documents using the new weights and submitted the top 2500.

5.3 Sabir Research

Once again Sabir Research submitted a very standard base-case set of runs for the relevance feedback track. There was nothing particularly new in what was done - all weighting and feedback approaches were developed at least 15 years ago. (The very basic approach was dictated in part due to Chris Buckley of Sabir Research being the only person who knew in advance the categories that each input document belonged to, and thus he wanted to ensure that nothing was done to take unfair advantage of that knowledge.)

All runs were made on the full category A set of English Clueweb documents. The base retrieval run was a SMART tf*idf run, with the documents being weighted with SMART Lnu weights and the queries with ltu weights. Feedback used a standard Rocchio algorithm with original query weights being given the same importance as weights from the single relevant document. The input relevant document was re-indexed with ltu weights and the top weighted 25 terms were added, where each term had to occur in at least 100 documents in the collection. A standard inner-product run was then done. All 10 input document sets were run, with each run taking about an hour of clock time – the added terms tended to be frequently occurring terms, and no query truncation was used here.

5.4 University of Amsterdam

We adopt a language modeling framework in which typical relevance feedback algorithms consider feedback documents as generative models from which to sample terms. We find that simply applying out-of-the-box relevance feedback algorithms to the single example document is not effective; such feedback algorithms degrade retrieval performance. To address this issue, we have implemented a novel model and our focus in our TREC participation this year is to evaluate its performance. Our proposed algorithm makes use of the moderated contents of Wikipedia as a pivot language. Wikipedia articles can be created by anyone, but they are typically moderated by a relatively small group of volunteers. Moreover, Wikipedia has extensive guidelines in place, pertaining to the correct use of grammar and style. As a consequence (and unlike common web pages), the language used in each article tends to be "clean" and to the point. It is this particular feature of Wikipedia that we use to influence the estimation of the language model of web pages. The expanded query language model is interpolated with the initial query to obtain a final representation of the user's information need.

5.5 University of Delaware

UD submitted a baseline indri Category B run using a weighted combination of Markov Random Field models—a full-text model, a title-only model, and a heading-only model. Each feedback run is generated by randomly reordering the baseline's top 40 ranked documents for each topic, using the provided relevant document to seed an RNG. Our goal was to explore hypotheses about topic definition and relevance judging in the ClueWeb corpus rather than about retrieval models per se.

5.6 University of Padova

The approach adopted by UPD aimed at investigating the effectiveness of relationship among terms in the feedback document modeled using local cooccurrence data.

The 10 terms with highest IDF were selected from the feedback document. The selected terms were adopted to build a symmetric term-by-term matrix computed by moving a window of text of size 7 centered on one of these terms at a time; if one of the selected terms appeared in the text window, the TF-IDF weight of the term was added to the two entries of the matrix involving both the term in the center of the text window and the considered term. The symmetric matrix was decomposed by SVD; in the experiments the first eigenvector was selected. Therefore term relationships were represented by the one dimensional subspace spanned the selected eigenvector, namely a vector in R^{k+10}, where k was the number of terms in the original query. Each of the $k+10$ elements represents a measure of the relationship with the other terms.

Document re-ranking was performed computing the distance between the obtained vector subspace representation for term relationships and the document representation. Documents were represented as vectors of BM25 weights. The distance computation resulted in altering the documents BM25 weights of the (expanded) query terms, specifically multiplying each BM25 weight by the corresponding term relationship weight in the selected eigenvector.

6 Conclusion

This year the relevance feedback track had a single document relevance feedback task. At this time, the submitted runs are still being assessed.

7 Acknowledgments

Special thanks to Catherine Grady and Chandra Prakash Jethani for their on-going assistance with the track and helping make the assessment process a reality.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), in part by an Amazon Web Services in Education Research Grant, and in part by the University of Waterloo. 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.

References

[1] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. `http://arxiv.org/abs/1004.5168`, 2010.