Description of the PRC CEO Algorithm for TREC-2

Paul Thompson
PRC Inc., Mail Stop 5S3
1500 PRC Drive
McLean, VA 22102
Phone: 612/687-4650
E-mail: thompson@research.westlaw.com
(current affiliation: West Publishing Company)

Abstract

This paper describes an application of the Combination of Expert Opinion technique to combine the results of multiple retrieval methods used on the TREC-2 collection. The methods being combined were weighted by their TREC-1 performance.

1. Introduction

This paper describes work done on the TREC-2 project at PRC Inc. in collaboration with Professor Edward Fox and his colleagues at Virginia Polytechnic Institute and State University (VPI&SU). The reader should refer to the description of their system included in these working notes for further details on the common processing of the TREC-2 data shared by PRC and VPI&SU (Fox et al. 1993). PRC used its algorithm, the Combination of Expert Opinion (CEO), to combine the results of VPI&SU's runs. VPI&SU used a different combination technique for their final results. Originally the intent was that the CEO algorithm would be integrated with the SMART system used by VPI&SU. Both upper and lower level combination of results would take place, i.e., at the lower level of individual document features within a particular retrieval method and the upper level of combination of the output of the individual methods themselves, i.e., the various cosine and p-norm methods used by VPI&SU. For TREC-1 we were not able to train the CEO algorithm, so that the weighting of the various methods would be optimized based on relevance judgments. For TREC-2 we used the 11 point average scores obtained by the various methods for TREC-1 for weighting. Again we only used the upper level of CEO. For TREC-1 we found that combining all methods resulted in lower performance than using the single best method. This year our first version combined the top two methods, based on TREC-1, while the second version used the top five methods.

2. Combination of Expert Opinion

The statistical technique of CEO provides a solution to the problem of combining different probabilistic models of document retrieval. This technique is expected to result in improved precision and recall over that provided by any one model, or method, since research has shown that various retrieval models retrieve different sets of more or less equally relevant documents (Katzer et al. 1982, Fox et al. 1988). In the

Bayesian formulation of the CEO problem (Lindley 1983) a decision maker is interested in some parameter or event for which he/she has a prior, or initial, distribution or probability. The decision maker revises the distribution upon consulting several experts, each with his/her own distribution or probability for the parameter or event. To effect this revision, the decision maker must assess the relative expertise of the experts and their interdependence, both with each other and the decision maker. The experts' distributions are considered as data by the decision maker, which is used to update the prior distribution.

For automatic document retrieval, the retrieval system is the decision maker, and different retrieval algorithms, or models, are the experts (Thompson 1990a,b, 1991). This is referred to as the upper level CEO. At the lower level the probabilities of individual features, e.g., terms, within a particular retrieval model can be combined using CEO. In lower level CEO the retrieval model is the decision maker and the term probabilities are viewed as lower level experts. The probability distributions supplied by these lower level experts can be updated, according to Bayes theorem, by user relevance judgments for retrieved documents. These same relevance judgments also give the system a way to evaluate the performance of each model, both in the context of a single search of several iterations and over all searches to date. These results can be used in a statistically sound way to weight the contributions of the models in the combined probability distribution used to rank the retrieved documents. Since various algorithms, such as p-norm, are expressed in terms of correlations rather than probability distributions, it was necessary to extend the CEO algorithm to handle correlations. So far this extension has been handled in a heuristic fashion. If a retrieval method, e.g., one of the cosine methods, returned a value between 0 and 1 as a retrieval status value; the logistic transformation of this weight was interpreted as an estimate of the mean of a logistically transformed beta distribution which was provided as evidence to the decision maker. Since there was no basis with which to assign a standard deviation to this distribution, as called for by the CEO methodology, an assumption was made that all standard deviations were .4045, a value corresponding to a standard deviation of .1 in terms of probabilities. The CEO code was written in g++.

For TREC-1 we used the CEO algorithm to combine all of the VPI&SU retrieval methods except for the Boolean, i.e., weighted and unweighted cosine and inner product measures as well as p-norm measures of 1.0, 1.5, and 2.0. For measures, such as the inner product and some of the p-norm results not giving a retrieval status value in the 0 to 1 range, the result was mapped to this interval by scaling the highest score of the method in question for a given topic to the highest score given by one of the cosine measures. Default scores half way between 0 and the lowest score achieved by a particular method were used for documents not retrieved in the top 200 in response to a given topic, since the actual score of these documents was unknown. For TREC-2 we followed the same approach except that only the results of methods with better TREC-1 performance were combined. Our first version used Cosine.atn and Cosine.nnn, the two best VPI&SU methods from TREC-1, weighted by their performance on TREC-1. The second version used these two methods and the next best three, Inner.atn, Inner.nnn, and Pnorm 1.0, also weighted by their TREC-1 performance (see VPI&SU report for details on these methods). Figures 1 and 2 show

our summary official TREC-2 results for versions 1 and 2 respectively.

```
Queryid (Num):       all  prceo1
Total number of documents over all queries
    Retrieved:      50000
    Relevant:       10785
    Rel_ret:         3561
Interpolated Recall - Precision Averages:
    at 0.00         0.4768
    at 0.10         0.2613
    at 0.20         0.1807
    at 0.30         0.1202
    at 0.40         0.0875
    at 0.50         0.0504
    at 0.60         0.0328
    at 0.70         0.0169
    at 0.80         0.0142
    at 0.90         0.0077
    at 1.00         0.0031
Average precision (non-interpolated) over all
rel docs
                    0.0904
Precision:
  At    5 docs:     0.2120
  At   10 docs:     0.2480
  At   15 docs:     0.2707
  At   20 docs:     0.2760
  At   30 docs:     0.2753
  At  100 docs:     0.2418
  At  200 docs:     0.1756
  At  500 docs:     0.1086
  At 1000 docs:     0.0712
R-Precision (precision after R (= num_rel for
a query) docs retrieved):
    Exact:          0.1703
```

Figure 1: Summary scores for PRC version 1 using two best experts

```
Queryid (Num):       all  prceo1
Total number of documents over all queries
    Retrieved:      50000
    Relevant:       10785
    Rel_ret:         3323
Interpolated Recall - Precision Averages:
    at 0.00         0.5963
    at 0.10         0.3270
    at 0.20         0.2306
    at 0.30         0.1430
    at 0.40         0.0866
    at 0.50         0.0425
    at 0.60         0.0323
    at 0.70         0.0246
    at 0.80         0.0209
    at 0.90         0.0131
    at 1.00         0.0041
Average precision (non-interpolated) over all
rel docs
                    0.1120
Precision:
  At    5 docs:     0.4120
  At   10 docs:     0.4000
  At   15 docs:     0.3920
  At   20 docs:     0.3740
  At   30 docs:     0.3527
  At  100 docs:     0.2722
  At  200 docs:     0.1974
  At  500 docs:     0.1090
  At 1000 docs:     0.0665
R-Precision (precision after R (= num_rel for
a query) docs retrieved):
    Exact:          0.1809
```

Figure 2: Summary scores for PRC version 2 using five best experts

3. Conclusion

Although selecting the best individual TREC-1 VPI&SU retrieval methods for combination and weighting them by their TREC-1 performance seemed to be a reasonable strategy which would yield better retrieval results than unweighted combination of all methods, in fact CEO performance was worse on TREC-2**.** This was due, in part, to changes made in the individual methods which made methods that had been best in TREC-1 less effective than other methods for TREC-2. These results suggest that selection and weighting of methods for combination based on performance of earlier versions of the methods is unwarranted.

References

Fox, E.A.; Koushik, P.; Shaw, J.; Modlin, R.; Rao, D. 1993. "Combining Evidence from Multiple Searches" this proceedings.

Fox, E.A.; Nunn, G.L.; Lee, W.C. 1988. "Coefficients for Combining Concept Classes in a Collection" *Proceedings of the 11th. International Conference on Research and Development in Information Retrieval*, June 13-15, Grenoble, France, p. 291-307.

Katzer, J.; McGill, M.J.; Tessier, J.A.; Frakes, W.; DasGupta, P. 1982. "A study of the overlap among document representations" *Information Technology: Research and Development*, vol. 2, p. 261-274.

Lindley, D.V. 1983. "Reconciliation of probability distributions" *Operations Research*, v. 31, no. 5, p. 866-880.

Thompson, P. 1990a. "A Combination of Expert Opinion Approach to Probabilistic Information Retrieval, Part 1: The Conceptual Model" *Information Processing and Management*, vol. 26, no. 3, p. 371-382.

Thompson, P. 1990b. "A Combination of Expert Opinion Approach to Probabilistic Information Retrieval, Part 2: Mathematical Treatment of CEO Model 3" *Information Processing and Management*, vol. 26, no. 3, p. 383-394.

Thompson, P. 1991. "Machine Learning in the Combination of Expert Opinion Approach to IR" In Birnbaum, L. and Collins, G. (eds.) *Machine Learning: Proceedings of the Eighth International Workshop (ML91)*, San Mateo: Morgan Kaufmann, p.270-274.