

Improving Relevance Feedback in Language Modeling with Score Regularization

Fernando Diaz^{*}

Yahoo! Inc.

1000 Rue de la Gauchetiere, Suite 2400

Montreal, QC

diazf@yahoo-inc.com

ABSTRACT

We demonstrate that regularization can improve feedback in a language modeling framework.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Relevance Feedback

General Terms: Algorithms

Keywords: regularization, relevance feedback, language modeling

1. INTRODUCTION

In many information retrieval scenarios, in addition to a query, a user also supplies sample relevant and non-relevant documents. These judgments may be provided with the query or in response to some initial query probing the collection for documents. The second scenario, referred to as *relevance feedback*, is the focus of this paper.

We would like to improve the performance of a language modeling baseline for relevance feedback. In particular, we are interested in augmenting the generative and one-class approaches with a model of the discriminative information conveyed by positive and negative feedback information. In this poster, we propose a method for introducing relevance information by forcing scores of documents related to relevant documents to be high and those related to non-relevant documents to be low. We accomplish this by using a document re-ranking technique known as *score regularization*. Given an initial set of retrieval scores, score regularization refers to a process of re-scoring documents in order to improve the consistency of scores of topically related documents [2]. We demonstrate that we can improve relevance feedback by conducting a post-retrieval re-ranking which incorporates relevance information and document similarities.

2. RELEVANCE FEEDBACK USING LANGUAGE MODELS

The language modeling approach to information retrieval ranks documents by comparing each document's smoothed language model, θ_d , to a language model estimated from the user's short query, θ_Q . Without relevance judgments,

^{*}Work conducted at the Center for Intelligent Information Retrieval.

the query model is usually a maximum likelihood estimate. When document relevance information is provided, we can estimate the *true relevance model* directly with binary weights [5, p. 69],

$$P(w|\theta_R) = \lambda P(w|\theta_Q) + (1 - \lambda) \sum_{d \in \mathcal{R}^+} \frac{1}{|\mathcal{R}^+|} P(w|\theta_d) \quad (1)$$

\mathcal{R}^+ is the set of documents judged relevant. Documents are then ranked according to cross entropy,

$$y_d = \sum_{w \in \mathcal{V}} P(w|\theta_Q) \log P(w|\theta_d) \quad (2)$$

where \mathbf{y} is vector of document scores.

We note that, unlike the vector space model and the probabilistic retrieval model, there is no formal model of *non-relevance* in relevance feedback based on true relevance models. True relevance models approach information retrieval from the perspective of density estimation. Relevant examples provide statistics for the true relevance model. The non-relevance model, by default, is the language model estimated using collection statistics, $P(w|\theta_C)$. Since the majority of the collection is non-relevant, the information from additional non-relevant documents is insignificant. This might be seen as a minor theoretical detail given the empirical evidence that negative feedback does not result in significant improvements [1, 3]. However, we believe that the information in explicitly non-relevant documents can be useful in situations where no relevant documents are retrieved and the system must filter non-relevant information. For example, if the only known keywords for a topic retrieve a cohesive, non-relevant cluster, we would like to provide information to remove the entire non-relevant cluster [7]. Furthermore, although the collection is a reasonable model of non-relevance, high-ranking non-relevant documents aid in refining the decision boundary between relevant and non-relevant documents.

3. REGULARIZED RELEVANCE FEEDBACK

Given an initial set of retrieval scores, \mathbf{y} , score regularization refers to a process of re-scoring documents in order to improve the consistency of scores of topically related documents [2]. The output of regularization is a vector of scores, \mathbf{f} , which minimizes two cost functions. One cost function, $\mathcal{S}(\mathbf{f})$, measures the *dissimilarity* of scores of related documents. Document relationships are encoded in a $n \times n$ matrix, \mathbf{W} , of inter-document similarities. The other cost function, $\mathcal{E}(\mathbf{f}, \mathbf{y})$, measures the *dissimilarity* of scores of the

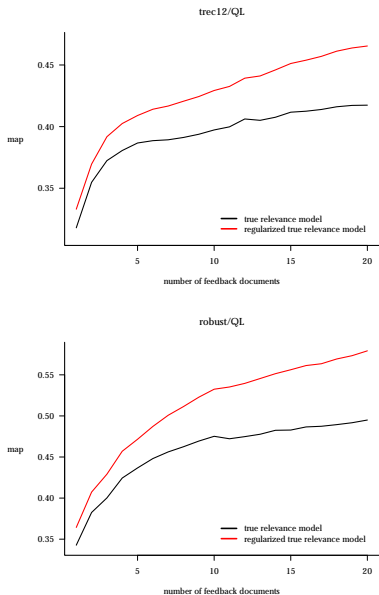


Figure 1: Relevance feedback results for trec12 and robust. Without feedback, trec12 QL performance is 0.2506 and 0.2800 if regularized. Robust QL performance is 0.2649 and 0.2955 if regularized.

documents with the original retrieval scores. We linearly combine these into a composite function,

$$\mathcal{Q}(\mathbf{f}, \mathbf{y}) = \mathcal{S}(\mathbf{f}) + \mu \mathcal{E}(\mathbf{f}, \mathbf{y}) \quad (3)$$

The constraints are defined as,

$$\mathcal{S}(\mathbf{f}) = \mathbf{f}^T \Delta \mathbf{f} \quad \mathcal{E}(\mathbf{f}, \mathbf{y}) = \|\mathbf{f} - \mathbf{y}\|_2^2$$

where, letting $D_{ii} = \sum_j W_{ij}$, $\Delta = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2}$ is known as the combinatorial Laplacian. The Laplacian has been empirically shown to be an effective method for measuring the consistency of the scores of related documents. The closed form solution for computing \mathbf{f}^* is,

$$\mathbf{f}^* = (1 - \alpha)(\alpha \Delta + (1 - \alpha) \mathbf{I})^{-1} \mathbf{y} \quad (4)$$

where $\alpha = \frac{1}{1 + \mu}$.

Whereas true relevance modeling is a non-parametric density estimation method, regularization is a non-parametric function approximation method. One advantage of approaching this as function approximation is that we can explicitly model non-relevance differently than we do uncertainty. Put another way, in Equation 2, a score of 0 represents both non-relevant documents and unjudged documents. In regularization, if we normalize scores to zero mean and unit variance, we can explicitly model relevant document scores (eg, $y_i > 1$), non-relevant document scores (eg, $y_i < -1$), and unjudged documents (eg, $y_i = 0$). In practice, for each relevant document, we replace its score with a value sampled from the region of the Gaussian greater than the maximum score. We do the same replacement for each non-relevant document by using samples from the bottom region of the Gaussian. Given these adjusted scores, we compute regularized scores according to Equation 4.

4. EXPERIMENTS

We performed experiments on two data sets. The first data set, which we will call “trec12”, consists of the 150 TREC Ad Hoc topics 51-200. We used only the news collections on Tipster disks 1 and 2 [4]. The second data set, which we will call “robust”, consists of the 250 TREC 2004 Robust topics [6]. We used only the news collections on TREC disks 4 and 5. For both data sets, we use only the title queries. We indexed collections using the Indri retrieval system, the Rainbow stop word list, and Krovetz stemming. We use TREC judgments (qrels) to simulate feedback.

We measure the performance of the system after receiving feedback on the top k documents. After the “user” marks the top k documents as relevant or non-relevant, we re-retrieve documents using the true relevance, $P(w|\theta_R)$. This is our baseline. We then normalize and regularize the scores and compare them to our baseline. We note that judged documents are included in the evaluation set so that we are testing the ability to re-retrieve relevant documents and penalize non-relevant documents; this intended to avoid problems with queries poorly represented in the corpus.

We present the results of these experiments in Figure 1. The horizontal axis of this graph represents the number of documents judged in the initial retrieval. Given the results in [2], we should not be surprised that regularization consistently improves the performance of retrieval. However, the amount of improvement grows with the number of relevance judgments. We suspect that, as the number of judgments increases, the regularization component of the system becomes more important because the additional data introduces a more discriminative component to standard true relevance models, allowing us to take advantage of additional data.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the author and do not necessarily reflect those of the sponsor.

6. REFERENCES

- [1] I. J. Aalbersberg. Incremental relevance feedback. In *SIGIR 1992*, pages 11–22, 1992.
- [2] F. Diaz. Regularizing query-based retrieval scores. *Information Retrieval*, 10(6):531–562, December 2007.
- [3] M. D. Dunlop. The effect of accessing nonmatching documents on relevance feedback. *TOIS*, 15(2):137–153, 1997.
- [4] D. K. Harman. The first text retrieval conference (TREC-1) Rockville, MD, U.S.A., 4-6 November, 1992. *Information Processing and Management*, 29(4):411–414, 1993.
- [5] V. Lavrenko. *A Generative Theory of Relevance*. PhD thesis, University of Massachusetts, 2004.
- [6] E. Voorhees. Overview of the TREC 2004 robust track. In *TREC 2004*, 2004.
- [7] X. Wang, H. Fang, and C. Zhai. Improve retrieval accuracy for difficult queries using negative feedback. In *CIKM*, pages 991–994, 2007.