Optimizing Parameters in a Ranked Retrieval System Using Multi-Query Relevance Feedback

Brian T. Bartell* Garrison W. Cottrell Richard K. Belew

Institute for Neural Computation, and
Department of Computer Science & Engineering
University of California, San Diego
La Jolla, California 92093-0114

Abstract

A method is proposed by which parameters in ranked-output text retrieval systems can be automatically optimized to improve retrieval performance. A ranked-output text retrieval system implements a ranking function which orders documents, placing documents estimated to be more relevant to the user's query before less relevant ones. The proposed method is to adjust system parameters to maximize the match between the system's document ordering and the user's desired ordering, given by relevance feedback. The utility of the approach is demonstrated by estimating the similarity measure in a vector space model of information retrieval. The approach automatically finds a similarity measure which performs equivalent to or better than all "classic" similarity measures studied. It also performs within 1% of an estimated theoretically optimal measure.

1 Introduction

State of the art document retrieval systems have a large number of free parameters, such as the weights on terms in documents, parameters of the similarity metric, retrieval thresholds, etc. One goal in the development of a retrieval system is to set and adjust these parameters to tune the system for better retrieval performance. For example, in ranked-output retrieval systems the system implements a *ranking function* which orders documents based on their estimated relevance to the user's query. Adjusting system parameters typically results in alternative orderings of documents made by the system. The goal then is to adjust the parameters so that, as much

*Brian Bartell is presently with the Advanced Technology Group, Encyclopædia Britannica, Inc., La Jolla, CA, 92037. This paper is reprinted from the Procedings of the Third Annual Symposium on Document Analysis and Information Retrieval with permission from the Information Science Research Institute, University of Nevada, Las Vegas.

as possible over a range of possible queries, the ordering of the documents given by the system corresponds well to the actual ordering of the documents' relevance to the user's need. This can be an extremely difficult task, both because of the large number of parameters and because of the often obscure relationship between parameter values and system performance.

We propose a computational framework which supports the explicit optimization of retrieval system parameters. The optimization improves the ability of the retrieval system to rank documents well for a finite set of users' queries. Using this set of queries along with the users' judgments of the desired ordering of retrieved documents, the system is optimized to match the desired ordering. The retrieval system is optimized by automatically adjusting certain free parameters in the system. Adjusting the parameter values changes how the system ranks documents; thus, our method changes the parameters to result in the preferred ranking.

2 Problem Statement

Let $R_{\Theta,q}(d)$ be the *ranking function* implemented by the text retrieval system to be optimized. $R_{\Theta,q}(d)$ is a function which assigns a real value (from the set of reals, $\Re$) to each document d in the collection of documents. The real value assigned to a particular document depends on both the user's query, q, and the parameters of the system, Θ. The ranking value for each document is meant to imply the document's estimated relevance to the query: Documents with large values are considered more relevant to the user's need than document's with smaller values. As $\Re$ is well ordered, $R_{\Theta,q}(d)$ provides an ordering over all documents. Typically the retrieval system will display the retrieved documents to the user in the order given by $R_{\Theta,q}(d)$. This formulation of the ranked retrieval problem is sufficiently general to include many popular methods for ranked retrieval, such as the vector space model [12] the probabilistic ranking model [3], and spreading activation models [2].

The goal of the method is to find values of the parameters Θ which result in the best ordering of documents. The target ordering is typically given by user feedback to the system. For each query issued by a user, the user indicates which of the retrieved documents are preferred over the others. This provides a target ordering over the documents for the set of past user queries. We formalize this target ordering with the preference notation of Wong, et. al.: the binary relation $\succ_q$ denotes the preference ordering over documents for the query q [14]. For two documents d and d',

$$d \succ_q d' \iff \text{the user prefers } d \text{ to } d' \tag{1}$$

We let Q denote the set of past user queries. The problem we are solving is to find parameter values resulting in document orderings which best match the constraints given by $\succ_q$ over the set of queries Q. Our method must find parameters Θ such that

$$\forall d, d' \in D, \quad R_{\Theta,q}(d) > R_{\Theta,q}(d') \text{ whenever } d \succ_q d' \tag{2}$$

where D is the collection of documents. When no Θ exists which can provide the perfect ranking, we desire a Θ which matches the target ordering $\succ_q$ as well as possible.

3 Method

Our approach is to optimize a criterion derived from Guttman's Point Alienation [6], a measure of rank correlation between two variables. An overview of the approach is provided here; further details can be found in other work [1]. The criterion which is optimized is:

$$J(R_\Theta) = \frac{-1}{\mid Q \mid} \sum_{q \in Q} \frac{\sum_{d \succ_q d'} (R_{\Theta,q}(d) - R_{\Theta,q}(d'))}{\sum_{d \succ_q d'} \mid R_{\Theta,q}(d) - R_{\Theta,q}(d') \mid} \tag{3}$$

where Q is the set of training queries, and d, d' are from the document set D. This function is an average, over the queries in Q, of the rank match between $\succ_q$ and $R_{\Theta,q}$. Note that when $R_{\Theta,q}$ perfectly orders the documents with respect to $\succ_q$, the numerator and denominator are equivalent, and the ratio is 1.0. When R is completely misordered, the numerator is the negative of the denominator, and the ratio is -1.0.

We minimize $J(R_\Theta)$ using Conjugate Gradient, a gradient-based optimization method [11]. J is mostly differentiable and is therefore suitable for gradient-based optimization. There are pathological cases in which the gradient of J is not defined. One such critical point occurs when all $R_{\Theta,q}(d)$ are equivalent for $d \in D$. This is an extreme condition, and has not occurred in our simulations. A more prevalent critical point occurs when $R_{\Theta,q}(d) = R_{\Theta,q}(d')$, as the derivative of $\mid x \mid$, $x = 0$, is required. In this case, we define $\partial \mid x \mid / \partial x = -1.0$ for $x = 0$; this has the effect of forcing $R_{\Theta,q}(d)$ to be strictly greater than $R_{\Theta,q}(d')$ when $d \succ_q d'$.

As we are using gradient optimization, this method can only be used to optimize parameters Θ by which the ranking function is differentiable. This can be an important limitation because many problems in IR do not satisfy this constraint, for example the selection of a stemming algorithm. A second limitation is that the gradient method may not find the globally optimal parameter settings if the optimization surface has many false (local) optima. We alleviate this problem by using multiple optimizations from different starting points, when necessary.

4 Application: Learning the Optimal Similarity Measure

To illustrate the utility of the proposed method, we apply the method to a well known problem in vector space text retrieval: Choosing the best vector similarity measure. The problem is interesting for both historical and pragmatic reasons. Historically, a large number of similarity measures have been proposed (McGill, et. al., [10] and Jones & Furnas [8] contain lengthy enumerations), and no single similarity measure has gained universal acceptance. Pragmatically, the selection of similarity measure can have a significant effect on the performance of the system; for example, Harman [7] demonstrates a 12% improvement in average precision by switching between two different normalized inner product measures.

We address this problem by optimizing the parameters of a general class of simi-

Θ_1	Θ_2	Standard Measure
-	0	Inner Product
2	0.5	Cosine
1	1	Pseudo-Cosine
0	1	Number of Terms

Table 1: For certain values of Θ, the parameterized similarity measure $R_{\Theta,q}(d)$ is identical to certain classic vector similarity measures typically used in text retrieval. Note that when $\Theta_2 = 0$, any finite value of Θ_1 yields the inner product measure.

larity measures. Let $R_{\Theta,q}$ be the similarity measure, defined as:

$$\begin{aligned} R_{\Theta,q}(d) &= \frac{\mathbf{q}^T\mathbf{d}}{(\sum_{\mathbf{d}_i \neq 0} \mathbf{d}_i{}^{\Theta_1})^{\Theta_2}} \\ &= \frac{\text{inner product}}{\text{doc length normalization}} \end{aligned} \tag{4}$$

where T is the transpose operator, and $\mathbf{q}$ and $\mathbf{d}$ are query and document column vectors derived from samples q and d (a domain constraint is that all $\mathbf{d}_i$ are non-negative). As Table 1 illustrates, various values of Θ correspond to common similarity measures in the text retrieval literature [8]. We have selected this model and parameterization because of its correspondence with well known measures and the vector L-norms. In addition, the model has only two parameters, which facilitates visualization and analysis of the optimization problem.

We validate the utility of optimizing our criterion function by comparing it to a more standard measure of performance in text retrieval, precision averaged over a range of recall levels. Like the criterion J, average precision measures the ranking performance of the system, though it is not a differentiable measure and therefore cannot be optimized using gradient-based numerical methods. We compare the criterion J to average precision by sampling both $J(R_\Theta)$ and average precision for different values of Θ. Figure 1 illustrates the surface of J using the CISI text collection of information science abstracts (a benchmark collection in the text retrieval literature). The complete set of 1460 documents and 76 relevance-tagged queries are used in evaluating J. Figure 2 demonstrates empirically that J is well correlated with average precision – visually, the two surfaces are similar, and the correlation between corresponding Θ is very high. Optimizing J will therefore likely improve average precision. We have performed this same comparison using a number of other criteria from the statistical, psychometrics, and text retrieval literatures, including Kruskal's stress function [9], a variation of Pearson's correlation coefficient [4], Wong's Perceptron criterion [13] and Fuhr & Buckley's Least Squared Error criterion [5]. Of all the alternative criteria studied, none matched average precision as closely as the criterion we have selected. Note that this comparison is done only to validate that our criterion is closely correlated with average precision, a standard measure of performance in IR; an application of the technique to a real-world problem need not duplicate this effort.

We now demonstrate that the approach can be used to *automatically* find parameters of a similarity measure which outperform or are equivalent to other standard

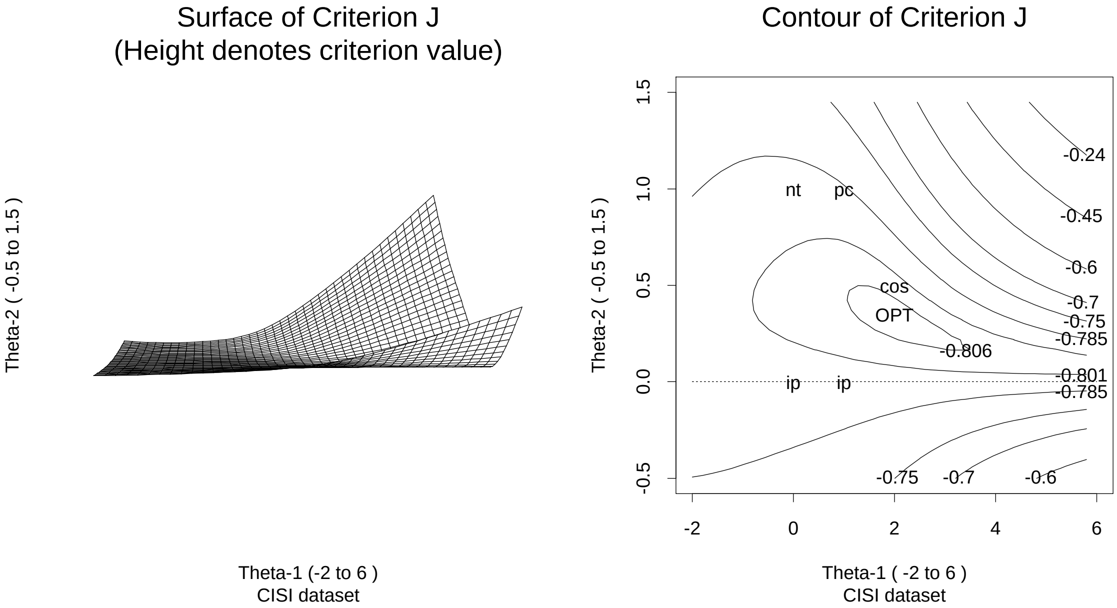

Figure 1: The surface of J is illustrated by sampling J at 1,681 (41×41) different values of Θ. The 3D perspective at left highlights J's general concave-up bowl shape. At right is a contour plot of the same surface. Parameter settings corresponding to classic similarity measures are identified ("nt" is number of terms, "pc" is pseudo-cosine, "ip" is inner product, and "cos" is cosine). The minimum of the surface (the optimal parameter setting) is identified by "OPT".

measures. J is optimized using conjugate gradient to determine the optimal parameter settings. In this case, the 76 relevance-tagged queries from the CISI collection are randomly partitioned into training and test sets. 51 queries are placed in the training set and are used in the optimization of J, and the remaining 25 queries are used to evaluate performance on "novel" queries. The results of the optimization for a single random partition of the queries and a comparison to the performance of other similarity measures are provided in Table 2. The similarity measure learned by our method is within 1% of the theoretical optimum average precision performance. This theoretical optimum is found by exhaustively sampling average precision over the 25 test queries to find the similarity measure which performs best on the test set. Our measure also outperforms all the classic measures which are realizable by the model. Finally, our measure outperforms the measure found by exhaustively sampling average precision over the 51 training queries. The difference in performance between these two measures may seem counter-intuitive because the two criteria (J and average precision) are so highly correlated (as was illustrated in Figure 2). However, the optima of the two criteria are not identical, and the performance of the two measures may therefore differ somewhat in both training set and test set performance.

This experiment is repeated 7 additional times, choosing different random test and training sets each time. This is done to factor out variance due to the particular random selection of training and test queries. In the repeated experiments, our

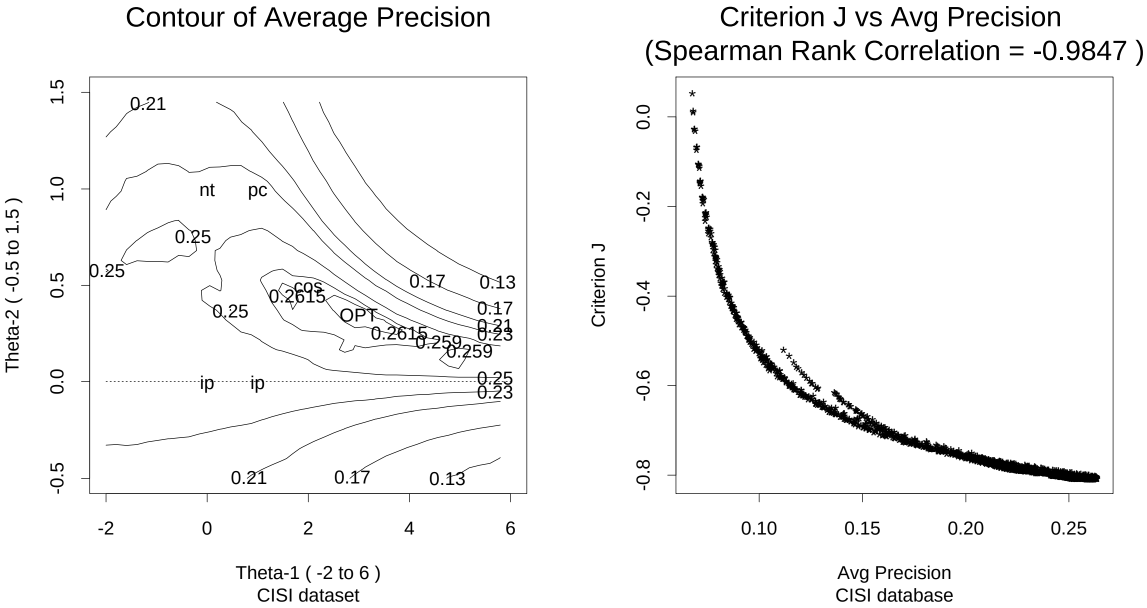

Figure 2: The average precision surface (at left) is structurally quite similar to the surface of J in the previous figure. In fact, samples of the two surfaces are highly correlated (at right). The correlation is negative because maxima of the average precision surface correspond to minima of J.

method is found to significantly outperform both the pseudo-cosine and number-of-terms measures (with $p < 0.0005$). It also outperforms both cosine and inner product, though the difference is not statistically significant. We are unable to compare the method to the exhaustive sampling procedures, due to the computational expense.

We emphasize that none of the similarity measures, not even the theoretical optimum, provide exceptional average precision performance. This fact illustrates a critical point: The method finds optimal parameters completely within the scope of a specific retrieval model. The optimal solution found by our method is therefore limited by the expressiveness of this parameterized model. In the model of similarity measures used here, even the theoretical optimum does not outperform the best classic measures by much. Within this restricted domain, our method is able to find a similarity measure which performs very close to optimal. Nevertheless, our method places important emphasis on the task of defining an adequate model. To achieve truly exceptional performance, the model must be rich enough to make the distinctions necessary to rank relevant documents before less relevant ones. Our method then provides an automatic approach to setting the parameters of the model to achieve this goal.

5 Summary

A novel method for automatically optimizing the differentiable parameters in ranked free-text retrieval systems has been proposed. The method is based on minimizing a

Metric	Θ_1	Θ_2	Test Set Avg Precision	Percent Off Optimal
Number of Terms	0	1	.2428	-14%
Pseudo-Cosine	1	1	.2439	-14%
Inner Product	-	0	.2439	-14%
Exhaustive Search (Training Set)	2.8	0.15	.2661	-6%
Cosine	2	0.5	.2727	-3%
Optimization of J	2.37087	.318704	.2778	-1%
Exhaustive Search (Test Set)	3.2	0.3	.2820	-

Table 2: Results of learning the optimal similarity measure using a single random partitioning of the query set into 51 training queries and 25 test queries. The learned similarity measure outperforms all classic similarity measures studied. It also performs within 1% of the estimated optimal similarity measure (found by exhaustive sampling of the parameter space in search of the best test-set average precision performance).

mostly differentiable criterion function, which measures how well the system is ranking documents relative to a target ordering given by user feedback to the system. The method is applied to a problem of interest in text retrieval: Identifying the optimal similarity measure in a vector space retrieval model. The method is able to select from an infinite range of possible similarity measures and find high quality measures which are better than or equivalent to all classic measures studied.

Acknowledgements: This work was supported in part by NSF grant IRI-9221276.
Address correspondence to the first author. Email: *bbartell@eb.com.* US mail: Advanced Technology Group, Encylopædia Britannica, Inc., 3252 Holiday Court, Ste. 208, La Jolla, CA 92037.

References

[1] Brian T. Bartell. *Optimizing Ranking Functions: A Connectionist Approach to Adaptive Information Retrieval.* PhD thesis, Department of Computer Science & Engineering, The University of California, San Diego, 1994.

[2] R.K. Belew. Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents. In *Proc. SIGIR 1989*, pages 11–20, Cambridge, MA, June 1989.

[3] A. Bookstein and D. R. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25:312–318, 1974.

[4] I. Borg and J. Lingoes. *Multidimensional Similarity Structure Analysis.* Springer-Verlag, New York, 1987.

[5] Norbert Fuhr and Chris Buckley. A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems*, 9(3):223–248, 1991.

[6] L. Guttman. What is not what in statistics. *The Statistician*, 26:81–107, 1978.

[7] Donna Harman. An experimental study of factors important in document ranking. In *Proceedings of the ACM SIGIR*, pages 186–193, Pisa, Italy, 1986.

[8] William P. Jones and George W. Furnas. Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38(6):420–442, 1987.

[9] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, March 1964.

[10] M. McGill, M. Koll, and T. Noreault. An evaluation of factors affecting document ranking by information retrieval systems. Technical report, School of Information Studies, Syracuse University, Syracuse, New York 13210, October 1979.

[11] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 1988.

[12] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., 1983.

[13] S. K. M. Wong, Y. J. Cai, and Y. Y. Yao. Computation of term associations by a neural network. In *Proceedings of SIGIR*, Pittsburgh, PA, June 1993.

[14] S. K. M. Wong and Y. Y. Yao. Query formulation in linear retrieval models. *Journal of the American Society for Information Retrieval*, 41(5):334–341, 1990.