# A Logistic Regression Approach to Distributed IR

Ray R. Larson
School of Information Management and Systems
University of California, Berkeley
Berkeley, California, USA, 94720-4600
ray@sherlock.berkeley.edu

## ABSTRACT

This poster session examines a probabilistic approach to distributed information retrieval using a Logistic Regression algorithm for estimation of collection relevance. The algorithm is compared to other methods for distributed search using test collections developed for distributed search evaluation.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.3.7 [**Digital Libraries**]: Systems issues

## General Terms

Algorithms, Performance

## Keywords

Distributed Information Retrieval

## 1. INTRODUCTION

This poster will present some recent effectiveness results for the application of Logistic Regression to the problem of distributed Information Retrieval (IR). This approach, and the rationale for its use, is described in this extended abstract of the work. During the poster session new experiments and evaluation currently underway will be described. In this proposal we briefly describe the problem, our approach and how this approach compares to some other well-known methods for distributed search using test collections developed for distributed search evaluation[5, 4, 8].

As increasing numbers of sites around the world make their databases available through protocols such as OAI or Z39.50 the problem arises of determining, for any given query, which of these databases are likely to contain information of interest to a world-wide population of potential users. This is the central problem of distributed IR and has been an area of active research interest for many years. Some of the best known work has been that of Gravano, et al. [6] on GlOSS and Callan's [1] application of inference networks to distributed IR (CORI). French and Powell, along with a number of collaborators [5, 4, 8], enabled comparative evaluation of distributed IR by defining test collections derived from TREC data, where the TREC databases are divided into sub-collections representing virtual distributed collections. In addition they defined a number of measures for evaluation of the performance of distributed IR[5] used in this paper to compare previously published results on collection selection with a probabilistic model based on logistic regression.

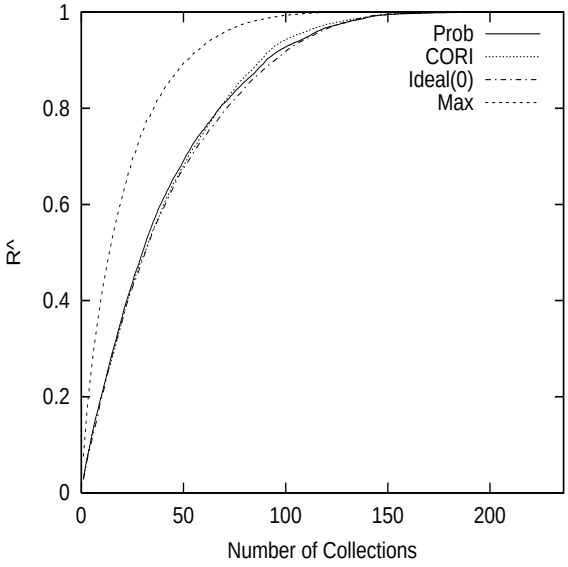


**Figure 1: Title Query Performance**

## 2. THE APPROACH

The algorithms developed for this research are based on the *logistical regression* method developed by researchers at U.C. Berkeley and tested in TREC evaluations [3, 2]. Since the "collection documents" used for this evaluation represent collections of documents and not individual documents, a number of differences from the usual logistic regression measures were used. In addition, analysis showed that different forms of the TREC queries (short titles only, longer queries including the concepts fields and the very long title, concepts, description and narrative) behaved quite differently in searching the distributed collection, so three different regression equations were derived and applied automatically based on the length of the query. In this paper only the "title" and "long" (title and concepts) queries are discussed, since they best approximate actual queries.

In the logistic regression model of IR, the estimated proba-


*SIGIR'02,* August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

bility of relevance for a particular query and a particular collection (or collection document) $P(R \mid Q, C)$, is calculated and collections are ranked in order of decreasing values of that probability. In the current system this is calculated as the "log odds" of relevance $\log O(R \mid Q, C)$, Logistic regression provides estimates for a set of coefficients, $c_i$, associated with a set of $S$ statistics, $X_i$, derived from the query and database of collection documents, such that:

$$\log O(R \mid Q, C) \approx c_0 \sum_{i=1}^{S} c_i X_i \quad (1)$$

where $c_0$ is the intercept term of the regression. For the set of $M$ terms that occur in both a particular query and a given collection document. This formula is similar to that used in TREC 3[2]. The statistics used in this study were:

$X_1 = \frac{1}{M} \sum_{j=1}^{M} logQAF_{t_j}$ . This is the log of the absolute frequency of occurrence for term $t_j$ in the query averaged over the $M$ terms in common between the query and the document.

$X_2 = \sqrt{QL}$ . This is square root of the query length (i.e., the number of terms in the query disregarding stopwords).

$X_3 = \frac{1}{M} \sum_{j=1}^{M} logCAF_{t_j}$ . This is is the log of the absolute frequency of occurrence for term $t_j$ in the collection averaged over the $M$ common terms.

$X_4 = \sqrt{\frac{CL}{10}}$ . This is square root of the collection size.

$X_5 = \frac{1}{M} \sum_{j=1}^{M} logICF_{t_j}$ . This is is the log of the *inverse collection frequency*(ICF) averaged over the $M$ common terms. ICF is calculated as the total number of collections divided by the number that contain term $t_j$

$X_6 = logM$ . The log of the number of terms in common between the collection document and the query.

For short (title only) queries, for example, the equation used in ranking was:

$$\begin{aligned} \log O(R \mid Q, C) &= -3.70 + (1.269 * X_1) + (-0.310 * X_2) \\ &\quad +(0.679 * X_3) + K \\ &\quad +(0.223 * X_5) + (2.01 * X_6); \end{aligned}$$

($K$ is a constant because query term frequency is always 1 in short queries).

# 3. EVALUATION

We used collections formed by dividing the documents on TIPSTER disks 1, 2, and 3 into 236 sets based on source and month[5]. Collection relevance information was based on whether *any* documents in the collection were relevant according to the relevance judgements for TREC queries 51-150. The relevance information was used both for estimating the logistic regression coefficients (using a sample of the data) and for the evaluation (with full data).

Figure 1 summarizes some preliminary results of the evaluation. The X axis is the number of collections in the ranking and the Y axis, $\hat{R}$, is a Recall analog that measures the proportion of the total possible relevant documents that have been accumulated in the top $N$ databases, averaged across all of the queries. The Max line is the optimal results based where the collections are ranked in order of the number of relevance documents they contain. Ideal(0) is an implementation of the GlOSS "Ideal" algorithm and CORI is an implementation of Callan's Inference net approach (described in [8]. The Prob line is the logistic regression method described above. For title queries (Figure 1) the described method performs slightly better than the CORI algorithm for up to about 100 collections, where CORI exceeds it. Both CORI and the logistic regression method outperform the Ideal(0) implementation.

We are continuing to examine probabilistic distributed IR using logistic regression. Further work is underway to refine the model and to apply it to actual systems in the U.K. using collection harvesting techniques described in [7]. This ongoing work will be described at the poster session.

# 4. ACKNOWLEDGMENTS

The author would like to thank James French and Allison Powell for kindly suppling the CORI and Ideal(0) results used in the evaluation for this paper. They were derived from versions of the algorithms as described in Allison Powell's dissertation[8]. This work was supported by the National Science Foundation and the Joint Information Systems Committee (U.K.) under *International Digital Libraries Program* award #IIS-9975164.

# 5. REFERENCES

[1] J. Callan. Distributed information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval: Recent research from the Center for Intelligent Information Retrieval*, chapter 5, pages 127–150. Kluwer, Boston, 2000.

[2] W. S. Cooper, F. C. Gey, and A. Chen. Full text retrieval based on a probabilistic equation with coefficients fitted by logistic regression. In D. K. Harman, editor, *The Second Text Retrieval Conference (TREC-2)*, pages 57–66, Gaithersburg, MD, 1994. NIST.

[3] W. S. Cooper, F. C. Gey, and D. P. Dabney. Probabilistic retrieval based on staged logistic regression. In *SIGIR '92*, pages 198–210, New York, 1992. ACM.

[4] J. C. French, A. L. Powell, J. P. Callan, C. L. Viles, T. Emmitt, K. J. Prey, and Y. Mou. Comparing the performance of database selection algorithms. In *SIGIR '99*, pages 238–245, 1999.

[5] J. C. French, A. L. Powell, C. L. Viles, T. Emmitt, and K. J. Prey. Evaluating database selection techniques: A testbed and experiment. In *SIGIR '98*, pages 121–129, 1998.

[6] L. Gravano, H. García-Molina, and A. Tomasic. GlOSS: text-source discovery over the Internet. *ACM Transactions on Database Systems*, 24(2):229–264, 1999.

[7] R. R. Larson. Distributed resource discovery: Using Z39.50 to build cross-domain information servers. In *JCDL '01*, pages 52–53. ACM, 2001.

[8] A. L. Powell. *Database Selection in Distributed Information Retrieval: A Study of Multi-Collection Information Retrieval*. PhD thesis, University of Virginia, Virginia, 2001.