Model-Averaged Latent Semantic Indexing*

Miles Efron
School of Information
University of Texas, Austin
miles@ischool.utexas.edu

ABSTRACT

This poster introduces a novel approach to information retrieval that uses statistical model averaging to improve latent semantic indexing (LSI). Instead of choosing a single dimensionality k for LSI , we propose using several models of differing dimensionality to inform retrieval. To manage this ensemble we weight each model's contribution to an extent inversely proportional to its AIC (Akaike information criterion). Thus each model contributes proportionally to its expected Kullback-Leibler divergence from the distribution that generated the data. We present results on three standard IR test collections, demonstrating significant improvement over both the traditional vector space model and single-model LSI.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Experimentation, Performance, Theory

Keywords

latent semantic indexing, model selection, model averaging

1. INTRODUCTION

This poster reports efforts to improve latent semantic indexing (LSI) by statistical model averaging. LSI projects documents and queries onto a low-dimensional subspace of the observed vector space by use of the singular value decomposition (SVD) [4]. According to its proponents, LSI's dimensionality reduction improves retrieval by accounting for linguistic ambiguity. But how aggressively to truncate the SVD is an open research question [5, 7]. We propose an ensemble approach, using many models weighted by their expected estimated Kullback-Leibler (KL) divergence from the distribution that generated the data.

*The author thanks Chris H. Q. Ding for generous advice during early work on this paper.

SIGIR '07 Amsterdam
ACM 978-1-59593-597-7/07/0007.

2. APPROACH

Recent work [2, 6] suggests that no single choice of k, the dimensionality of an LSI model, will be optimal for all queries. With this in mind, our approach, model-averaged LSI (MALSI), uses a set of M LSI models. The intuition behind MALSI is that choosing any one model is risky. If we reduce dimensionality too far, we lose important information. On the other hand, if we retain too many dimensions, we risk overfitting the data, incurring the vocabulary mismatch problem. MALSI compensates for this risk by allowing many models to "vote" on the relevance of a document to a query, weighting each vote according to our confidence in that model.

To quantify our confidence in a model, let U_k be a k-dimensional LSI model; thus U_k contains the first k left singular vectors of the n term by p document matrix $\mathbf{X}$. We assess the fit of U_k by the Akaike information criterion (AIC), an estimate of the expected KL divergence between U_k and the unknown true model [3]:

$$AIC = -2\log\left(\mathcal{L}(U_k|\mathbf{X})\right) + 2D \tag{1}$$

where $D = rk + 1 - k(k-1)/2$ and r is the rank of $\mathbf{X}$.

Although LSI does not strictly define a generative model, work by Chris Ding [5] has derived the following log-likelihood function for an LSI model of k dimensions:

$$\log\mathcal{L}(U_k) = \lambda_1 + \ldots + \lambda_k - n\log Z(U_k) \tag{2}$$

where λ_k is the kth eigenvalue of $\mathbf{X}'\mathbf{X}$ and Z is a partition function

$$Z_k = \int \ldots \int \exp\left[(x\cdot u_1)^2 + \ldots + (x\cdot u_k)^2\right] dx^1 \ldots dx^p. \tag{3}$$

Given a set of candidate models M (we use all models between k_{min} and k_{max}) we find the p-vector of query-document similarities by

$$\hat{r}(\mathbf{q}) = \sum_{k\in M} w_k(\mathbf{q}'\mathbf{U}_k)(\mathbf{\Sigma}_k\mathbf{V}_k') \tag{4}$$

where w_k is the weight of the kth model (inversely proportional to its AIC), $\mathbf{\Sigma}_k$ is diagonal, containing the first k singular values of $\mathbf{X}$, and $\mathbf{V}_k$ holds the first k right singular vectors.

Figure 1 shows the log-likelihood and AIC values for all possible dimensionalities on three standard test collections.

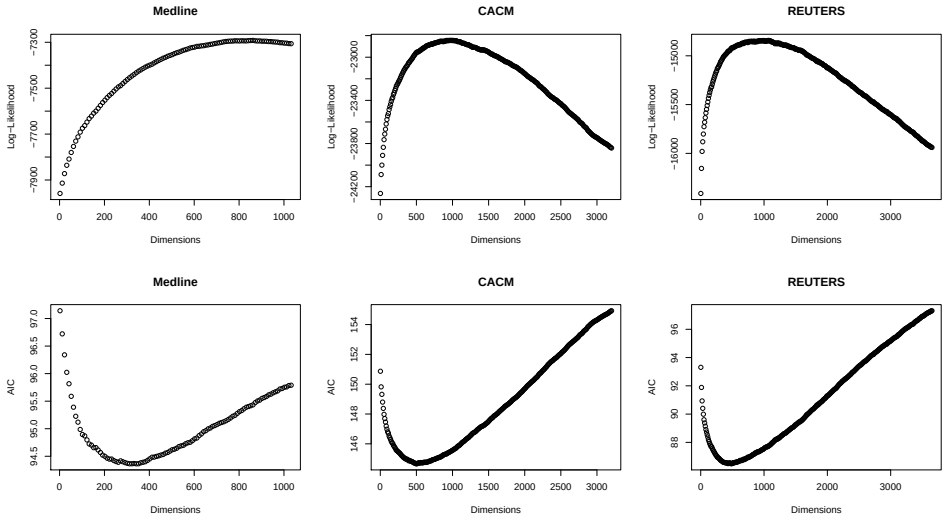

Figure 1: Log-Likelihood and AIC Values for LSI Models of Three Corpora

Table 1: Performance Averaged Over All Queries

	MED		CACM		REUT	
	MAP	R-Prec	MAP	R-Prec	MAP	R-Prec
VSM	0.466	0.455	0.158	0.155	0.486	0.490
LSI	0.483	0.467	0.143	0.146	0.516	0.509
MALSI	0.502	0.486	0.160	0.175	0.553	0.531

Our motivation for using AIC instead of the raw log-likelihood is evident from the different extrema that each function gives over the domain of candidate models. Due to its penalty for free parameters, AIC is optimized at a lower k than the log-likelihood; though more complex models may yield higher likelihood, AIC offers a better basis for model averaging [3].

3. EXPERIMENTS

We compared MALSI to the vector space model (VSM) and LSI (with single models chosen by minimum AIC) against three corpora: Medline (1033 documents), CACM (3204 documents) and a subset of Reuters-21578[1] [1]. Due to memory constraints, only the first 4000 Reuters documents were processed. Reuters queries were created by choosing `TOPIC` elements. Topics assigned to fewer than 10 documents were rejected, leaving 29. To gauge performance we used precision averaged across 11 levels of recall, and R-precision.

Table 1 shows each performance measure averaged over all queries. In all cases MALSI outperformed both LSI and VSM, even when LSI performed worse than the VSM.

To test the significance of these results, we conducted paired one-sided t-tests for each query (as opposed to the averaged results shown in Table 1). With respect to mean average precision (MAP) MALSI performed significantly better than LSI and VSM in most cases. The p-value of a test between MALSI and LSI was 0.058 for CACM, with $p \ll 0.01$ for the other corpora. MALSI decisively outperformed LSI on R-precision yielding p-values of 0.0007, 0.03, 0.09 for Medline, Reuters, and CACM, respectively.

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578

Several results from Table 1 are especially important. First, MALSI mitigated LSI's poor showing on CACM, supporting the hypothesis that model-averaging can lessen detrimental effects of dimensionality reduction. MALSI also improved overall accuracy, outperforming VSM and LSI at statistically significant levels ($p < 0.05$) for all cases except CACM.

4. CONCLUSION

Our use of AIC to assess model goodness of fit yielded promising results and puts MALSI on a strong information-theoretic footing. While this is sound and yields good results in future work we hope to compare the technique described here to a fully Bayesian approach, which admits a less restrictive notion of model uncertainty. Also, we plan to test MALSI on larger, more realistic corpora.

5. REFERENCES

[1] R. Baeza-Yates and B. Ribiero-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.

[2] H. Bast and D. Majumdar. Why spectral retrieval works. In *SIGIR '05*, pages 11–18, 2005.

[3] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference*. Springer, New York, 2nd edition, 2002.

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[5] C. H. Q. Ding. A similarity-based probability model for latent semantic indexing. In *SIGIR '99*, pages 58 – 65, 1999.

[6] G. Dupret. Latent concepts and the number of orthogonal factors in latent semantic analysis. In *SIGIR '03*, pages 221 – 226, 2003.

[7] M. Efron. Eigenvalue-based model selection during latent semantic indexing. *JASIST*, 56(9):969–988, 2005.