Abstract

MILES EFRON: Eigenvalue-based Estimators for Optimal Dimensionality Reduction in Information Retrieval

(Under the direction of Gregory B. Newby)

Latent Semantic Indexing (LSI) extends Salton's vector space model (VSM) of information retrieval, using dimensionality reduction to construct a statistical model of the relationships among the terms in a document collection. Though empirical studies have shown that such statistical models can improve retrieval over traditional key word-based approaches, dimensionality reduction raises an important question: if we are to reduce model dimensionality, how aggressively should we do so? Or conversely, what is the optimal value for k, the number of dimensions in an LSI system? In the unsupervised learning environment native to information retrieval, notions of model optimality and goodness of fit are difficult to define.

This dissertation pursues the viability of five statistical methods for estimating the optimal dimensionality of LSI systems. Though the five pursued methods entail different theoretical assumptions, they are all predicated on an analysis of the eigenvalues that arise naturally during LSI. This thesis contends that LSI's relation to principal component analysis makes an analysis of the eigenvalues the natural vehicle for dimensionality estimation.

To judge the utility of eigenvalue-based estimators for dimensionality estimation under LSI, two groups of experiments were performed. Both rounds of experimentation tested the accuracy of five dimensionality estimation criteria: eigenvalue-one, parallel analysis, percent-of-variance, Bartlett's test of isotrapy, and a novel method dubbed amended parallel analysis. During the first experiments, these criteria were applied to six standard information retrieval test collections and evaluated with regard to their accuracy. The second round of experiments applied each dimensionality estimator to simulated data.

The first round of experiments suggested that the family of estimation techniques based on the notion that dimensionality reduction is merited to the extent that the indexing features violate the VSM's assumption of statistical independence were especially accurate. Also during this round, amended parallel analysis yielded statistically significant improvements over traditional parallel analysis. The data simulations supported these findings. Applying parallel analysis and amended parallel analysis to simulated data of known intrinsic dimensionality yielded categorically superior estimates to all other tested estimation techniques.