TEMPLE UNIVERSITY
Department of Mathematics

Applied Mathematics and Scientific Computing Seminar

Room 617 Wachman Hall

Wednesday, 19 October 2005, 4:00 p.m.

Data Mining and Latent Semantic Indexing

by Judith Vogel
Richard Stockton College

Abstract. With the advent of the internet and the need for more powerful search engines, data mining and information retrieval has become an important topic of research in the field of Numerical Linear Algebra. The shortcomings of standard lexical searches become evident when users wish to retrieve documents based on conceptual content and not on word/document matches. This fundamental problem has shifted focus to vector based retrieval methods. Latent Semantic Indexing (LSI) overcomes the deficiencies of term-matching retrieval by treating the unreliability of term matching as a statistical problem. By assuming that there is an underlying latent semantic structure to the data, document/term matrices reflect the associated content without directly representing matching queries. Dimension reduction is imperative for efficiently manipulating the massive quantity of data and for eliminating the noise associated with standard lexical searches. To be useful, this lower-dimensional representation must be a good approximation of the original matrix representation of the document/term set. This talk discusses some of the techniques used to achieve appropriate dimension reduction in search analysis including Singular Value Decomposition, Clustering, and Term Weighting.