Reading Report - Week 9 **Varad Meru**
CS 221 - Information Retrieval Student # 26648958
Prof. Cristina Lopes Due Date: 03/07/2015

Indexing by latent semantic analysis.* - Summary

By Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman.

Summary

The authors start with the explanation of automatic systems for indexing and information retrieval. They introduce the problem of the present systems, such as *synonymy* and *polysemy*, and how reformulating the problem into a statistical problem helps applying known techniques such as *Latent Semantic Indexing*. They used singular-value decomposing to apply LSI analysis on a semantic space of a large term-document association matrix. The failure can be attributed to three factors - firstly, incomplete indexing; secondly, lack of adequate automatic method for dealing with polysemy; and thirdly, understanding the word type and attributes attached to it.

The discussion then heads to the explanation behind the use of latent semantic indexing. The illustration of the problems of retrieval start with an example of a fictional matrix of terms by documents, along with a fictional query. The illustration shows the *Relevant* and *Matching* case. In cases where the synonymy of the words are taken into account, it helps to actually find what the user "really" wanted to know, by considering the synonyms. The discussion then moves towards the choice of method to uncover the Latent Semantic Structure of the data. The goal is to find and fit a model which is useful in terms of modeling relationships between terms and documents. The model revolved around the idea of semantic similarity between documents and terms. This notion of proximity, i.e., putting similar items near each other in some space or structure, helps in finding the latent structure. The criteria for considering an alternative model were: (1) Adjustable representational richness,(2) Explicit representation of both terms and documents, and, (3) Computational tractability for large datasets. Only the *two-mode factor analysis* could satisfy the previously written criteria. It is a generalization of the familiar factor analytic model based on singular value decomposition (SVD).

The discussion dives deeper into singular value decomposition (SVD) or *two-mode factor analysis* which is used on a matrix of terms by documents to derive our particular latent semantic structure model. It is closely related to a number of mathematical and statistical techniques in a wide variety of other fields, including eigenvector decomposition, spectral analysis, and factor analysis. This matrix is again decomposed into three other matrices of a special nature, with the process called as "singular-value-decomposition". The technical details are given in Equation 1 and the description can be seen in 1(a) and 1(b).

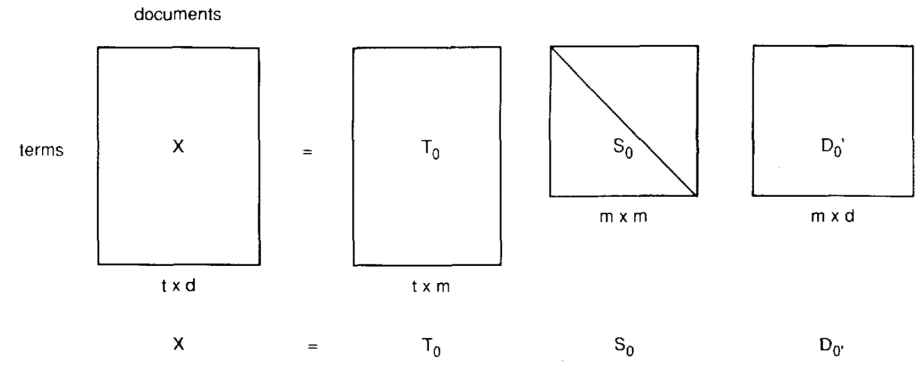
$$X = T_0 S_0 D_0{}' \tag{1}$$

X represents the $t \times d$ matrix and T_0 and $D_0{}'$ have orthonormal columns, along with S_0 is the diagonal matrix with the diagonal values in decreasing order.

Commentary

The paper gave me a good introduction on using latent information for finding structure in the information and using it in building a topic based indexing system. This can be using along with the unsupervised mechanisms from the machine learning domain.

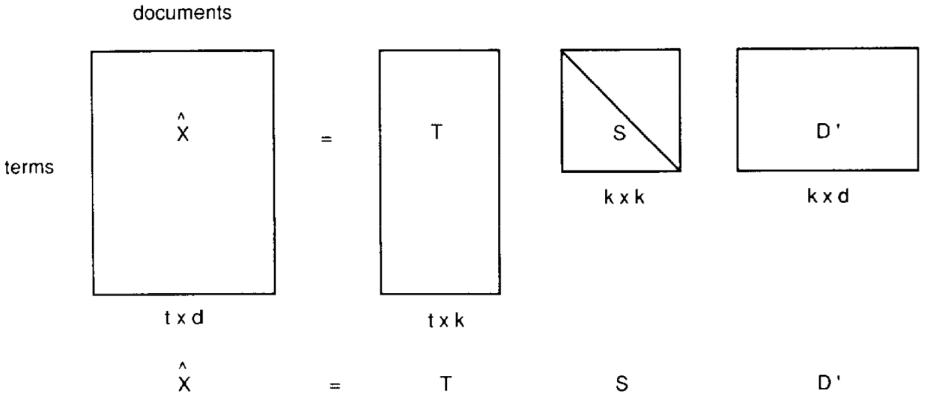
*Deerwester, Scott C., et al. *Indexing by latent semantic analysis.*. JAsIs 41.6 (1990): 391-407.



Singular value decomposition of the term x document matrix, X. Where:

T_0 has orthogonal, unit-length columns ($T_0' T_0 = I$)
D_0 has orthogonal, unit-length columns ($D_0' D_0 = I$)
S_0 is the diagonal matrix of singular values
t is the number of rows of X
d is the number of columns of X
m is the rank of X ($\leq$ min(t,d))

(a) Schematic of the Singular Value Decomposition (SVD) of a rectangular term by document matrix. The original term by document matrix is decomposed into three matrices each with linearly independent components.



Reduced singular value decomposition of the term x document matrix, X. Where:

T has orthogonal, unit-length columns (T' T = I)
D has orthogonal, unit-length columns (D' D = I)
S is the diagonal matrix of singular values
t is the number of rows of X
d is the number of columns of X
m is the rank of X ($\leq$ min(t,d))
k is the chosen number of dimensions in the reduced model ($k \leq m$)

(b) Schematic of the *reduced* Singular Value Decomposition (SVD) of a rectangular term by document matrix. The original term by document matrix is *approximated* using the k largest singular values and their corresponding singular vectors.