Latent Semantic Indexing is an Optimal Special Case of Multidimensional Scaling

Brian T. Bartell Garrison W. Cottrell Richard K. Belew

Department of Computer Science & Engineering-0114
University of California, San Diego
La Jolla, California 92093-0114

Abstract

Latent Semantic Indexing (LSI) is a technique for representing documents, queries, and terms as vectors in a multidimensional real-valued space. The representations are approximations to the original term space encoding, and are found using the matrix technique of Singular Value Decomposition. In comparison, Multidimensional Scaling (MDS) is a class of data analysis techniques for representing data points as points in a multidimensional real-valued space. The objects are represented so that inter-point similarities in the space match inter-object similarity information provided by the researcher. We illustrate how the document representations given by LSI are equivalent to the optimal representations found when solving a particular MDS problem in which the given inter-object similarity information is provided by the inner product similarities between the documents themselves. We further analyze a more general MDS problem in which the inter-document similarity information, although still in inner product form, is arbitrary with respect to the vector space encoding of the documents.

1 Introduction

There is currently a great deal of interest in automatic document indexing schemes which are not based simply on the matching of keywords in the documents. This is partly motivated by the observation [5] that individual keywords are not adequate discriminators of semantic content. Rather, the indexing relationship between word and document content is many-to-many: A number of concepts can be indexed by a single term, and a number of terms can index a single concept. When retrieval is based solely on the matching of terms between the query and documents, performance suffers as some relevant documents are missed (they are not indexed by the keywords used in the query, but by synonyms) and some irrelevant documents are retrieved (they are indexed by unintended senses of the keywords in the query).

Latent Semantic Indexing (LSI) [2] is a particular approach aimed at addressing these limitations. This technique maps each document from a vector space representation based on keyword frequency [8], to a vector in a lower dimensional space. Terms are also mapped to vectors in the reduced space. The claim is that the similarity between vectors in the reduced space, for example using the cosine similarity measure, may be a better retrieval indicator than similarity measured in the original term space. This is primarily because, in the reduced space, two related documents may be represented similarly even though they do not share any keywords. This may occur, for example, if the keywords used in each of the documents co-occur frequently in other documents.

Multidimensional Scaling (MDS) techniques have also received attention in the information retrieval literature. Most generally, MDS is a class of algorithms for analyzing inter-object similarity information. Given some measure of the similarity between pairs of objects in a domain, MDS represents the objects as points in a multidimensional real-valued space, such that the inter-point distances correspond well with the imposed similarity information. In the work of Everett and Pecotich [4], for example, MDS is used to visualize the interrelationships among journals by using citation frequency as a measure of journal "similarity".

The main goal of this paper is to illustrate the commonality between these two techniques. We present a

formal equivalence between the document representations found using Latent Semantic Indexing and the optimal representations found when solving a particular Multidimensional Scaling problem. This equivalence shows that the LSI document representations are optimal with respect to this MDS problem. This analysis is useful because

- it offers insights into both the strengths and weaknesses of Latent Semantic Indexing;
- it integrates LSI into the extensive literature on MDS; and
- it suggests directions for expansion of the technique.

Section 2 and section 3 outline the basic techniques which we will compare: Latent Semantic Indexing and Multidimensional Scaling. Section 4 presents the result that representations found by the two techniques are equivalent, and section 5 discusses a more general scaling formulation. Section 6 addresses the value of this analysis, and points to directions for future research.

2 Latent Semantic Indexing and the SVD

Latent Semantic Indexing begins with a vector space representation of documents [8], and attempts to improve retrieval performance by re-representing both documents and terms in a new vector space with smaller dimension. We will emphasize the re-representation of documents here, although corresponding arguments can be made for terms as well. According to Deerwester, et. al. [2], this reduced document representation has the advantages that

- the dimensions in the space are uncorrelated (i.e. they are orthogonal),
- the representations are less noisy, and
- the representations incorporate higher-order (latent) association structure among terms and documents.

These properties are a result of the technique used to re-represent the documents in the lower dimensional space. This technique is Singular Value Decomposition (SVD). We take care here to define certain properties of the SVD, as they will be useful in the remainder of the paper. SVD is a technique for uniquely[1] decomposing an arbitrary matrix $\mathbf{X}$ as the product of three matrices:

$$\mathbf{X} = \mathbf{ULA}^T \tag{1}$$

($\mathbf{A}^T$ denotes the transpose of $\mathbf{A}$). The three matrices have a special restricted form. $\mathbf{U}$ and $\mathbf{A}$ are both column orthonormal; that is, the columns are orthogonal (i.e. the inner product of two different columns is 0) and are unit length. Thus, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, and $\mathbf{UU}^T = \mathbf{P}$, where $\mathbf{P}$ is a projection matrix on to the space spanned by the columns of $\mathbf{X}$. $\mathbf{L}$ is a diagonal matrix of *singular values*: all off-diagonal elements are zero; all diagonal elements (the singular values) are non-negative real values, typically ordered in decreasing value. When $\mathbf{X}$ has t rows and d columns (denoted $t \times d$) and is rank r, then we allow $\mathbf{U}$ to be $t \times r$, $\mathbf{A}$ to be $d \times r$, and $\mathbf{L}$ to be $r \times r$ with no zero singular values.

A useful property of SVD is that it provides the best lower rank approximation of a matrix $\mathbf{X}$ in terms of the Euclidean matrix norm (or Frobenius norm, calculated by taking the square root of the sum of all squared entries of a matrix) [9]. More precisely, let $\mathbf{U}_k$ be the $t \times k$ ($k \leq r$) matrix found by removing $r - k$ columns from $\mathbf{U}$. The k columns remaining in $\mathbf{U}_k$ correspond to the largest singular values in $\mathbf{L}$ (similar versions of $\mathbf{L}_k$ and $\mathbf{A}_k$ can be defined). Then $\hat{\mathbf{X}} = \mathbf{U}_k\mathbf{L}_k\mathbf{A}_k^T$ minimizes $\|\hat{\mathbf{X}} - \mathbf{X}\|_F$ over all rank-k $\hat{\mathbf{X}}$, where $\|\cdot\|_F$ denotes the Frobenius norm.

Returning now to the topic of Latent Semantic Indexing, LSI uses SVD to derive the reduced dimension representation of the documents. If $\mathbf{X}$ ($t \times d$) is a matrix of d documents represented using t terms, and $\mathbf{X} = \mathbf{ULA}^T$ is the singular value decomposition of $\mathbf{X}$, then row i of $\mathbf{A}_k\mathbf{L}_k$ gives the representation of document i in k-space. These re-representations of the documents are used in place of the original t-space representations when making similarity judgments. If queries are represented as vectors in the original term space, then queries can also be mapped into this new k dimensional document space. Let the query be encoded as a row vector $\mathbf{q}$ in $\Re^t$. Then the query in k-space would be $\mathbf{qU}_k$. Thus, the similarity of each document to a query $\mathbf{q}$ is found by measuring the similarity between rows of $\mathbf{A}_k\mathbf{L}_k$ and the vector $\mathbf{qU}_k$. The similarity can be measured using, for example, the cosine measure, as was done in the original work [2].

3 Multidimensional Scaling

Multidimensional Scaling (MDS) [1] is a general class of data analysis, data reduction, and modeling techniques. MDS is used to find representations in $\Re^k$ of objects, such that the similarities between the objects

[1] Unique up to certain trivial re-arrangements of columns, and sub-space rotations in the case of duplicated singular values.

in the k-space correspond to known inter-object similarity information. The similarity information can come from a variety of sources: in psychology, MDS is often used to model human judgments concerning the similarity between pairs of concepts; in data reduction, the similarity information can be derived directly from the data, and MDS is used to find a lower dimension representation which best preserves that structure.

In applying MDS in a particular domain, the researcher must specify a number of "free variables" in the technique. For example, the similarity function over $\Re^k$ must be specified. This can be the simple inner product metric or euclidean distance metric[2], or an asymmetric or parameterized measure, depending on the requirements on the model. Furthermore, the similarity information can be metric (indicating the exact target similarities for the configuration of points), or non-metric (indicating only the relative ordering of inter-object similarities [7]). Related to the type of similarity information is the choice of fitness measure used to evaluate how well a particular configuration of points corresponds to the similarity information. Also, the mapping from an object to a point in $\Re^k$ may be arbitrary or functionally constrained. When functionally constrained, MDS finds a mapping (out of a restricted set of mappings) from object to point. For example, one might constrain the problem so as to only allow linear re-combinations of the original data points in determining the reduced representations.

As an example of MDS, consider a data matrix $\mathbf{X}$ $(t \times d)$ of d observations on t variables. The goal may be to represent the observations as points in $\Re^k$, $k \leq t$, such that the euclidean distances between the points best matches the original euclidean distances in $\Re^t$, in terms of least sum-of-squares error. This is a form of dimensionality reduction. We define the configuration error (measuring the disagreement between the similarity structure in $\Re^k$ and the structure in $\Re^t$) as

$$E_{\mathbf{X}}(f) = \sum_{i=1}^{d}\sum_{j=1}^{d}[\delta_k(f(\mathbf{x}_i), f(\mathbf{x}_j)) - \delta_t(\mathbf{x}_i, \mathbf{x}_j)]^2 \quad (2)$$

where $\delta_n()$ is the euclidean distance metric in $\Re^n$, $\mathbf{x}_i$ is the i'th column of $\mathbf{X}$, and $f()$ maps observations represented in $\Re^t$ to points in $\Re^k$. MDS attempts to find an f (which specifies the configuration of points in $\Re^k$) which minimizes $E_{\mathbf{X}}(f)$.

4 Equivalence

We now demonstrate that the document representations found using Latent Semantic Indexing are equivalent to the optimal representations for a Multidimensional Scaling problem in which the similarity information is derived from the data using the inner product similarity measure. We proceed by deriving the optimal solution to the MDS problem. This solution is well known in the Multidimensional Scaling literature (e.g., [10]). We then demonstrate its correspondence with Latent Semantic Indexing.

4.1 Scaling of Data-Derived Inner Product Similarities

Let $\mathbf{X}(t \times d)$ be a matrix of d objects represented in $\Re^t$. The similarity $s_{i,j}$ between a pair of objects i, j (represented by columns $\mathbf{x}_i$ and $\mathbf{x}_j$ of $\mathbf{X}$) using the inner product similarity measure is then $s_{i,j} = \mathbf{x}_i^T\mathbf{x}_j = \sum_{k=1}^{t}\mathbf{x}_{k,i}\mathbf{x}_{k,j}$. The matrix $\mathbf{S}(d \times d)$ of all pairwise similarities is given by $\mathbf{S} = \mathbf{X}^T\mathbf{X}$.

Assume we are trying to solve a Multidimensional Scaling problem in which $\mathbf{X}$ is our set of observations, and $\mathbf{S} = \mathbf{X}^T\mathbf{X}$ is a matrix of desired inter-observation similarity information. The problem is to find a mapping from $\mathbf{X}$ to a representation of the observations in a reduced space $\Re^k$, such that the inner product similarities between vectors in this space best match the entries in $\mathbf{S}$.

First consider only the optimal $\Re^k$ representation of the objects, ignoring the mapping from input representation to $\Re^k$. Let the matrix $\mathbf{H}$ $(k \times d)$ be a candidate optimal representation; i.e., the columns of $\mathbf{H}$ are d object vectors in $\Re^k$. To solve the representation problem, we seek the $\hat{\mathbf{H}}$ which minimizes

$$E(\mathbf{H}) = \|\mathbf{H}^T\mathbf{H} - \mathbf{S}\|_F^2 \quad (3)$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm. This problem permits a simple analytic solution. We first note that the SVD of $\mathbf{S} = \mathbf{X}^T\mathbf{X}$ is $\mathbf{A}\mathbf{L}^2\mathbf{A}^T$. This follows directly from $\mathbf{X}$ having SVD $\mathbf{U}\mathbf{L}\mathbf{A}^T$ and from general uniqueness properties of SVD. Furthermore, the best rank-k approximation of $\mathbf{S}$ in Frobenius norm is $\hat{\mathbf{S}} = \mathbf{A}_k\mathbf{L}_k^2\mathbf{A}_k^T$, as was stated in section 2. Observe that $\hat{\mathbf{S}}$ can be trivially factored as $\hat{\mathbf{S}} = (\mathbf{L}_k\mathbf{A}_k^T)^T(\mathbf{L}_k\mathbf{A}_k^T)$. Thus, representing the data $\mathbf{X}$ as $\hat{\mathbf{H}} = \mathbf{L}_k\mathbf{A}_k^T$ should provide an optimal k-space preservation of $\mathbf{S}$.

We now return to the problem of finding the best mapping from the data to this optimal representation. If we assume the mapping is linear, we need only solve the least squares problem

$$\mathbf{W}\mathbf{X} = \hat{\mathbf{H}} = \mathbf{L}_k\mathbf{A}_k^T \quad (4)$$

where $\mathbf{W}$ is a $(k \times t)$ matrix mapping $\mathbf{X}$ to the optimal k-space representations. One solution is $\hat{\mathbf{W}} = \mathbf{U}_k^T$.

[2] We are suppressing the issue of transforming "similarities" into "distances", and vice-versa.

The matrix of inner product similarities for the data represented in k-space is given by $(\hat{\mathbf{W}}\mathbf{X})^T\hat{\mathbf{W}}\mathbf{X}$,

$$\begin{aligned}(\hat{\mathbf{W}}\mathbf{X})^T\hat{\mathbf{W}}\mathbf{X} &= (\mathbf{U}_k^T\mathbf{X})^T(\mathbf{U}_k^T\mathbf{X}) \\ &= \mathbf{A}\mathbf{L}\mathbf{U}^T\mathbf{U}_k\mathbf{U}_k^T\mathbf{U}\mathbf{L}\mathbf{A}^T \\ &= \mathbf{A}\mathbf{L}_k^2\mathbf{A}^T \\ &= \hat{\mathbf{S}}\end{aligned} \tag{5}$$

As $\hat{\mathbf{S}}$ is known to be the best rank-k approximation of $\mathbf{S}$, we find that this linear mapping is optimal.

Thus, SVD tells us how to decompose the symmetric similarity matrix $\mathbf{S}$ into the product of the transpose of a real matrix and itself. This is a well known procedure in the scaling of inner product similarities [6, p. 209] [1, pp. 270–291]. In the current special case, with $\mathbf{S}$ restricted to $\mathbf{X}^T\mathbf{X}$, we get a particularly simple linear relationship between the original data and its optimal re-representation in k dimensions.

4.2 Latent Semantic Indexing Performs Optimal Scaling

We have seen how to best re-represent a set of vectors $\mathbf{X}$ in a lower dimensional space such that the original inner product similarities are best preserved by the inner product similarities in the new space. In the context of Information Retrieval, we can let $\mathbf{X}$ be a matrix of d documents represented over t terms. Then, a reduced representation of the documents which best preserves the original inner product similarities is $\mathbf{H} = \mathbf{U}_k^T\mathbf{X}$. However, this is the equivalent document representation we would get if we performed Latent Semantic Indexing on these documents (as was presented in section 2). Thus, the optimal representation given by the Multidimensional Scaling problem is equivalent to the representation given by LSI.

This demonstrates that the document representations found using LSI are optimal with respect to this particular scaling problem. That is, the inner product similarities between the documents in the original space are optimally preserved by the inner products between corresponding vectors in the reduced space. However, this is a special kind of optimality. If we are to delight in this result, we must trust that inner product similarities in the original t-space are actually the right quantities to preserve in order to achieve good performance on novel queries. In the case that inner product in $\Re^t$ is a good measure of document relatedness, this analysis supports that this measure has been optimally preserved. This further suggests that the inner product measure might be most appropriate in the reduced space, rather than some other measure. On the other hand, if relatedness is not well predicted by inner products in the t-space, this analysis suggests that the reduced representations will reflect this. Of course, a similarity measure other than inner product can be used in the reduced space in this case, and our analysis cannot predict the performance for an arbitrary similarity measure. More specifically, the equivalence illustrated here is only known to hold for inner product similarities; other measures are not known by the authors to be optimally preserved by the reduction.

We will examine these issues in more detail later (see the Discussion, section 6.1). In the next section, we will pursue an alternative implication of this equivalence result. Our approach is to generalize Latent Semantic Indexing using this link to Multidimensional Scaling. This generalization will allow for more arbitrary similarity information than the simple inner products between the t-space document representations.

5 Scaling of Arbitrary Inner Product Similarities

The previous section described an exact analytic solution for a very special case of MDS, in which the inner product similarity information $\mathbf{S}$ is derived directly from the data as $\mathbf{S} = \mathbf{X}^T\mathbf{X}$. We now look at a more general case, in which $\mathbf{S}$ represents arbitrary inner product similarity information. That is, we assume $\mathbf{S} = \mathbf{B}^T\mathbf{B}$ for some real $\mathbf{B}$ not necessarily related to $\mathbf{X}$.[3]

We consider this to be an interesting generalization of the previous restricted case. Whereas before the desired similarity information was derived trivially from the data, the current generalization allows other sources of similarity information to be modeled in the re-representation of documents. Sources might include co-citation information or relevance feedback. We will return to the idea of alternative sources of information in the discussion (section 6).

The problem is to find the best linear mapping $\mathbf{W}$ from the d vectors in $\Re^t$ given by $\mathbf{X}$, to d new vectors in $\Re^k$ given by $\mathbf{H} = \mathbf{W}\mathbf{X}$, such that $\mathbf{S}$ is preserved in Frobenius norm. Let $\hat{\mathbf{S}} = \mathbf{H}^T\mathbf{H} = (\mathbf{W}\mathbf{X})^T(\mathbf{W}\mathbf{X})$. The MDS problem is to find $\hat{\mathbf{W}}$ which minimizes the configuration error

$$E(\mathbf{W}) = \|\hat{\mathbf{S}} - \mathbf{S}\|_F^2 \tag{6}$$

In section 4.1, a similar problem was solved by using Singular Value Decomposition to find a matrix $\mathbf{G}$ which factors $\mathbf{S}$ into $\mathbf{S} = \mathbf{G}^T\mathbf{G}$. The k largest singular vectors of $\mathbf{G}$ provided the optimal representation of the data in k-space to preserve $\mathbf{S}$. Finding $\mathbf{W}$ such that $\mathbf{W}\mathbf{X} = \mathbf{G}_k$ in this special case was particularly easy.

[3] Borg & Lingoes [1, pp. 292–295] discuss the necessary conditions (i.e. positive semi-definiteness) for an arbitrary symmetric matrix to be decomposed as $\mathbf{B}^T\mathbf{B}$, for real $\mathbf{B}$.

Unfortunately, this procedure does not provide an optimal solution to the current more general problem. Although $\mathbf{G}_k$ is the best possible k-space representation to preserve an arbitrary $\mathbf{S}$, the least squares problem $\mathbf{WX} = \mathbf{G}_k$ may not be perfectly solvable. That is, there may be no linear mapping ($\mathbf{W}$) from the original data ($\mathbf{X}$) to their optimal k-space representation ($\mathbf{G}_k$). This will be the case when the row space of $\mathbf{G}_k$ is not entirely a subspace of the row space of $\mathbf{X}$.

One option is to find the best least squares solution to $\mathbf{WX} = \mathbf{G}_k$. However, there may actually exist better minima to equation (6) than this! In particular, if r is the rank of $\mathbf{G}$, there may be a better minimum to equation (6) if the row space of the $r-k$ smallest singular vectors of $\mathbf{G}$ intersects the row space of $\mathbf{X}$. In this case, the eigenvectors which correspond to the smaller eigenvalues have information which correlates with the target similarity information. As an example, consider an extreme case: the row spaces of $\mathbf{G}_k$ and $\mathbf{X}$ are orthogonal, but the row spaces of $\mathbf{G}_{r-k}$ and $\mathbf{X}$ are identical (here, $\mathbf{G}_{r-k}$ denotes the $(r-k)\times d$ matrix constructed from the $r-k$ smallest singular vectors and values of $\mathbf{G}$). In this case, the least squares solution to $\mathbf{WX} = \mathbf{G}_k$ is $\hat{\mathbf{W}} = \mathbf{0}$. However, there is an exact solution to $\mathbf{WX} = \mathbf{G}_{r-k}$ which provides a better solution to (6).

Solving this MDS problem therefore requires an alternate approach. Consider again the configuration error, $E(\mathbf{W})$:

$$\begin{aligned} E(\mathbf{W}) &= \|\hat{\mathbf{S}} - \mathbf{S}\|_F^2 \\ &= Tr\{(\hat{\mathbf{S}} - \mathbf{S})(\hat{\mathbf{S}} - \mathbf{S})^T\} \\ &= Tr\{(\mathbf{WX})^T\mathbf{WX}(\mathbf{WX})^T\mathbf{WX} \\ &\qquad -2\mathbf{S}(\mathbf{WX})^T\mathbf{WX} + \mathbf{SS}\} \end{aligned} \tag{7}$$

where $Tr\{\mathbf{M}\}$ denotes the sum of the diagonal entries of $\mathbf{M}$, and $Tr\{\mathbf{MM}^T\}$ is equivalent by definition to the squared Frobenius norm $\|\mathbf{M}\|_F^2$. Calculating $\partial E(\mathbf{W})/\partial\mathbf{W}$ and setting this partial to 0, we find that any minimum $\hat{\mathbf{W}}$ of $E(\mathbf{W})$ must satisfy

$$\mathbf{WX}(\mathbf{WX})^T\mathbf{WXX}^T = \mathbf{WXSX}^T \tag{8}$$

We prove in the appendix that one class of solutions to equation (8) is of the form

$$\hat{\mathbf{W}} = \mathbf{RM}_k^T\mathbf{CX}^+ \tag{9}$$

Here, $\mathbf{C}$ is any real matrix such that $\mathbf{C}^T\mathbf{C} = \mathbf{S}$. $\mathbf{X}^+$ is the pseudo-inverse of $\mathbf{X}$. $\mathbf{R}$ is any full-rank $k \times k$ rotation matrix (i.e. $\mathbf{R}^T\mathbf{R} = \mathbf{I}$). $\mathbf{M}_k$ requires a more lengthy specification: let $\mathbf{P}^{\mathbf{X}} = \mathbf{AA}^T$ be a projection matrix on to the row space of $\mathbf{X}$ (recall, the SVD of $\mathbf{X} = \mathbf{ULA}^T$). Then $\mathbf{M}_k$ is provided by the Singular Value Decomposition of $\mathbf{CP}^{\mathbf{X}} = \mathbf{M\Sigma N}^T$ (where $\mathbf{M}$ and $\mathbf{N}$ are column orthonormal, and $\mathbf{\Sigma}$ is a diagonal matrix of singular values). Note that $\mathbf{M}$ is a rotation matrix (it has orthonormal columns), and $\mathbf{M}_k$ is a rotation on to a rank-k subspace. Although equation (9) specifies a class of solutions, a particular solution can easily be chosen by letting $\mathbf{R} = \mathbf{I}$ and $\mathbf{C} = \mathbf{G}$, where $\mathbf{G}$ is the matrix described above found by Singular Value Decomposition of $\mathbf{S}$.

The solution $\hat{\mathbf{W}} = \mathbf{M}_k^T\mathbf{GX}^+$ is a local minimum of the configuration error (equation 6), but it is unknown whether this solution is necessarily globally optimal. That is, there may be other solutions to the equality in equation (8) which result in a lower configuration error. Towards this concern, we have performed experiments in which $E(\mathbf{W})$ in (6) is minimized using numerical techniques (conjugate gradient) for random matrices $\mathbf{S}$ and $\mathbf{X}$. No cases have been found having a better linear solution than that given by $\hat{\mathbf{W}} = \mathbf{RM}_k^T\mathbf{CX}^+$ (where a specific solution is instantiated from this set of solutions in the manner described in section 5). This provides some evidence that the proposed solution is globally optimal in many cases.

It should be emphasized that finding the best *linear* mapping from data to k-space representation places a strong constraint on the quality of the re-representation. That is, the re-representation of the data by $\hat{\mathbf{H}} = \hat{\mathbf{W}}\mathbf{X}$, where $\hat{\mathbf{W}} = \mathbf{M}_k^T\mathbf{GX}^+$, is not generally equivalent to $\mathbf{G}_k$, the k-space representations which best preserve $\mathbf{S}$. This is why we could not simply solve the least squares problem $\mathbf{WX} = \mathbf{G}_k$. Thus, it often will be possible to find some other function f which maps data objects to k-space representations such that the re-representations are closer to the optimal ones (the columns of $\mathbf{G}_k$). For example, let $(\mathbf{G}_k)_i$ denote the i'th column of $\mathbf{G}_k$. It may be possible to find an f such that $f(\mathbf{x}_i) = (\mathbf{G}_k)_i$ for all data objects i. If there is no linear solution such that $\mathbf{WX} = \mathbf{G}_k$, then this *nonlinear* function f would provide representations which preserve the inner product similarity structure better than the best linear solution.

6 Discussion

Section 4 demonstrated the equivalence between representations found with LSI and those which optimize a particular MDS problem. This correspondence was generalized in section 5 to permit alternate sources of similarity information other than inner products of the original document vectors. We now discuss some implications of these results.

6.1 Implications for Latent Semantic Indexing

This analysis is intended to have a complementary relationship with previous work on Latent Semantic Indexing. Specifically, previous work [2] has illustrated beneficial properties of the technique (such as reduction of noise, orthogonalization of the vector space, and incorporation of associational relationships in the representation), and the equivalence with Multidimensional Scaling does not detract from these results. Rather, the current analysis adds new terminology and an alternative perspective to the discussion.

We have found that Latent Semantic Indexing finds an optimal representation of documents with respect to a particular Multidimensional Scaling problem. This optimality illustrates both strengths and weaknesses of the technique, and suggests directions for enhancement of the method.

One major insight, given by this analysis, is the importance of the inner product similarities in the original term space. When these similarities are a good, though perhaps noisy, estimate of the relatedness between documents, the LSI reduction should yield good results. In this case, our analysis suggests that inner product is a well motivated metric to use in the reduced space.

The analysis further suggests that certain term weightings in the original term space (which is permitted in LSI [2]) should improve the performance of the technique. That is, weighting the elements of the vector representation, to incorporate the frequency of each term throughout the documents for example, may improve the retrieval performance using the inner product measure. Essentially, these modifications yield a new document matrix $\bar{\mathbf{X}}$ with its own similarity structure $\bar{\mathbf{S}} = \bar{\mathbf{X}}^T\bar{\mathbf{X}}$ based on inner products. When $\bar{\mathbf{S}}$ is a better measure of relatedness than $\mathbf{S}$, we may anticipate that LSI will perform better when operating on $\bar{\mathbf{X}}$ than on $\mathbf{X}$. Dumais' empirical study of LSI using various term weightings seems to agree with this suggestion [3]: Term weightings which tend to improve inner product retrieval in the original term space also tend to improve retrieval performance in the reduced space.

In past applications of Latent Semantic Indexing, similarity in the reduced space has been measured by the cosine measure rather than by inner product. Our result does not appear to generalize when cosine is used: We have not shown that cosines in the original term space are preserved in terms of cosines in the reduced space. We certainly have not shown that cosine is an inappropriate measure. Rather, we have shown that inner product is perhaps a more natural measure, when it is desirable to preserve the inner products from the original space. Furthermore, it is possible to preserve cosine similarities in the original space in terms of inner products in the reduced space. This is achieved by length-normalizing the documents first in the term space (resulting in a new document matrix $\bar{\mathbf{X}}$). Since cosines and inner products are identical with unit length vectors, cosines are preserved as inner products by LSI.

Just as our analysis did not generalize to prove that cosine similarities are best preserved by cosines in the reduced space, we also cannot comment on the effectiveness of other similarity measures in the reduced space. The analysis does not preclude their use, and empirical work may demonstrate other measures to be more useful. Rather, the analysis indicates that these alternate measures will be applied to a reduced representation which has much in common with the original space, in terms of their inner product similarity structures.

6.2 Directions for Future Work

In terms of applications to Information Retrieval, our generalization of the scaling problem to more arbitrary similarity matrices begs the question: where will this similarity information come from? That is, we allow for similarity information which is not derived directly from the inner products between the original documents. What are some possible sources for this similarity information?

It may be possible to start with similarities given by the inner products between documents as an initial estimate, but then to modify the scores with any available information. Sources might include, for example, document co-citations and relevance feedback from users of the system. We have not yet performed empirical studies in this direction. It is certainly a provocative course for future study.

A technical concern for future study is whether $\hat{\mathbf{W}}$, given by equation (9), provides the optimal linear solution for the general MDS problem. Our analysis has provided one class of solutions, but others may exist which both satisfy equation (8) and which result in a lower configuration error. Although experiments have been performed indicating that our solution is robust for a set of random matrices (as discussed previously), further analysis is nevertheless required.

7 Conclusion

We have demonstrated that the document representations found by Latent Semantic Indexing are equivalent to the representations found in solving a very restricted Multidimensional Scaling problem, in which the target similarity information is simply the inner products between documents. In this sense, LSI is calculating an optimal scaling solution. The analysis is ex-

tended to allow for more arbitrary similarity information. This extension illustrates how Multidimensional Scaling provides a complementary theoretical framework for Latent Semantic Indexing, in which possible enhancements to the method can be explored.

Acknowledgements: The authors thank James Bunch, Chris Cole, Dean Inada, and Bob Clarke for useful discussions. The first author gratefully acknowledges support from Peregrine Systems, Inc., Carlsbad, Ca. *Address all correspondence to the first author.* Email: *bbartell@cs.ucsd.edu* Surface mail: Dept. of Computer Science & Engineering-0114, UC San Diego, La Jolla, CA 92093.

References

[1] I. Borg and J. Lingoes. *Multidimensional Similarity Structure Analysis*. Springer-Verlag, New York, 1987.

[2] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[3] Susan T. Dumais. Enhancing performance in latent semantic indexing (LSI) retrieval. Technical Report Technical Memorandum, Bellcore, September 1990.

[4] James E. Everett and Antony Pecotich. A combined loglinear/MDS model for mapping journals by citation analysis. *Journal of the American Society for Information Science*, 42(6):405–413, 1991.

[5] George W. Furnas, Thomas K. Landauer, L. M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communications. *Communications of the ACM*, (30):964–971, 1987.

[6] Michael J. Greenacre and Leslie G. Underhill. Scaling a data matrix in a low-dimensional euclidean space. In Douglas M. Hawkins, editor, *Topics in Applied Multivariate Analysis*, pages 183–268. Cambridge University Press, 1982.

[7] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, March 1964.

[8] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1983.

[9] Gilbert W. Stewart. *Introduction to Matrix Computations*. Academic Press, 1973.

[10] W. S. Torgerson. *Theory and Methods of Scaling*. New York: John Wiley, 1958.

A Details of Solution

The proof that $\hat{\mathbf{W}} = \mathbf{R}\mathbf{M}_k^T\mathbf{C}\mathbf{X}^+$ is a solution to equation (8) is straight forward. Consider first the left side of (8):

$$\begin{aligned}(\mathbf{W}\mathbf{X})(\mathbf{W}\mathbf{X})^T(\mathbf{W}\mathbf{X})\mathbf{X}^T &= (\mathbf{R}\mathbf{M}_k^T\mathbf{C}\mathbf{P}^{\mathbf{X}})(\mathbf{P}^{\mathbf{X}}\mathbf{C}^T\mathbf{M}_k\mathbf{R}^T)(\mathbf{R}\mathbf{M}_k^T\mathbf{C}\mathbf{P}^{\mathbf{X}})\mathbf{X}^T \\ &= \mathbf{R}\mathbf{M}_k^T\mathbf{M}\mathbf{\Sigma}\mathbf{N}^T\mathbf{C}^T\mathbf{M}_k\mathbf{M}_k^T\mathbf{M}\mathbf{\Sigma}\mathbf{N}^T\mathbf{X}^T \\ &= \mathbf{R}\mathbf{\Sigma}_k\mathbf{N}^T\mathbf{C}^T\mathbf{M}_k\mathbf{\Sigma}_k\mathbf{N}^T\mathbf{X}^T\end{aligned} \tag{10}$$

These derivations make use of the facts that $\mathbf{X}\mathbf{X}^+ = \mathbf{P}^{\mathbf{X}}$, $\mathbf{P}^{\mathbf{X}}\mathbf{P}^{\mathbf{X}} = \mathbf{P}^{\mathbf{X}}$, $\mathbf{R}^T\mathbf{R} = \mathbf{I}$, and $\mathbf{C}\mathbf{P}^{\mathbf{X}} = \mathbf{M}\mathbf{\Sigma}\mathbf{N}^T$ by definition.

Now consider the right side of equation (8):

$$\begin{aligned}\mathbf{W}\mathbf{X}\mathbf{S}\mathbf{X}^T &= \mathbf{R}\mathbf{M}_k^T\mathbf{C}\mathbf{P}^{\mathbf{X}}\mathbf{C}^T\mathbf{C}\mathbf{X}^T \\ &= \mathbf{R}\mathbf{M}_k^T\mathbf{M}\mathbf{\Sigma}\mathbf{N}^T\mathbf{C}^T\mathbf{C}\mathbf{X}^T \\ &= \mathbf{R}\mathbf{\Sigma}_k\mathbf{N}^T\mathbf{C}^T\mathbf{C}\mathbf{P}^{\mathbf{X}}\mathbf{X}^T \\ &= \mathbf{R}\mathbf{\Sigma}_k\mathbf{N}^T\mathbf{C}^T\mathbf{M}\mathbf{\Sigma}\mathbf{N}^T\mathbf{X}^T \\ &= \mathbf{R}\mathbf{\Sigma}_k\mathbf{N}^T\mathbf{C}^T(\mathbf{M}_{k,t\times r} + \mathbf{M}_{r-k,t\times r})\mathbf{\Sigma}\mathbf{N}^T\mathbf{X}^T\end{aligned} \tag{11}$$

where $\mathbf{M}_{k,t\times r}$ denotes the matrix $\mathbf{M}$ retaining the first k orthonormal columns, but with the other columns zero. $\mathbf{M}_{r-k,t\times r}$ similarly denotes the matrix $\mathbf{M}$ retaining the last $r-k$ orthonormal columns, but with the first k columns zero. Obviously, $\mathbf{M} = \mathbf{M}_{k,t\times r} + \mathbf{M}_{r-k,t\times r}$. The derivation in (11) makes use of $\mathbf{X}^T = \mathbf{P}^{\mathbf{X}}\mathbf{X}^T$.

Comparing the final expressions derived in equations (10) and (11), we see that to verify the equality in equation (8) for $\hat{\mathbf{W}} = \mathbf{R}\mathbf{M}_k^T\mathbf{C}\mathbf{X}^+$, we need only show that $\mathbf{R}\mathbf{\Sigma}_k\mathbf{N}^T\mathbf{C}^T\mathbf{M}_{r-k,t\times r}\mathbf{\Sigma}\mathbf{N}^T\mathbf{X}^T = \mathbf{0}$. Noting that $\mathbf{N}^T\mathbf{C}^T = \mathbf{\Sigma}\mathbf{M}^T$,

$$\begin{aligned}\mathbf{R}\mathbf{\Sigma}_k\mathbf{N}^T\mathbf{C}^T\mathbf{M}_{r-k,t\times r}\mathbf{\Sigma}\mathbf{N}^T\mathbf{X}^T &= \mathbf{R}\mathbf{\Sigma}_k\mathbf{\Sigma}_k\mathbf{M}_k^T\mathbf{M}_{r-k,t\times r}\mathbf{\Sigma}\mathbf{N}^T\mathbf{X}^T \\ &= \mathbf{0}\end{aligned} \tag{12}$$

since $\mathbf{M}_k^T\mathbf{M}_{r-k,t\times r} = \mathbf{0}$. Thus, $\hat{\mathbf{W}} = \mathbf{R}\mathbf{M}_k^T\mathbf{C}\mathbf{X}^+$ is indeed a solution.