# Lower Dimensional Representation of Text Data based on Centroids and Least Squares

Haesun Park, * Moongu Jeon, † and J. Ben Rosen ‡

July, 2001
February, 2003, revised


## Abstract

Dimension reduction in today's vector space based information retrieval system is essential for improving computational efficiency in handling massive amounts of data. A mathematical framework for lower dimensional representation of text data in vector space based information retrieval is proposed using minimization and a matrix rank reduction formula. We illustrate how the commonly used Latent Semantic Indexing based on the Singular Value Decomposition (LSI/SVD) can be derived as a method for dimension reduction from our mathematical framework. Then two new methods for dimension reduction based on the centroids of data clusters are proposed and shown to be more efficient and effective than LSI/SVD when we have a priori information on the cluster structure of the data. Several advantages of the new methods in terms of computational efficiency and data representation in the reduced space, as well as their mathematical properties are discussed.

Experimental results are presented to illustrate the effectiveness of our methods on certain classification problems in a reduced dimensional space. The results indicate that for a successful lower dimensional representation of the data, it is important to incorporate a priori knowledge in the dimension reduction algorithms.


*The work of all three authors was supported in part by the National Science Foundation grant CCR-9901992. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF). Dept. of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, MN 55455, U.S.A., e-mail: hpark@cs.umn.edu.

†Dept. of Computer Science and Engineering, Univ. of California, Santa Barbara, Santa Barbara, CA 93106, U.S.A., e-mail: jeon@cs.ucsb.edu.

‡Dept. of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, MN 55455 and Dept. of Computer Science and Engineering, Univ. of California, San Diego, La Jolla, CA 92093, U.S.A. e-mail: jbrosen@cs.ucsd.edu.

# 1 Introduction

Today's exponential growth of the internet and computing power make it possible to accumulate tremendous amounts of data while users demand more efficient techniques to obtain useful information from this data. In information retrieval systems, the data takes various forms such as text, image and multimedia. In this paper we will concentrate on text or document data, and especially on vector space based methods. The vector space based information retrieval system, originated by G. Salton [33, 34], represents documents as vectors in a vector space.

Specifically, a term-document matrix

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n} \tag{1}$$

is formed based on the collection of documents, where $m$ is the total number of terms in the document collection and $n$ is the number of documents. Each column of $A$ represents a document, and in the matrix $A = (a_{ij})$, $a_{ij}$ are weighted frequencies of each word in a specific document representing the importance of term $i$ in document $j$. The simplest $a_{ij}$ is binary, but to improve the retrieval performance, various weighting methods have been developed [12, 34]. For other related topics such as stemming and removing stop lists, see [12, 34]. The SMART system [34] is one of the most influential test beds where the vector space based method is successfully implemented.

One major advantage of a vector space based method is that the algebraic structures of the term-document matrix can be exploited using the techniques developed in linear algebra. In particular, we believe that incorporation of a priori knowledge in the data such as cluster structures in its vector space representation is important in building an effective information retrieval system.

For achieving higher efficiency and effectiveness in manipulating these data, it will be necessary to find a lower dimensional representation of the data [9]. A vector space based information retrieval system needs to solve the following three problems frequently: document retrieval, classification, and clustering. Document retrieval is used to extract relevant documents from a text database given a query. Classification is the process of assigning new data to its proper group. The group is also called class or category [1]. A common classification system is composed of data collection, feature generation, feature selection, classifier design, and finally system evaluation and feedback [15, 29, 35]. Among them feature selection is of great importance for the quality of classification and computational cost of the classifier. Dimension reduction can be considered as a feature selection process in a vector space based method. Clustering is the process to find homogeneous groups (clusters) of data based on the values of their vector components and a predefined measure [20]. While the category structure is known in classification, in cluster analysis little or nothing is known about the category structure. All that is available is a collection of data whose category memberships are unknown. The objective is to discover a category structure in the data set [1].

In document retrieval, classification, and in some clustering algorithms, the major computation involves comparison of two vectors, which will be affected by different weighting schemes and the similarity measures [22, 34]. With dimension reduction of the given text collection, the complexity of subsequent computations involved in these problems can be substantially reduced. To achieve higher efficiency in computation, often it is necessary to reduce the dimension severely, and in the process, we may lose much information which was available in the original data. Therefore, it is important to achieve a *better representation* of data in the lower dimensional space according to specific tasks to be performed after the dimension reduction, such as classification, rather than simply reducing the dimension of the data to best approximate the full term-document matrix. The significance of this has been recognized by Hubert, et.al. [19], for example. The difficulty involved is that it is not easy to measure how well a certain dimension reduction method provides a good representation of the original data. It seems that this can only be estimated using experimental results.

The dimension reduction methods that we will discuss in this paper are based on the vector subspace computation in linear algebra. Unlike other probability and frequency based methods, where a set of representative words are chosen, the vector subspace computation will give reduction in the dimension of the term space where for each dimension in the reduced space we cannot easily attach corresponding words or a meaning. In Latent semantic indexing (LSI) [2, 3, 8, 13, 31], the singular value decomposition (SVD) of the term-document matrix is utilized for conceptual retrieval and lower dimensional representation. The dimension reduction by the optimal lower rank approximation from the SVD has been successfully applied in numerous applications, e.g. in signal processing. In these applications, often what the dimension reduction achieves is the effect of removing noise in the data. In case of information retrieval, often the term-document matrix has either full rank or close-to full rank. Also the meaning of *noise* in the text data collection is not well understood, unlike in other applications such as signal processing [32] or image processing. In addition, in information retrieval, the lower rank approximation is not only a tool for rephrasing a given problem into another which is easier to solve, but the data representation in the lower dimensional space itself is important [19] for further data processing.

In this paper, we propose a mathematical framework for lower dimensional representation of text data using the tools of minimization and matrix rank reduction formula. In our method, a lower dimensional representation of the original data is achieved by finding a lower rank approximate decomposition of the data matrix. This approximation is realized as a minimization problem, or by using matrix rank reduction formula. When the minimization is achieved using the matrix Frobenius norm, the lower dimensional representation of the data becomes the projected representation. How successfully we choose the projection will certainly influence the quality of the lower dimensional representation of the data. In particular, it will be important to choose the projection so that a priori knowledge on the data collection is reflected as much as possible. However, it is not always clear how to represent a priori knowledge mathematically to obtain better lower dimensional data. We attempt to present a general mathematical framework of dimension reduction in vector space based information retrieval, propose two dimension reduction methods based on the centroids and the least squares formulation, and illustrate the importance of incorporating a priori knowledge of the cluster structure in the data.

## 2 Lower Dimensional Representation of Term-Document Matrix

In a vector space based text retrieval system, each document is treated as a vector. Each term corresponds to one component in the vector, therefore, occupies one dimension in the Cartesian coordinate system. This means that implicitly, it is assumed that the terms are independent from each other. In fact, this assumption is not necessarily true, for some words are more closely related. In addition, in handling the vast amounts of data, allowing one extra dimension for each term will make the computation complexity extremely high even after preprocessing with stemming and removing the stop lists [12, 24]. Therefore, the dimension reduction in vector space based information retrieval is important for higher efficiency and effectiveness.

To mathematically understand the problem of lower dimensional representation of the given document sets, we will first assume that the reduced dimension, which we will denote as $k$ ($k << min(m, n)$), is given or determined in advance. Then given a term-document matrix $A \in \mathbb{R}^{m\times n}$ and an integer $k$, the problem is to find a transformation $G^T \in \mathbb{R}^{k\times m}$ that maps each vector $a_i$ in the $m$ dimensional space to a vector $y_i$ in the $k$ dimensional space :

$$G^T : a_i \in \mathbb{R}^{m\times 1} \rightarrow y_i \in \mathbb{R}^{k\times 1}, 1 \leq i \leq n. \tag{2}$$

Once the transformation $G^T$ is found, any vector $q \in \mathbb{R}^{m\times 1}$ can find its $k$ dimensional representation as $\hat{q} = G^T q \in \mathbb{R}^{k\times 1}$. Rather than looking for the mapping that achieves this explicitly, one can rephrase this as an approximation problem where the given matrix $A$ has to be decomposed into two matrices $B$ and $Y$ as

$$A \approx BY \tag{3}$$

where both $B \in \mathbb{R}^{m\times k}$ with rank($B$) = $k$ and $Y \in \mathbb{R}^{k\times n}$ with rank($Y$) = $k$ are to be found. This lower rank approximate factorization is not unique since for any nonsingular matrix $Z \in \mathbb{R}^{k\times k}$,

$$A \approx BY = (BZ)(Z^{-1}Y),$$

and $rank(BZ) = k$ and $rank(Z^{-1}Y) = k$. This problem of approximate decomposition (3) can be recast in two different but related ways. The first is in terms of matrix rank reduction formula and the second is as a minimization problem. The matrix rank reduction formula has been studied extensively in numerical linear algebra as well as psychometrics and applied statistics [6, 7, 16, 17, 19]. Here, we summarize the results that are relevant to our problem of lower dimensional representation of the term-document matrix.

THEOREM 1 *(Matrix Rank Reduction Theorem) Let $A \in \mathbb{R}^{m\times n}$ be a given matrix with $rank(A) = r$. Then the matrix*

$$E = A - (AS)(PAS)^{-1}(PA) \tag{4}$$

*where $P \in \mathbb{R}^{k\times m}$ and $S \in \mathbb{R}^{n\times k}$, $k \leq r$, satisfies*

$$rank(E) = rank(A) - rank((AS)(PAS)^{-1}(PA)) \tag{5}$$

*if and only if $PAS \in \mathbb{R}^{k\times k}$ is nonsingular.*

The only restriction on the premultiplier $P$ and the postmultiplier $S$ is on their dimensions and that the product $PAS$ be nonsingular. It is this choice for $P$ and $S$ that makes the dimension reduction flexible and makes incorporation of a priori knowledge possible. For our purpose, we will concentrate mostly on Eqn. (4), which we call *matrix rank reduction formula*. In fact, in [6] it is shown that many of the matrix decompositions can be derived using the matrix rank reduction formula. It is easy to see that the rank $k$ approximation from the truncated SVD of $A$ provides a solution that minimizes $||E||_2$ or $||E||_F$ [4, 11, 14]. We will discuss more on this in the next subsection.

Minimizing the error matrix $E$ in a certain norm $l$ is equivalent to solving a minimization problem

$$\min_{B,Y} ||A - BY||_l \tag{6}$$

where $B \in \mathbb{R}^{m\times k}$ with rank($B$) = $k$ and $Y \in \mathbb{R}^{k\times n}$ with rank($Y$) = $k$. Eqn. (4) and Eqn. (6) are related as

$$BY = (AS)(PAS)^{-1}PA.$$

It is well known that with $l = 2$ or $F$, the best approximation is obtained from the SVD of $A$. The commonly used LSI exploits the SVD of the term-document matrix [2, 3, 8, 10]. We emphasized that for a successful rank reduction scheme, it is important to exploit a priori knowledge. The incorporation of a priori knowledge can be translated to choosing the matrix $P$ and $S$ in (4), or adding a constraint in the minimization problem (6). However, mathematical formulation of the a priori knowledge as a constraint is not always easy or even possible.

In the following, we discuss various ways to choose the matrices $B$ and $Y$. In Section 3, we also show some interesting test results that illustrate that although the SVD of $A$ gives the best *approximation* in terms of minimizing the distance between $A$ and $BY$ for $l = 2$ or $F$, some other choices of $B$ and $Y$ based on the clustering of the data matrix $A$ may give a better reduced dimensional *representation* of the original documents. The choice of reduced dimension $k$ will be discussed later, and for now we will assume that the integer $k$ is given. The LSI/SVD has been a topic of active research [2, 3, 27] and the efforts have been made to facilitate the usage of the SVD by either parallelizing, preserving sparsity of the term-document matrix, or finding a faster decomposition that can approximate the SVD. The SVD certainly gives the best lower rank approximation. However, we show in this paper that the SVD may not give the ultimate solution in the case of dimension reduction for the task of classification. We summarize the LSI/SVD in the next subsection.

## 2.1 Latent Semantic Indexing with SVD

It is well known that for any matrix $A \in \mathbb{R}^{m\times n}$, its singular value decomposition (SVD) exists [4, 14]. The SVD is defined as

$$A = U\Sigma V^T \tag{7}$$

where

$$U \in \mathbb{R}^{m\times m}, \quad \Sigma \in \mathbb{R}^{m\times n}, \quad V \in \mathbb{R}^{n\times n}$$

$$U^TU = I_m, \quad V^TV = I_n, \quad \Sigma = diag(\sigma_1 \cdots \sigma_p)$$

with $p = min(m, n)$, $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_p \geq 0$ which are the singular values, and the columns of $U$ and $V$ are left and right singular vectors, respectively. It is widely known that the noise filtering can be achieved via a truncated SVD. If we replace a trailing diagonal components of $\Sigma$ with zeros and eliminate the corresponding left and right singular vectors, then the rank $k$ representation can be obtained as

$$A \approx A_k = U \begin{pmatrix} \Sigma_k & 0 \\ 0 & 0 \end{pmatrix} V^T = \begin{pmatrix} U_k & \hat{U} \end{pmatrix} \begin{pmatrix} \Sigma_k & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_k^T \\ \hat{V}^T \end{pmatrix} = U_k \Sigma_k V_k^T, \tag{8}$$

where $U_k \in \mathbb{R}^{m\times k}, \Sigma_k \in \mathbb{R}^{k\times k}$, and $V_k \in \mathbb{R}^{n\times k}$. The rank $k$ approximation (8) can also be obtained when the matrix $E$ in the matrix rank reduction formula shown in (4) is minimized using matrix $L_2$ norm or Frobenius norm. It can be easily shown that the minimum error is obtained with $P = U_k^T$ and $S = V_k$, which gives

$$\begin{aligned} (AS)(PAS)^{-1}(PA) &= (AV_k)(U_k^T AV_k)^{-1}(U_k^T A) \\ &= (U_k\Sigma_k)(\Sigma_k)^{-1}(\Sigma_k V_k^T) = U_k\Sigma_k V_k^T. \end{aligned}$$

The LSI is based on the assumption that there is some underlying latent semantic structure in the data of the term-document matrix that is corrupted by the wide variety of words used in documents and queries for the same objects (the problem of polysemy and synonymy, see [8]). It is claimed that the SVD and statistics of the frequency of association terms make it possible that two documents of similar topic can belong to the same cluster or be retrieved simultaneously for a query even if they do not share the same keywords [2]. The basic idea of the LSI/SVD is that if two document vectors represent the same topic, they will share many associating words with a keyword, and they will have very close semantic structures after dimension reduction via the SVD.

In classification, clustering, and document retrieval, the fundamental operation is to compare a document (or pseudo-document) to another document (or pseudo-document). The choice of similarity measure plays an important role [22]. In the vector space based information retrieval, the most commonly used similarity measures are, $L_2$ norm (Euclidean distance), inner product, cosine, or variations of these [22]. When the inner product is used as a measure, the documents are compared as,

$$A^T A \approx A_k^T A_k = V_k \Sigma_k^T U_k^T U_k \Sigma_k V_k^T = (V_k \Sigma_k^T)(\Sigma_k V_k^T), \tag{9}$$

i.e., the inner product between a pair of columns of $A$ can be approximated by the inner product between a pair of columns of $\Sigma_k V_k^T$. Accordingly, $V_k \Sigma_k^T \in \mathbb{R}^{k \times n}$ is considered a representation of the document vectors in the reduced dimension. This argument holds for the cosine similarity measure as well when the vectors are properly normalized. In general the above derivation is valid only for the inner product measure.

Consider the inner product between a new vector $q \in \mathbb{R}^{m \times 1}$ and the document vectors in $A$:

$$q^T A \approx q^T A_k = (q^T)(U_k \Sigma_k V_k^T) = (q^T U_k)(\Sigma_k V_k). \tag{10}$$

Eqn. (10) shows that a new vector $q \in \mathbb{R}^{m \times 1}$ can be represented as

$$\hat{q} = U_k^T q \tag{11}$$

in the $k$ dimensional space, since the columns of $\Sigma_k V_k^T$ represent the columns of $A$ in the $k$ dimensional space. In LSI/SVD, it has also been proposed that $q$ be reduced to a vector in $\mathbb{R}^{k \times 1}$ as

$$\hat{q} = \Sigma_k^{-1} U_k q. \tag{12}$$

In Section 2.2.4, we give a different derivation and interpretation of the transformations (11) and (12) of $q$ to $\hat{q}$. In fact, our derivation will clearly illustrate how one can obtain $\hat{q} \in \mathbb{R}^{k \times 1}$ shown in Eqn. (12).

## 2.2 Dimension Reduction of Cluster Structured Data

Although the SVD provides the optimal approximation $BY$ of $A$ that gives the minimum distance when $l = 2$ or $F$ in Eqn. (6), the SVD does not take into account that the data matrix $A$ is *cluster structured*, i.e., its columns can be grouped into a number of clusters. In a cluster structured matrix, each column is more closely related to a certain set of columns than to others. We will show that there are other approximation schemes that are often superior to the SVD in producing better reduced dimensional *representation* of the text data for the *classification* task when the data has a cluster structure.

### 2.2.1 Representation of Each Cluster

First, we will assume that the data set is cluster structured and already grouped into a number of clusters. This assumption is not a restriction, since we can cluster the data set if it is not already clustered, using one of the several existing clustering algorithms such as k-means [9, 20]. Also especially when the data set is huge, we can assume that the data has a cluster structure and it is often necessary to cluster the data first to utilize the tremendous amount of information in an efficient way.

Suppose we are given a data matrix $A$ whose columns are grouped into $k$ clusters. We want to find the matrices $B$ and $Y$ with $k$ columns and $k$ rows, respectively, with $A \approx BY$, so that the $k$ clusters are represented well in a reduced dimensional space. For this purpose, we want to choose each column of $B$ so that it *represents* the corresponding cluster. For any given scalar data set with the data items $\alpha_1, \alpha_2, \cdots, \alpha_n$, a commonly used representative of the data set is the *mean* value

$$m_\alpha = \frac{1}{n}\sum_{i=1}^{n} \alpha_i. \tag{13}$$

The mean gives the minimum variance

$$\sum_{i=1}^{n}(\alpha_i - m_\alpha)^2 = \min_{\delta\in\mathbb{R}} \sum_{i=1}^{n}(\alpha_i - \delta)^2 = \min_{\delta\in\mathbb{R}} \|(\alpha_1 \cdots \alpha_n) - \delta(1 \cdots 1)\|_2^2. \tag{14}$$

The mean value can be extended to data sets in a vector space. Suppose $a_1, a_2, \cdots, a_n \in \mathbb{R}^{m\times 1}$. Then its *centroid* defined as

$$c = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{1}{n}Ae \in \mathbb{R}^{m\times 1} \tag{15}$$

where $e = (1, 1, \cdots, 1)^T \in \mathbb{R}^{n\times 1}$, is commonly used as a vector that represents the vector data set. The centroid is the vector which achieves the minimum variance in the following sense:

$$\sum_{i=1}^{n} \|a_i - c\|_2^2 = \min_{x\in\mathbb{R}^{m\times 1}} \sum_{i=1}^{n} \|a_i - x\|_2^2 = \min_{x\in\mathbb{R}^{m\times 1}} \|A - xe^T\|_F^2. \tag{16}$$

Eqn. (16) shows that the centroid vector gives the smallest distance in Frobenius norm between the matrix $A$ and the rank one approximation $xe^T$ where $x$ is to be determined. For other alternatives for cluster representatives, such as *medoid*, see [30]. In the following three subsections, we describe how we find the approximation $BY$ that exploits the cluster structure using the centroids, and also illustrate that the LSI/SVD is a special case of our model.

### 2.2.2 Minimization with Centroid Vectors

Assume that the data matrix $A$ is partitioned into $k$ clusters and define the *centroid matrix*

$$C = \begin{bmatrix} c_1 & c_2 & \cdots & c_k \end{bmatrix} \in \mathbb{R}^{m\times k},$$

where the $i$th column is the centroid of the $i$th cluster. In our Centroid algorithm for dimension reduction, we find the rank reducing decomposition $A \approx BY$ by first taking $B$ as the centroid matrix $C$ that represents the $k$ clusters in the data, and then solving the least squares problem

$$\min_{Y\in\mathbb{R}^{k\times n}} ||CY - A||_F.$$

The solution matrix $Y \in \mathbb{R}^{k\times 1}$ gives a $k$ dimensional representation of $A$, which is the projected representation of $A$ in the range space of the matrix $C$.

**Algorithm 1** : Centroid algorithm for Dimension Reduction

Given a data set $A \in \mathbb{R}^{m \times n}$ with $k$ clusters and a vector $q \in \mathbb{R}^{m \times 1}$, this algorithms computes a $k$ dimensional representation $\hat{q} \in \mathbb{R}^{k \times 1}$ of $q$.

1. Compute the centroid $c_i$ of the $i$th cluster, $1 \leq i \leq k$
2. Set $C = \begin{bmatrix} c_1 & c_2 & \cdots & c_k \end{bmatrix}$
3. Solve $\min_{\hat{q}} \|C\hat{q} - q\|_2$

This can also be explained using the matrix rank reduction formula. Let $N_i$ denote the set of indices of the columns of $A$ that belong to the cluster $i$. Defining a *grouping* matrix $H \in \mathbb{R}^{n \times k}$ as

$$H = F \cdot (diag(diag(F^T F)))^{-1} \text{where} \quad F \in \mathbb{R}^{n \times k} \quad \text{and} \quad F(i,j) = \begin{cases} 1 & \text{if } i \in N_j \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

the centroid matrix $C$ can be represented as

$$C = AH. \tag{18}$$

In addition, the solution $Y$ for

$$\min_Y \|CY - A\|_F$$

obtained with $C = AH$ is $Y = (C^T C)^{-1} C^T A$. This in turn, gives the matrix rank reduction expression

$$E = A - CY = A - (AH)(C^T C)^{-1} C^T A \tag{19}$$

$$= A - (AH)(H^T A^T AH)^{-1}(H^T A^T A), \tag{20}$$

with the prefactor $P = H^T A^T$ and the postfactor $S = H$ in Eqn. (4). After obtaining the decomposition $A \approx CY$ as discussed above, any new data $q \in \mathbb{R}^{m \times 1}$ can be transformed to the lower dimensional space by solving the minimization problem

$$\min_{\hat{q} \in \mathbb{R}^{k \times 1}} \|C\hat{q} - q\|_2. \tag{21}$$

This dimension reduction method is summarized in Algorithm 1, the Centroid algorithm for dimension reduction.

In the Centroid-based classification algorithm which is summarized in Algorithm 2, the similarity between a vector $q$ to be classified and the centroid vectors $c_i$, $1 \leq i \leq k$, is compared and $q$ is determined to belong to the cluster $j$ when the similarity between $q$ and $c_j$ is the highest. Since the solution for the minimization problem

$$min\|Cy - c_i\|_l, \quad 1 \leq i \leq k, \tag{22}$$

is the $i$th column of the identity matrix $e_i \in \mathbb{R}^{k \times 1}$, the centroids are represented as the unit vectors. Therefore, the centroids in the reduced space are orthogonal and the independence of the vectors that represent

**Algorithm 2** : Centroid-based Classification

Given a data matrix $A$ with $k$ clusters and $k$ corresponding centroids, $c_i$, $1 \le i \le k$, and a vector $q \in \mathbb{R}^{m \times 1}$, it finds the index $j$ of the cluster in which the vector $q$ belongs.

- find the index $j$ such that $sim(q, c_i)$, $1 \le i \le k$, is minimum (or maximum), where $sim(q, c_i)$ is the similarity measure between $q$ and $c_i$. (For example, $sim(q, c_i) = \|q - c_i\|_2$ using the $L_2$ norm, and we take the index with the minimum value. Using the cosine measure,

$$sim(q, c_i) = cos(q, c_i) = \frac{q^T c_i}{\|q\|_2 \|c_i\|_2},$$

and we take the index with the maximum value.)

the clusters is maximized as a result of projection on to the space spanned by the columns in the matrix $B$. After obtaining the $k$ dimensional representation $\hat{q}$ of the vector $q$ from the Centroid algorithm, in the Centroid-based classification algorithm $\hat{q}$ is simply compared to the unit vectors $e_i$, $1 \le i \le k$, which are the centroids of the clusters in the reduced space. When the Euclidean distance is used as the similarity measure, we look for

$$arg \min_{1 \le i \le k} \|\hat{q} - e_i\|_2. \tag{23}$$

The minimum will be achieved when $i$ is the index of the largest component of $\hat{q}$. Therefore, the computation of $k$ of the 2-norms in Eqn. (23) becomes unnecessary, and we simply need to find out where the largest component of $q$ lies and it will be the cluster index where $q$ belongs. The situation is similar when the cosine measure is used:

$$arg \max_{1 \le i \le k} \frac{\hat{q}^T e_i}{\|\hat{q}\|_2 \|e_i\|_2} \tag{24}$$

is obtained again when $i$ is the index of the largest component of $\hat{q}$. Therefore, the result of classification will be the same whether we use 2-norm or cosine as similarity measures. The computational complexity is lower since the computation of $k$ of the 2-norm (or cosine) will be replaced by finding the largest of the $k$ components in $\hat{q}$.

#### 2.2.3 Minimization with an Orthogonal Basis of the Cluster Representatives

If the factor $B$ has orthonormal columns in a rank $k$ approximation $A \approx BY$, then the correlation in $A$ is well approximated with the correlation in $Y$ since

$$A^T A \approx Y^T B^T B Y = Y^T Y.$$

In addition, most of the common similarity measures can directly be inherited from the full dimensional space to the reduced dimensional space, since for any vector $y \in \mathbb{R}^{k \times 1}$,

$$\|y\|_B = \|By\|_2 = \|y\|_2,$$

where $B$ has orthonormal columns.

**Algorithm 3** : Orthogonal Centroid algorithm for Dimension Reduction

Given a data set $A \in \mathbb{R}^{m\times n}$ with $k$ clusters and a vector $q \in \mathbb{R}^{m\times 1}$, it computes a $k$ dimensional representation $\hat{q}$ of $q$.

1. Compute the centroid $c_i$ of the $i$th cluster, $1 \le i \le k$
2. Set $C = \begin{bmatrix} c_1 & c_2 & \cdots & c_k \end{bmatrix}$
3. Compute the reduced QR decomposition of $C$, which is $C = Q_k R$
4. $\hat{q} = Q_k^T q$

No matter how the matrix $B$ is chosen, this can be achieved by computing the reduced $QR$ decomposition of the matrix $B$. The QR decomposition of $B$ gives an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$ and an upper triangular matrix $R \in \mathbb{R}^{k\times k}$ such that

$$B = Q \begin{pmatrix} R \\ 0 \end{pmatrix}. \tag{25}$$

With $Q = (Q_k \quad Q_r)$, $Q_k \in \mathbb{R}^{m\times k}$, and $Q_r \in \mathbb{R}^{m\times(m-k)}$, we have $B = Q_k R$, which is often called the *reduced QR decomposition* of $B$. Premultiplying $Q^T = (Q_k, Q_r)^T$ to Eqn. (25) gives

$$Q_k^T B = R \quad \text{and} \quad Q_r^T B = 0. \tag{26}$$

In our Orthogonal Centroid algorithm for dimension reduction, the matrix $Q_k$ is from the reduced orthogonal decomposition of the centroid matrix $C$, and the least squares solution for

$$\min_z ||Q_k z - q||_F \tag{27}$$

gives the $k$-dimensional representation $z \in \mathbb{R}^{k\times 1}$ of $q \in \mathbb{R}^{m\times 1}$. Our algorithm is summarized in Algorithm 3: Orthogonal Centroid algorithm for dimension reduction. The $k$ dimensional representations $y$ from the Centroid algorithm and $z$ from the Centroid algorithm are related since $z = Q_k^T q$ and $z = Ry$ where $y$ is the solution for

$$\min_y ||Cy - q||_F. \tag{28}$$

In both expressions $A \approx CY$ and $A \approx Q_k(RY)$, since $CY = Q_k RY$, the error $E_d$ in approximation of the data $A$ is

$$E_d = A - CY = A - AH(H^T A^T AH)^{-1} H^T A^T A,$$

and the error in correlation is

$$\begin{aligned} E_c &= A^T A - ZZ^T \\ &= A^T A - Q_k^T A A Q_k^T \\ &= A^T A - (A^T AH)(H^T A^T AH)^{-1}(H^T A^T A). \end{aligned}$$

Although $L_2$ norm is invariant under orthogonal transformation, this does not hold for the transformation in the Orthogonal Centroid method since $Q_k Q_k^T \neq I$. However, we now show that the transformation by $Q_k$ still has a certain invariance property which makes the Centroid-based classification results in the full space identical to those in the reduced space obtained by the Orthogonal Centroid algorithm [21].

**Definition: Order of Similarity**: Given any vector $q \in \mathbb{R}^{m\times 1}$ and any matrix $B \in \mathbb{R}^{m\times k}$, the order of similarity $S(q, B)$ is an ordered indices of columns of $B$ according to the similarity between $q$ and the $k$ columns of $B$. For example, when $B = \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix} \in \mathbb{R}^{m\times 3}$ and $\|q - b_1\|_2 \leq \|q - b_3\|_2 \leq \|q - b_2\|_2$, $S(q, B) = (1, 3, 2)$ in $L_2$ norm similarity measure.

The following two theorems show that the order of similarity between any data item and the centroid matrix is preserved after dimension reduction by the Orthogonal Centroid algorithm.

THEOREM 2 (ORDER PRESERVING IN $L_2$ NORM) *The order of similarity $S(q, C)$ with $L_2$ norm measure in the full dimensional space between any given vector q and the centroid matrix $C \in \mathbb{R}^{m\times k}$ is completely preserved in the reduced space obtained by the Orthogonal Centroid algorithm. That is, $S(q, C) = S(\hat{q}, \hat{C})$ where $\hat{q} = Q_k^T q$ and $\hat{C} = Q_k^T C$, and the reduced QR decomposition of C is $Q_k R$.*

Proof: Since $Q_r^T c_j = 0$ from Eqn. (26),

$$\|q - c_j\|_2^2 = \|Q^T(q - c_j)\|_2^2 = \|Q_k^T(q - c_j)\|_2^2 + \|Q_r^T(q - c_j)\|_2^2 = \|Q_k^T(q - c_j)\|_2^2 + \|Q_r^T q\|_2^2. \quad (29)$$

If

$$\|q - c_j\|_2 \leq \|q - c_l\|_2,$$

then

$$\|Q_k^T(q - c_j)\|_2^2 + \|Q_r^T q\|_2^2 \leq \|Q_k^T(q - c_l)\|_2^2 + \|Q_r^T q\|_2^2$$

and

$$\|Q_k^T(q - c_j)\|_2 \leq \|Q_k^T(q - c_l)\|_2.$$

This means that the Orthogonal Centroid reduction method preserves the order of the $L_2$ norm similarity between any vector and the centroids in the full dimensional space after dimension reduction. $\square$

THEOREM 3 (ORDER PRESERVING IN COSINE MEASURE) *The order of similarity $S(q, C)$ with cosine measure in the full dimensional space between any given vector q and the centroid matrix $C \in \mathbb{R}^{m\times k}$ is completely preserved in the reduced space obtained by the Orthogonal Centroid algorithm. That is, $S(q, C) = S(\hat{q}, \hat{C})$ where $\hat{q} = Q_k^T q$ and $\hat{C} = Q_k^T C$, and the reduced QR decomposition of C is $Q_k R$.*

Proof: Let $cos(q, c_j)$ be cosine between vectors $q$ and $c_j$. Since $Q_r^T c_j = 0$ from Eqn. (26),

$$cos(q, c_j) = cos(Q^T q, Q^T c_j) = \frac{(Q^T q)^T Q^T c_j}{\|Q^T q\|_2 \|Q^T c_j\|_2} = \frac{q^T Q_k Q_k^T c_j}{\|Q^T q\|_2 \|Q_k^T c_j\|_2}$$

and

$$cos(\hat{q}, \hat{c}_j) = cos(Q_k^T q, Q_k^T c_j) = \frac{q^T Q_k Q_k^T c_j}{\|Q_k^T q\|_2 \|Q_k^T c_j\|_2}. \quad (30)$$

If $cos(q, c_j) \leq cos(q, c_l)$, then

$$\frac{q^T Q_k Q_k^T c_j}{\|Q^T q\|_2 \|Q_k^T c_j\|_2} \leq \frac{q^T Q_k Q_k^T c_l}{\|Q^T q\|_2 \|Q_k^T c_l\|_2}.$$

Accordingly,

$$\frac{q^T Q_k Q_k^T c_j}{\|Q_k^T q\|_2 \|Q_k^T c_j\|_2} \leq \frac{q^T Q_k Q_k^T c_l}{\|Q_k^T q\|_2 \|Q_k^T c_l\|_2}, \tag{31}$$

i.e., $cos(\hat{q}, \hat{c}_j) \leq cos(\hat{q}, \hat{c}_l)$. $\square$

The above two theorems show that we can completely recover the orders of both the $L_2$ and cosine similarities between any vector and the centroids, when the original dimension is reduced to the number of categories $k$ by the Orthogonal Centroid algorithm. Therefore, Centroid-based classification produces exactly the same classification results using the reduced data from the Orthogonal Centroid algorithm as with the full dimensional data. Note that the order preserving property of the dimension reduction obtained by the Orthogonal Centroid algorithm holds regardless of the cluster quality. In fact, the order preserving property in either $L_2$ norm or cosine measure does not depend on the fact that the columns of the matrix $C$ are the centroids of the clusters. As far as the dimension reduction is achieved by an orthogonal basis $Q_B$ for the range space of *any* matrix $B$, the order between a datum and the columns of $B$ after dimension reduction by $Q_B^T$ will be preserved. However, in the above theorems, we specifically discussed the case when $B$ is the centroid matrix due to its relevance to the Orthogonal Centroid dimension reduction algorithm in conjunction with the the centroid-based classification method. The savings in computational cost from dimension reduction is obvious. In Section 3, some experimental results are given to show the performance of our algorithms.

#### 2.2.4 LSI/SVD using the Minimization Model

Now we show how the LSI/SVD can be explained using our least squares model and achieve the results shown in Eqn. (11) and Eqn. (12). A solution for the minimization problem (6) can be obtained from the SVD of $A$ with $B = U_k$ and $Y = \Sigma_k V_k^T$. In fact, this is the optimal solution when $l = 2$ or $F$. Then solving the least squares problem

$$\min_{\hat{q}} \| U_k \hat{q} - q \|_2 \tag{32}$$

we obtain

$$\hat{q} = U_k^T q. \tag{33}$$

This produces the same result as Eqn. (11). In LSI/SVD, $\hat{q}$ is compared with the document vectors in the reduced dimension, which are $\Sigma_k V_k^T$. Note that the columns of $\Sigma_k V_k^T$ are the solution vectors $\hat{y}_i$ which we would obtain from solving

$$\min_{\hat{y}} \| U_k \hat{y} - a_i \|_2 \tag{34}$$

for $1 \leq i \leq n$.

Using our minimization model, it can easily be shown that the representation $\hat{q}$ in Eqn. (12) is the solution for

$$\min_{y} \| U_k \Sigma_k \hat{y} - q \|_2, \tag{35}$$

**Algorithm 4** : k Nearest Neighbor (knn) Classification

Given a data matrix $A = [a_1, \ldots, a_n]$ with $k$ clusters and a vector $q \in \mathbb{R}^{m\times 1}$, it finds the cluster in which the vector $q$ belongs.

1. From the similarity measure $sim(q, a_j)$ for $1 \le j \le n$, find the $k^*$ nearest neighbors of $q$. (We use $k^*$ to distinguish the algorithm parameter from the number of clusters.)
2. Among these $k^*$ vectors, count the number belonging to each cluster.
3. Assign $q$ to the cluster with the greatest count in the previous step.

therefore, in this case, the matrix $B$ is considered to be $\Sigma_k U_k$ and accordingly, the document vectors in the reduced dimension are the columns of $V_k^T$. Our derivation also makes it possible for any similarity measure such as cosine and $L_2$ norm to be used in comparing the vectors in the reduced dimension. It is not restricted only to the inner product for similarity measure as in the original derivation of the LSI/SVD.

It has been reported that the LSI/SVD gives the best performance in text retrieval when the reduced dimension is between 100 and 300 [2]. Considering the size of the text documents these days, even this range of 100 to 300 can be quite a dramatic reduction of the dimension. In fact, the reduced dimension will rarely be the numerically estimated rank of the original term-document matrix, unlike in other applications such as signal processing where the SVD is often utilized.

# 3 Experimental Results

In this section, we present several experimental test results that illustrate the performance of our dimension reduction methods. We compare the classification results obtained using the data in the full space and the reduced dimensional data from our dimension reduction methods, the Centroid algorithm and the Orthogonal Centroid algorithm. For classification, we use the centroid-based classification algorithm as well as the k nearest neighbor (knn) algorithms presented in Algorithm 4. We also compare the classification results obtained using the reduced dimensional data produced by our algorithms and by the SVD.

In preprocessing the text data the stop lists such as “the” and “a” which appear in almost every document are eliminated, since they are not considered as important words differentiating documents from each other. Then the Porter’s stemming algorithm [12] was applied to process the words into their stems in order to reduce the data dimension and increase the effectiveness of classification. For example, “computer”, “computation”, and “computing” are processed into one stem word “comput”. We used the term frequencies in generating the term-document matrix.

**Test I**: In the first test, the purpose is to examine the relationship between the data items and the centroids in the full space and the reduced dimensional space. We use a relatively low dimensional clustered data set which is artificially generated by the algorithm presented in [20]. The program generated 200 data items of three classes in a 10 dimensional space. There are 63, 79 and 58 data items in class 1, class 2, and class 3, respectively, forming a data matrix $A \in \mathbb{R}^{10\times 200}$. Then the dimension of the data vectors are reduced to 3 from 10 using the Orthogonal Centroid algorithm and the classification was performed using the Centroid-based classification algorithm in the full space as well as in the reduced space. Then we compare classification in the full and reduced space. Table 1 shows the distances between each data item and the centroids in the full dimensional space and the reduced space with the $L_2$ norm measure and with

Table 1: Test I, $L_2$ norm and cosine similarity values between the data and the three centroids, in the reduced space and in the full space.

| Data | $\|Q_3^T(a_i - c_j)\|_2$ | | | $\|(a_i - c_j)\|_2$ | | | $cos(Q_3^T a_i, Q_3^T c_j)$ | | | $cos(a_i, c_j)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $c_1$ | $c_2$ | $c_3$ | $c_1$ | $c_2$ | $c_3$ | $c_1$ | $c_2$ | $c_3$ | $c_1$ | $c_2$ | $c_3$ |
| $a_1$ | $\underline{0.67}$ | 1.23 | 1.12 | $\underline{1.42}$ | 1.75 | 1.68 | $\underline{0.97}$ | 0.89 | 0.89 | $\underline{0.87}$ | 0.80 | 0.80 |
| $a_2$ | $\underline{0.17}$ | 1.46 | 0.88 | $\underline{0.88}$ | 1.70 | 1.24 | $\underline{1.00}$ | 0.74 | 0.92 | $\underline{0.92}$ | 0.69 | 0.85 |
| $a_3$ | 1.04 | $\underline{0.66}$ | 1.02 | 1.35 | $\underline{1.09}$ | 1.34 | 0.88 | $\underline{0.92}$ | 0.89 | 0.78 | $\underline{0.81}$ | 0.78 |
| $a_4$ | $\underline{0.60}$ | 1.34 | 1.53 | $\underline{0.83}$ | 1.50 | 1.63 | $\underline{0.97}$ | 0.85 | 0.79 | $\underline{0.95}$ | 0.83 | 0.77 |
| $a_5$ | $\underline{0.84}$ | 1.86 | 1.15 | $\underline{1.24}$ | 2.07 | 1.46 | $\underline{0.92}$ | 0.46 | 0.84 | $\underline{0.83}$ | 0.42 | 0.76 |
| $a_{64}$ | 1.36 | $\underline{0.33}$ | 1.48 | 1.49 | $\underline{0.68}$ | 1.59 | 0.77 | $\underline{0.98}$ | 0.73 | 0.73 | $\underline{0.94}$ | 0.70 |
| $a_{65}$ | 0.86 | $\underline{0.54}$ | 1.17 | 1.43 | $\underline{1.26}$ | 1.63 | 0.92 | $\underline{0.96}$ | 0.84 | 0.78 | $\underline{0.81}$ | 0.71 |
| $a_{66}$ | 1.09 | $\underline{0.57}$ | 1.25 | 1.31 | $\underline{0.93}$ | 1.45 | 0.89 | $\underline{0.96}$ | 0.83 | 0.78 | $\underline{0.85}$ | 0.73 |
| $a_{67}$ | 1.18 | $\underline{0.52}$ | 1.34 | 1.18 | $\underline{0.80}$ | 1.47 | 0.91 | $\underline{0.97}$ | 0.79 | 0.84 | $\underline{0.89}$ | 0.72 |
| $a_{68}$ | 1.57 | $\underline{0.87}$ | 1.38 | 1.85 | $\underline{1.31}$ | 1.70 | 0.73 | $\underline{0.93}$ | 0.80 | 0.67 | $\underline{0.85}$ | 0.73 |
| $a_{143}$ | 0.94 | 1.40 | $\underline{0.44}$ | 1.32 | 1.68 | $\underline{1.03}$ | 0.90 | 0.67 | $\underline{1.00}$ | 0.79 | 0.59 | $\underline{0.88}$ |
| $a_{144}$ | 1.22 | 1.62 | $\underline{0.28}$ | 1.50 | 1.84 | $\underline{0.92}$ | 0.83 | 0.65 | $\underline{0.99}$ | 0.76 | 0.60 | $\underline{0.91}$ |
| $a_{145}$ | 0.61 | 1.52 | $\underline{0.47}$ | 0.92 | 1.70 | $\underline{0.89}$ | 0.96 | 0.74 | $\underline{0.98}$ | 0.91 | 0.70 | $\underline{0.93}$ |
| $a_{146}$ | 0.73 | 1.30 | $\underline{0.38}$ | 0.84 | 1.36 | $\underline{0.57}$ | 0.94 | 0.75 | $\underline{0.99}$ | 0.92 | 0.74 | $\underline{0.97}$ |
| $a_{147}$ | 1.20 | 1.53 | $\underline{0.67}$ | 1.40 | 4.81 | $\underline{1.69}$ | 0.84 | 0.57 | $\underline{0.99}$ | 0.75 | 0.51 | $\underline{0.89}$ |

cosine measure. Since the space is limited, we only show the results for the first five data items of each cluster. The numbers with the underlines show the best similarity values (the smallest for the $L_2$ norm, and the largest for the cosine) and therefore the classification results from the Centroid-based classification. The first column of the table contains the label of each data, the numerical values in the next three columns are Euclidean distances between data $a_i$ and centroids $c_j$ in the full dimensional space, and the next three columns represent those in the reduced space. Then the next six columns show these values measured using the cosine similarity. Note that $a_3$ is misclassified in the full space as well as in the reduced space in both cases of $L_2$ and cosine similarity measures. In the next tests, we use larger data sets and show the classification accuracy.

In Tests II-IV, three cases are tested using the $L_2$ norm and cosine measures, and using the dimension reduction methods by the Centroid algorithm and the Orthogonal Centroid algorithm. Then using the Centroid-based classification algorithm, each item was classified in the reduced space, as well as in the full space.

**Test II**: A data set with three clusters was formed using the data from the SMART system: 82 documents from computer and information science (ADI), 1400 documents from aerospace engineering (CRANFIELD), and 1033 documents from medical science (MEDLINE). Input data and the classification results from the Centroid-based classification in the full space and in the reduced dimensional space are presented in Table 2. This is a relatively easy classification problem since the three classes are rather disjoint. The classification results from the Centroid-based classification applied to the dimension reduced data of the Centroid algorithm are identical when we used the $L_2$ norm and the cosine measure, as shown in Section 2.2.2. Also the Centroid-based classification results in the full space and the reduced space given by the Orthogonal Centroid method are identical, which corroborates Theorems 2 and 3. This phenomenon is shown

Table 2: Test II, classification accuracy in the full space and in the reduced dimensional spaces. Classification was performed using the Centroid-based classification algorithm.

| | Data from the SMART system | | |
|---|---|---|---|
| | category | number of data items | |
| 1 | ADI | 82 | |
| 2 | CRANFIELD | 1400 | |
| 3 | MEDLINE | 1033 | |
| | Classification Accuracy (in %) from Centroid-based Classification algorithm (in %) | | |
| | Full | Centroid | Orthgonal Cent. |
| Dimension | $12140 \times 2515$ | $3 \times 2515$ | $3 \times 2515$ |
| $L_2$ | 89.0 | 96.9 | 89.0 |
| Cosine | 98.4 | 96.9 | 98.4 |

throughout our tests.

**Test III**: The data set for this test is constructed using the MEDLINE database available from the National Institute of Health. We randomly selected five clusters from MEDLINE and obtained 500 documents from each cluster, forming a data set with 2500 documents. The total number of terms are 22095 after preprocessing of the stop list removal and stemming. The categories have many common words related to cancer. Using the Centroid and Orthogonal Centroid algorithms, the dimension 22095 is dramatically reduced to 5, the number of classes. Two different classification algorithms, the Centroid-based classification algorithm and the knn algorithm with various values of the parameter $k^*$, were used in this test. The classification results obtained from the Centroid-based classification algorithm are identical in the full space as well as in the 5-dimensional space given by the Orthogonal Centroid algorithm. The computational time for the knn algorithm is dramatically reduced after the dimension reduction since the comparisons between the data to be classified and all other data points are now made in the 5-dimensional space, instead of in the full 22095-dimensional space.

Table 3 shows that the classification results with both measures of $L_2$ norm and cosine are identical in the full and reduced dimensional space obtained from the Orthogonal Centroid method. The results for various values of $k^*$ with the $L_2$ norm in knn were much worse than those with the cosine measure. It is interesting that the classification accuracy of the knn algorithm was better overall after dimension reduction. In this particular data set, classification accuracy as well as the CPU time of the Centroid-based classification method was much better than those of the knn method.

**Test IV**: The data set here is an extension of the data set used in Test III. It has three extra categories of MEDLINE data in addition to five from Test III. As in the second data set, 500 documents are chosen from each category, making a data set of 4000 documents. After preprocessing, the total number of terms was 29152. Again our Centroid and Orthogonal Centroid algorithms are tested and two different similarity measures are examined to determine which measure is more appropriate for classifying the text data. The test procedure is the same as that in the previous tests. The results are shown in Table 4.

In all of the above three tests, the overall classification accuracy with the cosine measure is better. This is consistent with other substantial test results we obtained comparing several similarity measures in text processing [23]. Since the Centroid-based classification results with the cosine and $L_2$ norm measures are

Table 3: Test III, classification accuracy from two classification algorithms, the Centroid-based classification as well as the k nearest neighbor on the data in the full space and in the reduced spaces obtained by the Centroid algorithm and the Orthogonal Centroid algorithm.

| | | Data from MEDLINE | | |
|---|---|---|---|---|
| | | category | no. of data | |
| 1 | | heart attack | 500 | |
| 2 | | colon cancer | 500 | |
| 3 | | diabetes | 500 | |
| 4 | | oral cancer | 500 | |
| 5 | | tooth decay | 500 | |
| | | Classification Accuracy(in %) | | |
| | | Full | Centroid | Orthgonal Cent. |
| Dimension | | $22095 \times 2500$ | $5 \times 2500$ | $5 \times 2500$ |
| Centroid-based classification | $L_2$ | 88.2 | 91.5 | 88.2 |
| | Cosine | 92.2 | 91.5 | 92.2 |
| 5nn | $L_2$ | 72.9 | 93.2 | 92.8 |
| | Cosine | 87.6 | 93.8 | 94.1 |
| 15nn | Cosine | 88.0 | 94.1 | 94.1 |
| 35nn | Cosine | 88.6 | 93.4 | 93.6 |
| 55nn | Cosine | 88.0 | 93.1 | 93.3 |
| 100nn | Cosine | 87.2 | 92.3 | 92.7 |

the same when the data dimension was reduced by the Centroid algorithm, the classification in the reduced space appears to outperform the classification in the full space when the $L_2$ norm is used. The classification accuracy of Test II is higher than that of Test III and IV, due to the fact that the clusters in Test II are almost disjoint, while the data in Tests III and IV have many documents that belong to more than one cluster, and therefore the cluster boundary becomes more vague.

**Test V**: The purpose of this last experiment was to compare the performance of the three different dimension reduction methods, Centroid, Orthogonal Centroid, and the SVD based reduction, when used in the Centroid-based classification algorithm with the cosine measure. The experiment was conducted on the small set from the MEDLINE database. The data has 5 clusters as in Test II. The training set of 200 documents from 5 clusters were used in computing the centroids. The classification of the training data set in the reduced space was compared and presented under the columns denoted as “self”. In addition, 200 new documents were classified after their dimension was reduced to a 5 dimensional space. The results are shown in Table 5.

It is remarkable that in all our tests, even though the dimension reduction is quite severe, the two centroid based dimension reduction methods presented in this paper worked extremely well. On the other hand, when the dimension reduction is severe, the SVD based method does not perform as well as these centroid based dimension reduction methods when used in classification. However, as the reduced dimension was increased, the performance of the SVD based method became better.

Table 4: Test IV, Classification accuracy in in the full space and in the reduced spaces obtained by the Centroid algorithm and the Orthogonal Centroid algorithm. Classification was performed using the Centroid-based classification algorithm.

| | Data from MEDLINE | | |
|---|---|---|---|
| | category | no. of data | |
| 1 | heart attack | 500 | |
| 2 | colon cancer | 500 | |
| 3 | diabetes | 500 | |
| 4 | oral cancer | 500 | |
| 5 | tooth decay | 500 | |
| 6 | prostate cancer | 500 | |
| 7 | breast cancer | 500 | |
| 8 | diet weight loss | 500 | |
| | Classification Accuracy(in %) | | |
| | Full | Centroid | Orthgonal Cent. |
| Dimension | $29152 \times 2500$ | $8 \times 2500$ | $8 \times 2500$ |
| $L_2$ | 77.4 | 83.4 | 77.4 |
| Cosine | 84.0 | 83.4 | 84.0 |

# 4 Concluding Remarks

A mathematical framework for dimension reduction and new dimension reduction algorithms are presented. The Centroid algorithm is based on the projection by the matrix whose columns are the centroids of the clusters, and the Orthogonal Centroid algorithm is based on the projection via an orthonormal basis for the space spanned by the centroids. The new methods have an advantage that they are computationally far less costly than the SVD based method. In addition, they seem to give more effective results in classification for the clustered data. The algorithms were tested on several data sets. Although our experimental test results clearly illustrate that the new approaches are promising, there are still many open problems. We have illustrated our new dimension reduction model and its use for clustered data assuming that the data have already been clustered and the reduced dimension is the same as the number of clusters. Important questions include the choice of the clustering algorithm when the data set is not already clustered as well as the reduced dimension $k$. In terms of reducing the computational complexity, the smaller $k$ is, the better. However, severe reduction in dimension may cause loss of information in the original data, and the optimal choice of the dimension needs to be determined. Our scheme fixes reduced dimension $k$ as the number of clusters. If a different value of $k$ is desired, the existing clusters can either be combined or further clustered to decrease or increase the values of $k$, respectively.

We assumed that the matrix $B$ consisting of the centroids of the clusters has full column rank. This assumption cannot be guaranteed although we have not experienced any counter example in practice. Especially when the number of clusters is relatively small compared to the full dimension, the matrix $B$ was always well conditioned. Currently, we are studying the effect of dimension reduction on substantial test data sets using various classification methods including the support vector classifiers [25]. We also plan to investigate the effect of different norms such as the $L_1$ norm in minimization.

Table 5: Test V, Classification Accuracy in the full space and in three reduced spaces. Training data set of $7519 \times 200$, and new data set of $7519 \times 200$, are classified using the Centroid-based classification.

| | Centroid (in %) | | Ortho. Centroid (in %) | | SVD (in %) | |
|---|---|---|---|---|---|---|
| reduced dimension | self | new | self | new | self | new |
| 5 | 98.0 | 83.0 | 96.5 | 84.0 | 74.5 | 72.5 |
| 10 | | | | | 89.5 | 80.5 |
| 20 | | | | | 93.5 | 83.0 |
| 50 | | | | | 94.0 | 83.5 |
| 100 | | | | | 94.0 | 85.5 |
| 200 | | | | | 96.5 | 84.5 |
| full(7519) | | | | | 96.5 | 84.0 |

**Acknowledgement:**

The work of the first author was conducted in part when she was visiting the Department of Mathematics, Linkoping University, Linkoping, Sweden, during August, 2000. She would like to thank the department and Prof. Lars Eldén for their kind invitation for her to visit the department. She would also like to thank Prof. Gene Golub who was also visiting the department in August, for valuable discussions and the references which made this work more interesting. We would also like to thank the anonymous referees and Prof. Axel Ruhe whose thoughtful comments made it possible to improve the presentation significantly.

## References

[1] M. R. Anderberg. Cluster analysis for applications. *Academic Press, New York and London*, 1973.

[2] M.W. Berry, S.T. Dumais, and G.W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573-595, 1995.

[3] M.W. Berry, Z. Drmac, and E.R. Jessup. Matrices, Vector Spaces, and Information Retrieval. *SIAM Review*, 41:335-362, 1999.

[4] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, PA, 1996.

[5] J.R. Colon and S.J. Colon. Optimal Use of an Information Retrieval System. *J. of the American Society of Information Science*, 47(6):449-457, 1996.

[6] M.T. Chu, R.E. Funderlic, and G.H. Golub. A rank-reduction formula and its applications to matrix factorizations. *SIAM Rev.*, 37: 512-530, 1995.

[7] R.E. Cline and R.E. Funderlic. The rank of a difference of matrices and associated generalized inverses. *Linear Algebra Appl.*, 24:185-215, 1979.

[8] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. of the Society for Information Science*, 41:391-407, 1990.

[9] I. S. Dhillon and D.S. Modha. Concept Decompositions for Large Sparse Text Data using Clustering. *Machine Learning* 42(1):143-175, 2001.

[10] S. T. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23:229-236, 1991.

[11] C. Eckart and G. Young. The approximation of one matrix by another lower rank. *Psychometrika*, 1:211-218, 1936.

[12] W.B. Frakes and R. Baeza-Yates. Information Retrieval: Data Structures and Algorithms. *Prentice Hall*, Englewood Cliffs, New Jersey, 1992.

[13] M.D. Gordon. Using Latent Semantic Indexing for Literature Based Discovery. *J. of the American Society for Information Science*, 49(8):674-685, 1998.

[14] G.H. Golub and C.F. Van Loan. *Matrix Computations*, third edition. Johns Hopkins University Press, Baltimore, 1996.

[15] E. Gose, R. Johnsonbaugh and S. Jost. Pattern Recognition and Image Analysis. *Prentice Hall*, Upper Saddle River, New Jersey, 1996.

[16] L. Guttman. A necessary and sufficient formula for matrix factoring. *Psychometrika*, 22:79-81, 1957.

[17] S. Harter. Psychological Relevance and Information Science. *J. of the American Society of Information Science*, 43(9):602-615, 1992.

[18] H.S. Heaps. Information Retrieval, Computational and Theoretical Aspects. *Academic Press*, 1978.

[19] L. Hubert, J. Meulman, and W. Heiser. Two Purposes for Matrix Factorization: A Historical Appraisal. *SIAM REVIEW*, Vol. 42, No. 1, pp.68-82, 2000.

[20] A.K. Jain, and R.C. Dubes. Algorithms for Clustering Data. *Prentice Hall*, 1988.

[21] M. Jeon. Centroid-Based Dimension Reduction Methods for Classification of High Dimensional Text Data, *Ph.D. Dissertation, University of Minnesota*, June 2001.

[22] Y. Jung, H. Park, and D. Du. An Effective Term-Weighting Scheme for Information Retrieval, Technical Report TR00-008. Department of Computer Science and Engineering, University of Minnesota, Twin Cities, U.S.A., 2000.

[23] Y. Jung, H. Park, and D. Du. A Balanced Term-Weighting Scheme for Improved Document Comparison and Classification, *preprint*, 2001.

[24] G. Kowalski. Information Retrieval System: Theory and Implementation, *Kluwer Academic Publishers*, 1997.

[25] H. Kim, P. Howland, and H. Park. Text Categorization using Support Vector Machines with Dimension Reduction, preprint, 2003.

[26] T. G. Kolda. Limited-memory matrix methods with applications. *Dissertation*, Applied Mathematics, University of Maryland, 1997.

[27] T. G. Kolda and D. P. O'Leary. A Semi-Discrete Matrix Decomposition for Latent Semantic Indexing in Information Retrieval. *ACM Transactions on Information Systems*, 16:322-346, 1996.

[28] R. Krovetz and W.B. Croft. Lexical Ambiguity and Information Retrieval. *ACM Transactions on Information Systems*, 10(2):115-241, 1992.

[29] M. Nadler and E.P. Smith. Pattern Recognition Engineering, *John Wiley & Sons*, 1993.

[30] R.T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining, *Proceedings of the 20th International Conference on Very Large Databases*, 144-155, 1994.

[31] A.M. Pejtersen. Semantic Information Retrieval. *Communications of the ACM*, 41(4):90-92, 1998.

[32] J.B. Rosen, H. Park, and J. Glick. Total least norm formulation and solution for structured problems. *SIAM Journal on Matrix Anal. Appl.*, 17:110-128, 1996.

[33] G. Salton. The SMART retrieval system, *Prentice Hall*, 1971.

[34] G. Salton, and M.J. McGill. Introduction to Modern Information Retrieval, *McGraw-Hill*, 1983.

[35] S. Theodoridis and K. Koutroumbas. Pattern Recognition, *Academic Press*, 1999

[36] D. Zhang, R. Ramakrishan, and M. Livny. An efficient data clustering method for very large databases. *Proc. of the ACM SIGMOD conference on management of data, Montreal, Canada*, June 1996.