LBNL-43202

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

A Similarity-Based Probability Model for Latent Semantic Indexing

Chris H.Q. Ding

National Energy Research Scientific Computing Division

May 1999

To be presented at the
22nd International ACM SIGIR Conference on Research and Development in Information Retrieval,
Berkeley, CA,
December 15–19, 1999,
and to be published in
the Proceedings

REFERENCE COPY
Does Not
Circulate
Bldg. 50 Library - Ref.
Lawrence Berkeley National Laboratory
Copy 1
LBNL-43202

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBNL-43202

A Similarity-Based Probability Model for Latent Semantic Indexing

Chris H.Q. Ding

National Energy Research Scientific Computing Division
Ernest Orlando Lawrence Berkeley National Laboratory
University of California
Berkeley, California 94720

May 1999

This work was supported by the Director, Office of Computational and Technology Research, Division of Mathematical, Information, and Computational Sciences, and the Director, Office of Science, Office of Laboratory Policy and Infrastructure, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

A Similarity-based Probability Model for Latent Semantic Indexing

Chris H.Q. Ding
NERSC Division, Lawrence Berkeley National Laboratory
University of California, Berkeley, CA 94720

Abstract

A dual probability model is constructed for the Latent Semantic Indexing (LSI) using the cosine similarity measure. Both the document-document similarity matrix and the term-term similarity matrix naturally arise from the maximum likelihood estimation of the model parameters, and the optimal solutions are the latent semantic vectors of of LSI. Dimensionality reduction is justified by the statistical significance of latent semantic vectors as measured by the likelihood of the model. This leads to a statistical criterion for the optimal semantic dimensions, answering a critical open question in LSI with practical importance. Thus the model establishes a statistical framework for LSI. Ambiguities related to statistical modeling of LSI are clarified.

1 Introduction

Automatic document retrieval to a user query, such as searching documents on Internet search engines, often matches the keywords in the query to the index words for all documents in the database, following the vector-space model for information retrieval (IR)[1]. Latent semantic indexing (LSI) [2, 3, 4, 5] is a successful scheme to go beyond lexical matching to address the well-known problem of using individual keywords to identify the content of documents. LSI attempts to capture the underlying or latent semantic structures, which better index the documents than individual indexing terms, by the truncated singular value decomposition (SVD) of the term-document matrix X. The effectiveness of LSI has been demonstrated empirically in several text collections as increased average retrieval precision.

Clearly, a theoretical and quantitative understanding beyond empirical evidences is desirable. To date, several theoretical results or explanations [6, 7, 8, 9] have appeared, and these studies provide better understanding of LSI. However, many fundamental questions remain unresolved.

In this paper, we outline a dual probabilistic model for LSI based on the similarity concept widely used in vector-space model. For this model, the term-similarity matrix XX^T and document similarity matrix X^TX naturally arise during the maximum likelihood estimation, and LSI is the optimal solution of the model.

From statistical point of view, LSI amounts to an effective dimensionality reduction, i.e., reduce the problem dimension to k-dim LSI space. Dimensions with small singular values are thus often viewed as representing semantic noises and thus are ignored. This generic argument, considering its fundamental importance, needs to be clarified, quantified and verified. Our model provides a mechanism to do so by checking the statistical significance of the semantic dimensions: If a few semantic dimensions can effectively characterize the data statistically, as indicated by the likelihood of the model, we believe they also effectively represent the

semantic meanings/relationships as defined by the cosine similarity.

Thus the likelihood is the key to the verification of optimal semantic subspace that LSI advocates. We give some theoretical results and an illustrative example to support the existence of such optimal semantic subspace. In doing so, a criterion for determining the optimal semantic dimensions can be defined, addressing a critical open question in LSI with practical importance.

2 Latent Semantic Indexing

In vector-space model of information retrieval, the term to document association relation is represented as a term-document matrix

$$X = \begin{pmatrix} x_1^1 & \dots & x_n^1 \\ \vdots & \ddots & \vdots \\ x_1^d & \dots & x_n^d \end{pmatrix} \equiv (\mathbf{x}_1 \cdots \mathbf{x}_n) \equiv \begin{pmatrix} \mathbf{t}^1 \\ \vdots \\ \mathbf{t}^d \end{pmatrix} \quad (1)$$

containing the frequence of the d index terms occurring in the n documents and properly weighted by other factors[3, 10]. Here $\mathbf{x}_i$ is a d-component column vector representing a document $((\mathbf{x}_i)^\alpha \equiv x_i^\alpha)$. $\mathbf{t}^\beta$ is a n-component row vector representing a term $((\mathbf{t}^\beta)_j \equiv x_j^\beta)$. (In this paper, capital letters refer to matrices, bold face lower-case letters to vectors; subscripts refer to documents, superscripts refer to terms; α, β sum over all d terms and i, j sum over all n documents.) Given a user query $\mathbf{q}$, consisting of a set of terms (keywords), the system calculate a n-component score vector $\mathbf{s} = \mathbf{q}^T X$ as the relevance of each document to the query. The relevant documents are sorted according to the score and returned to user.

LSI re-represents both terms and documents in a new vector space with smaller k dimensions, in order to capture the underlying or latent structures (indices). This is done through the truncated singular value decomposition, $X \simeq U_k \Sigma_k V_k^T$, or explicitly,

$$X \simeq (\mathbf{u}_1 \cdots \mathbf{u}_k) \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix} \begin{pmatrix} \mathbf{v}^1 \\ \vdots \\ \mathbf{v}^k \end{pmatrix} \quad (2)$$

where $\mathbf{u}_1 \cdots \mathbf{u}_k$ and $\mathbf{v}^1 \cdots \mathbf{v}^k$ are left and right singular vectors. $\sigma_1, \cdots, \sigma_k$ are singular values. Mathematically, the truncated SVD is the best approximation of X in the reduced k-dimensional subspace. In this k-dim LSI subspace, query is represented as $\mathbf{q}^T U_k$, and documents are represented as columns of $\Sigma_k V_k^T$. The score vector is calculated as $\mathbf{s} = (\mathbf{q}^T U_k)(\Sigma_k V_k^T)$.

Here we point out an important feature of LSI: if document vectors (columns of X) are normalized to one in the original d-dim space, their representations (columns of $\Sigma_k V_k^T$) in the reduced k-dim LSI subspace are also normalized to one. To prove this, we have

$$(\Sigma_k V_k^T)^T (\Sigma_k V_k^T) = (U_k \Sigma_k V_k^T)^T (U_k \Sigma_k V_k^T) \simeq X^T X \quad (3)$$

Since columns of X are normalized, diagonal elements of $X^T X$ are all one's, which implies the normalization of columns of $\Sigma_k V_k^T$. Therefore, LSI will preserve the uniform scale if we start with normalized document vectors. For this reason, we believe that documents should be normalized before LSI is applied. We assume so in this paper, without loss of generality. Therefore, cosine similarity is equivalent to dot-product similarity in both spaces. Note that Eq.3 also indicates that the document-document similarities are preserved in the LSI k-dim subspace. [6] further proved that this preservation is an optimal one.

Typically taking $k = 200 - 300$ (far more less than either d, or n), LSI increases the retrieval precision for the query. The optimal k to achieve best precision is currently determined by exhaustive evaluation. How to calculate it directly from X remains an open question[4].

3 Similarity matrices

Our starting point is the understanding of matrices $X^T X$ and $X X^T$, since they determine the SVD and give arise to latent semantic vectors $\mathbf{u}_1 \cdots \mathbf{u}_k$ and $\mathbf{v}^1 \cdots \mathbf{v}^k$. $X^T X$ is the similarity matrix between documents, using the cosine or dot-product

similarity measure in vector-space IR models:

$$\mathrm{sim}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2 \equiv \sum_{\alpha=1}^{d} x_1^\alpha x_2^\alpha \qquad (4)$$

Note the document-document similarity is defined in the space spanned by the d indexing terms (term space). This similarity measure is of fundamental important in vector-space IR models. The dot-product between two term vectors $\mathbf{t}^1, \mathbf{t}^2$ (rows of X):

$$\mathrm{sim}(\mathbf{t}^1, \mathbf{t}^2) = \mathbf{t}^1 \cdot \mathbf{t}^2 = \sum_{i=1}^{n} t_i^1 \cdot t_i^2 \qquad (5)$$

measures their co-occurrences through all documents in the collection, therefore their closeness or similarity[1,2]. XX^T contains dot-products of all pairs of term and is the term-term similarity matrix. In several automatic text categorization methods, terms are often first clustered according their co-occurrences in documents using XX^T . A statistical models of LSI in document space should involve both X^TX and XX^T . We will show later they indeed arise naturally in our model.

Here we emphasize the dual relationship between documents and terms. As discussed above, similarity between documents are defined in term-space, and similarity between terms are defined in document-space. This fundamental relationship between documents and terms naturally corresponds to the occurrence of right and left singular vector in SVD, and is a key feature of our statistical modeling.

4 Dual Probability Model

If we view each document as a data entry in the d-dimensional term-space (index space), there are reasons to believe that documents do not occur entirely randomly. Thus we assume they occur according to certain probability distribution, and can be modeled by standard statistical methods. This idea is similar to, e.g., the Naive Bayes document classification approach where documents are assumed to be governed or generated by a probability distribution.

Consider a column vector $\mathbf{c}$ representing a document, characterizing a Latent semantic structure in LSI. The probability of the occurrence of a document $\mathbf{x}_i$ is related to its similarity (cf. Eq.4) to the latent structure vector $\mathbf{c}$. Motivated by the widespread use of Gaussian distribution, we assume the documents are distributed according to the probability

$$p(\mathbf{x}_i|\mathbf{c}) = e^{(\mathbf{x}_i \cdot \mathbf{c})^2}/Z(\mathbf{c}) \qquad (6)$$

where the normalization constant $Z(\mathbf{c}) = \int \exp(\mathbf{x} \cdot \mathbf{c})^2 d\mathbf{x}$. The next step is to find $\mathbf{c}$ as the optimal parameter for the probability model subject to the constraint $|\mathbf{c}| = 1$. For this purpose, we use the maximum likelihood estimation (MLE), a standard method in statistics. In MLE, we try to find the $\mathbf{c}$ that maximize the following log-likelihood function:

$$\ell(\mathbf{c}) = \log \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{c}) = \mathbf{c}^T XX^T \mathbf{c} - n \log Z(\mathbf{c}) \quad (7)$$

assuming data are independently, identically distributed. Here we have used

$$\sum_{i=1}^{n} (\mathbf{x}_i \cdot \mathbf{c})^2 = \sum_{\alpha,\beta=1}^{d} c^\alpha (XX^T)^{\alpha\beta} c^\beta \qquad (8)$$

The term-term similarity matrix XX^T arises as natural consequence of the model. We point out that it is term similarity matrix XX^T arise here, not the document similarity matrix X^TX (as one might had expected). Here documents are data which live in the index space (term space). XX^T measures the "correlation" between components of data when properly normalized, and would not change much if more data are included, thus serves a role similar to the covariance matrix in a principle component analysis. The conventional covariance matrix on the data set $\mathbf{u}_1 \cdots \mathbf{u}_n$ must subtract the averages and would look qualitatively different from XX^T due to the sparsity of the data matrix X. Thus LSI is not principle component analysis, although they are similar.

In general, finding $\mathbf{c}$ that maximizes $\ell(\mathbf{c})$ involves rather complicated numerical procedure, particularly difficult because $Z(\mathbf{c})$ is an integral in

$d = 10^3 - 10^5$ dimensional space and is analytically intractable. However, note that $n\log Z(\mathbf{c})$ is a very slow changing function in comparing to the first term $\mathbf{c}^T XX^T\mathbf{c}$, and thus can be ignored. In essence, $\mathbf{c}$ is similar to the mean vector μ in Gaussian distribution where the normalization Z is independent of μ. Thus we expect $Z(\mathbf{c})$ to be nearly independent of $\mathbf{c}$.

Therefore, we need only to maximize the first term, $\mathbf{c}^T XX^T\mathbf{c}$. The symmetric positive-definite matrix XX^T has the spectral decomposition: $XX^T = \sum_{\alpha=1}^{d} \lambda_\alpha \mathbf{u}_\alpha \mathbf{u}_\alpha^T$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, here $\lambda_\alpha, \mathbf{u}_\alpha$ are the αth eigenvalue and eigenvector. Thus the optimal solution is $\mathbf{c} = \mathbf{u}_1$.

We can improve the statistical modeling of the data by using k characteristic document vectors, and generalizing Eq.6 to

$$p(\mathbf{x}|\mathbf{c}_1 \cdots \mathbf{c}_k) \propto e^{(\mathbf{x}\cdot\mathbf{c}_1)^2 + \cdots + (\mathbf{x}\cdot\mathbf{c}_k)^2} \qquad (9)$$

with the constraint that they are mutually orthogonal. Following the same maximum likelihood estimation procedure, the optimal solution for model parameters $\mathbf{c}_1 \cdots \mathbf{c}_k$ are the first k eigenvectors of XX^T , $\mathbf{u}_1 \cdots \mathbf{u}_k$, exactly the left singular vectors of LSI/SVD (cf Eq.2). The optimal model is therefore

$$p(\mathbf{x}|\mathbf{u}_1 \cdots \mathbf{u}_k) = e^{(\mathbf{x}\cdot\mathbf{u}_1)^2 + \cdots + (\mathbf{x}\cdot\mathbf{u}_k)^2}/Z_k \qquad (10)$$

where $Z_k = Z(\mathbf{u}_1 \cdots \mathbf{u}_k)$ is the normalization constant.

The above analysis of modeling documents are carried out in term-space. We can also model terms $\mathbf{t}^1 \cdots \mathbf{t}^d$ as defined by their co-occurrences in the document collection, the document space. In this model, the data are the terms, indexed by documents. Consider a (normalized) row vector $\mathbf{r}$ representing a term. Using the term similarity measure Eq.5, we assume $\mathbf{r}$ characterizes the data according to the probability

$$p(\mathbf{t}|\mathbf{r}) = e^{(\mathbf{t}\cdot\mathbf{r})^2}/Z(\mathbf{r}), \qquad (11)$$

similar to Eq.6. To find optimal $\mathbf{r}$, we calculate the log-likelihood,

$$\ell(\mathbf{r}) = \log \prod_{\alpha=1}^{d} p(\mathbf{t}^\alpha|\mathbf{r}) = \mathbf{r}^T X^T X \mathbf{r} - d\log Z(\mathbf{r}) \qquad (12)$$

after some algebra, and noting

$$\sum_{\alpha=1}^{d} t_i^\alpha t_j^\alpha = (X^T X)_{ij} \qquad (13)$$

The document-document similarity matrix $X^T X$ arise again as direct consequence of the model. The symmetric positive-definite matrix $X^T X$ has the spectral decomposition: $A^T A = \sum_{i=1}^{n} \xi_i (\mathbf{v}^i)^T \mathbf{v}^i$, $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_n$, here $\xi_i, \mathbf{v}^i$ are the ith eigenvalue and eigenvector. Thus the optimal solution is $\mathbf{r} = \mathbf{v}^1$.

We may also use k characteristic row vectors to model the data, and the optimal solution is the right singular vectors $\mathbf{v}^1 \cdots \mathbf{v}^k$ of the SVD. Thus we obtain the final probability representation

$$p(\mathbf{t}|\mathbf{v}^1 \cdots \mathbf{v}^k) = e^{(\mathbf{t}\cdot\mathbf{v}^1)^2 + \cdots + (\mathbf{t}\cdot\mathbf{v}^k)^2}/Z_k. \qquad (14)$$

We have constructed a dual probability model, one for documents in term-space using the document similarity, and another for terms in document-space using term similarity. For both models, the optimal solutions for the model parameters are found to be exactly the LSI/SVD vectors. Thus LSI is the optimal solution of the model, and we refer to $\mathbf{u}_1 \cdots \mathbf{u}_k$ and $\mathbf{v}^1 \cdots \mathbf{v}^k$ as latent semantic or index vectors, meaning they identify the latent structures in LSI.

Eqs.10,14 are dual probability representations of the LSI. This dual relationship is further enhanced by the following facts: (a) XX^T and $X^T X$ have the same eigenvalues

$$\lambda_j = \xi_j = \sigma_j^2, \; j = 1, \cdots, k;$$

(b) left and right LSI vectors are related by

$$\mathbf{u}_j = (1/\sigma_j) X \mathbf{v}_j, \; j = 1, \cdots, k.$$

Thus both probability models have the same maximum log-likelihood

$$\ell_k = \sigma_1^2 + \cdots + \sigma_k^2 - n\log Z_k \qquad (15)$$

with the only difference in the normalization constants. This is a direct consequence of dual relationship between terms and documents. In particular, for statistical modeling of the observed term-text co-occurrence data, both probability models

should be considered with same number k, as is the case in the SVD. Eq.15 also indicates that the statistical significance of each LSI vector is proportional to the square of its singular values (σ_i^2 in the likelihood). Therefore, contributions of LSI vectors with small singular values is much smaller than σ_i itself as it appear in the SVD (cf. Eq.2). This is an important result of the theory.

5 Optimal Semantic Subspace

The central theme in LSI is that the LSI subspace captures the essential mainingful semantic associations while reducing redundant and noisy semantic information. Our model provide the means to verify this claim, by measuring the statistical significance of the LSI vectors. We can compute the numerical values of the likelihood and verify that as more latent index vectors are included in the probability density Eq.10, the likelihood of the model increases, indicating the improvement of the quality of statistical model and hence the effectiveness of LSI. We further conjecture that latent index vectors with small eigenvalues contain statistically insignificant information, and their inclusion in the probability density will not increase the the likelihood. In LSI, they represent redundant and noisy semantic information.

Thus the likelihood is the key to verify the existence of the optimal semantic subspace. The log-likelihood for the k latent vectors is

$$\ell_k \equiv \ell(\mathbf{u}_1 \cdots \mathbf{u}_k) = \lambda_1 + \cdots + \lambda_k - n\log Z_k \quad (16)$$

In general, $Z_k \equiv Z(\mathbf{u}_1 \cdots \mathbf{u}_k)$ is difficult to calculate, because it is an integral in a high d-dimensional space ($d = 10^3 - 10^5$):

$$Z_k = \int \cdots \int e^{(\mathbf{x}\cdot\mathbf{u}_1)^2 + \cdots + (\mathbf{x}\cdot\mathbf{u}_k)^2} dx^1 \cdots dx^d.$$

Fortunately in maximum likelihood analysis, what matters is the relative variation of log-likelihood vs k, not the absolute values. To this goal, we may proceed using statistical sampling. In the statistical modeling: data (documents) are samples drawn randomly from the population, thus

$$Z_k \approx \sum_{i=1}^{n} e^{(\mathbf{x}_i\cdot\mathbf{u}_1)^2 + \cdots + (\mathbf{x}_i\cdot\mathbf{u}_k)^2} d\mathbf{x}_i \quad (17)$$

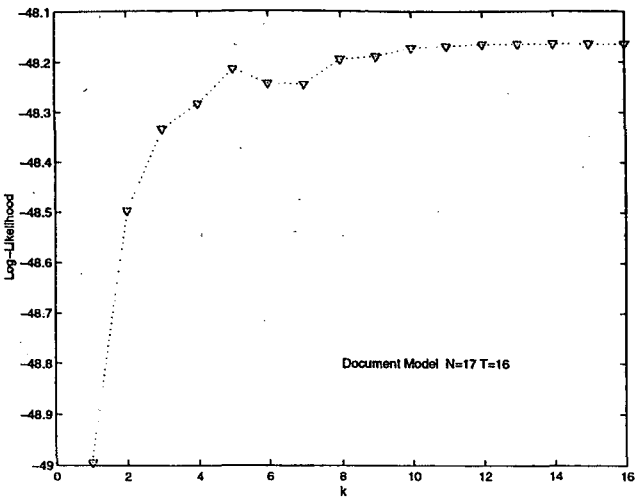

Figure 1: The log-likelihood for modeling documents in term space.

This is an unbiased estimate, and the approximation improves as n increases. If all $d\mathbf{x}_i$ have same size, we can take them out of the sum. In general case, we may also factor them out by a properly defined average $\langle d\mathbf{x}\rangle = (\eta/n)\sum_{i=1}^{n} d\mathbf{x}_i$, where $\eta \sim 1$ and is weakly dependent on k. We may further absorb the difference in the discrete summation (a proportional constant of Eq.17) into $\langle d\mathbf{x}\rangle$, and obtain

$$Z_k = \langle d\mathbf{x}\rangle \sum_{i=1}^{n} e^{(\mathbf{x}_i\cdot\mathbf{u}_1)^2 + \cdots + (\mathbf{x}_i\cdot\mathbf{u}_k)^2} = \langle d\mathbf{x}\rangle \tilde{Z}_k \quad (18)$$

The key point here is that $\langle d\mathbf{x}\rangle$ depends on the given text collection, but independent (or very weakly dependent) of k. In the following likelihood analysis, we will ignore $\langle d\mathbf{x}\rangle$, and compute $\tilde{Z}_k$ only. Thus we have a practical method to calculate Z_k.

5.1 An illustrative example

Here we use a concrete example to illustrate some of the useful concepts. For this goal, we adopt the example of 17 book titles reviewed in SIAM Review[5]. They are indexed by 16 terms, resulting in the 16x17 term-document matrix. After normalizing each document vector (column of X) to 1, we compute the left singular vectors (eigenvectors of XX^T), and the log-likelihood (cf. Eq.16), as shown in Figure 1. The likelihood increases steadily as k

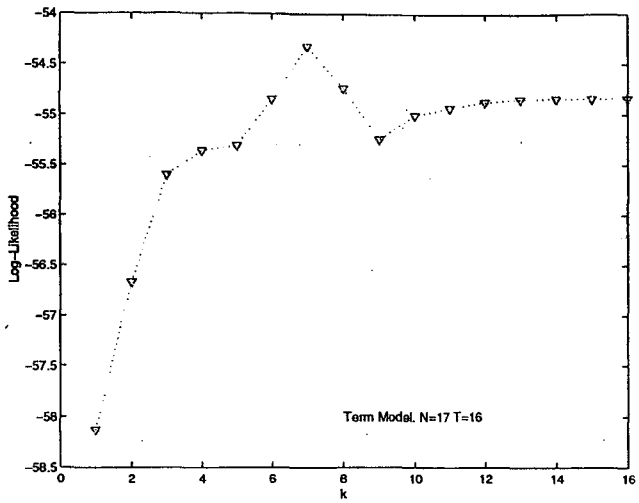

Figure 2: The log-likelihood for modeling terms in document space.

increases from 1 to 5, clearly indicating that the probability model provides better statistical modeling of the documents as more LSI vectors are included; the likelihood fluctuate for $k > 6$, indicating no meaningful statistical information are represented by those latent index vectors with smaller eigenvalues.

To model the terms, we computed the right singular vectors (eigenvectors of X^TX), and the log-likelihood, as shown in Figure 2. One see the likelihood peaked at around $k = 7$ and fluctuate afterwards. These two likelihood curves behave qualitatively similarly, indicating the kind of feature we expect if we believe LSI vectors with small singular values are statistically unimportant. It would be very interesting to repeat these calculations on much large text collections. Clearly the optimal k can be determined by this statistical model. In this collection, $k_{\text{opt}} = 5 \sim 7$.

5.2 Likelihood Analysis

One may ask if the likelihood curves for the book title collection will hold for general cases. After all, as more parameters are included in the model, one would *expect* the likelihood continue to increase. The answer is that even though $Z_k(\mathbf{c})$ changes very slowly indeed, an approximation is still made in finding the optimal analytic solution to Eq.9 in the MLE procedure. Thus the likelihood is not guaranteed to monotonically increase in our model.

Given this clarification, we have some theoretical indications that the likelihood behavior of the book titles example is likely true for general cases. We can prove the following relation

$$\ell_{k+1} \simeq \ell_k \tag{19}$$

for reasonably large k. By "reasonably large k", we mean that in

$$\tilde{Z}_{k+1} = \sum_{i=1}^{n} e^{(\mathbf{x}_i\cdot\mathbf{u}_1)^2+\cdots+(\mathbf{x}_i\cdot\mathbf{u}_k)^2+(\mathbf{x}_i\cdot\mathbf{u}_{k+1})^2}$$

$e^{(\mathbf{x}_i\cdot\mathbf{u}_{k+1})^2}$ is statistically independent of $e^{(\mathbf{x}_i\cdot\mathbf{u}_1)^2+\cdots+(\mathbf{x}_i\cdot\mathbf{u}_k)^2}$ so we can write

$$\begin{aligned}\tilde{Z}_{k+1} &\simeq [\sum_{i=1}^{n} e^{(\mathbf{x}_i\cdot\mathbf{u}_1)^2+\cdots+(\mathbf{x}_i\cdot\mathbf{u}_k)^2}][\frac{1}{n}\sum_{i=1}^{n} e^{(\mathbf{x}_i\cdot\mathbf{u}_{k+1})^2}] \\ &\simeq \tilde{Z}_k(1+\lambda_{k+1}/n),\end{aligned}$$

after expanding $e^{(\mathbf{x}_i\cdot\mathbf{u}_{k+1})^2} \simeq 1 + (\mathbf{x}_i \cdot \mathbf{u}_{k+1})^2$ since $|\mathbf{x}_i \cdot \mathbf{u}_{k+1}| \leq 1$, and using Eq.8. Substituting this into Eq.16 for Z_{k+1}, we obtain Eq.19.

This relation indicates a plateau or a peak in the likelihood curve, instead of a monotonic increase. The theory does not predict whether it will be a peak or a plateau.

6 Invariance Properties

We have outlined the dual model and worked out a few results. Here we mention the invariance properties of the model. First, the model is invariant with respect to (w.r.t.) the order that terms or documents are indexed, since they depend on the dot-product which is invariant w.r.t. the order. The SVD and singular vector and values are also invariant, since they depends on XX^T and X^TX , both of which are invariant w.r.t. the order.

Second, the projections in the k-dim subspace preserve the dot-product similarity. The document projections, columns of $U_k^TX = \Sigma_kV_k^T$, preserve the similarity as shown in Eq.3. The term projections, columns of $V_k^TX^T = \Sigma_kU_k^T$, also preserve the

term-term similarity,

$$(\Sigma_k U_k^T)^T(\Sigma_k U_k^T) = (U_k\Sigma_k V_k^T)(U_k\Sigma_k V_k^T)^T \simeq XX^T \quad (20)$$

up to the minor difference due to the truncation in SVD. In particular, if these quantities are normalized in the original space, they will remain normalized in the LSI subspace.

Third, the model is invariant with respect to incorporating a scale parameter s, an average similarity, in Eq.9,

$$p(\mathbf{x}_i|\mathbf{c}_1\cdots\mathbf{c}_k) \propto e^{[(\mathbf{x}_i\cdot\mathbf{c}_1)^2+\cdots+(\mathbf{x}_i\cdot\mathbf{c}_k)^2]/s^2}, \quad (21)$$

similar to the standard deviation in Gaussian distributions. We obtain same LSI vectors and same likelihood curves except that the vertical scale is enlarged.

7 Related work

Traditional IR probabilistic models, such as the binary independence retrieval model [11, 12] focus on relevance to queries. There, relevance to a specific query is pre-determined or iteratively determined in the relevance feedback, on individual query basis. Our new approach focuses on the data, the term-document matrix X, ignoring query-specific information at present.

As discussed above, similarity matrices XX^T, X^TX are key considerations of our model. X^TX is used as the primary target in the multi-dimensional scaling interpretation[6] of LSI. where it is shown that LSI/SVD is the best approximation to X^TX in the reduced k-dimensional subspace. There, the document-document similarity are also generalized to include arbitrary weighting, which can be similarly carried out in our model.

Minimum description length principle is used in [9] to determine optimal k which is rather close to the experimentally determined value. The relations between the model and the term-document matrix there require further clarifications, however.

8 Concluding remarks

In this paper, we introduced a dual probability model for LSI based on the fundamental cosine (dot-product) similarity measures. Similarity matrices are then direct consequences of the model, and latent semantic vectors of LSI/SVD are the optimal solutions of the model. The latent semantic relationship, as characterized by the latent semantic vectors, are then related to the statistical significance as they are used in characterize (parametrize) the probability distribution. The likelihood is then proposed to quantify this significance. Both the illustrative example and our theoretical understanding (cf. Eq.19) indicate a plateau (or peak) in the likelihood curve. This signals the existence of an optimal semantic subspace with much smaller dimensions that effectively capture the essential associative semantic relationship between terms and documents, consistent with the empirical evidences and the general intuition.

LSI/SVD techniques have been used in information filtering (document classification) and computational linguistics (e.g., [4, 13, 14]). Our model applies to these cases too, as long as the semantic structures defined by the dot-product similarity remains the essential relationship there. In text classification[4, 14], documents are projected into the LSI subspace; the same semantic relations remain in this new feature space as in the retrieval cases. In word sense disambiguation[13], the relevant relationship is the cosine between two vectors in the context space and thus our theory applies here also. In all these cases, it is the appropriate similarity matrix, not the conventional covariance matrix, that determine the meaningful reduced subspace.

The dual probability model outlined here is a constructive model. It can be further extended. One may try to add the query-related information for IR, or other factors relevant for the particular application.

In summary, we believe this model establishes a sound theoretical framework for LSI and LSI/SVD related dimensionality reduction methods, and answer some of the fundamental questions in infor-

mation retrieval and filtering.

Acknowledgement. I thank Drs. Hongyuan Zha, Osni Marques, Horst Simon, Inderjit Dhillon for valuable conversations. This work is supported by the Director, Office of Computational and Technology Research, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract number DE-AC03-76SF00098, and by the Director, Office of Science, Office of Laboratory Policy and Infrastructure, of the U.S. Department of Energy under contract number DE-AC03-76SF00098.

References

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

[2] S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman. Indexing by latent semantic analysis. J.Amer.Soc.Info.Sci, 41(6), pp.391-407. 1990.

[3] S.T. Dumais. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments and Computers, 23(2), 229-236. 1991.

[4] S.T. Dumais. Using LSI for information filtering: TREC-3 experiments. D. Harman (Ed.), Overview of TREC-3, National Institute of Standards and Technology Special Publication, Tech-report 500-335, 1995.

[5] M.W. Berry, S.T. Dumais, G.W. O'Brien. Using linear algebra for intelligent information retrieval. SIAM Review, 37(4), pp.573-595, 1995.

[6] B.T. Bartell, G.W. Cottrell, and R.K. Belew. Representing Documents Using an Explicit Model of Their Similarities J.Amer.Soc.Info.Sci, **46**, 251-271, 1995.

[7] C.H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala. Proc. of Symposium on Principles of Database Systems (PODS), Seattle, Washington, June 1998. ACM Press.

[8] H. Zha, O. Marques and H. Simon. A Subspace-Based Model for Information Retrieval with Applications in Latent Semantic Indexing, Proceedings of Irregular '98, Lecture Notes in Computer Science, Vol. 1457. pp.29-42, 1998, Springer-Verlag.

[9] R.E. Story. An explanation of the effectiveness of latent semantic indexing by means of a Bayesian regression model. Information Processing & Management, 32(03), pp. 329-344. 1996.

[10] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513-523, 1988.

[11] C.J. van Rijsbergen. Informational Retrieval. 2nd Ed. Butterworths. 1979.

[12] N. Fuhr. Probabilistic Models in Information Retrieval. Computer Journal, v.35, pp.243-255, 1992.

[13] H. Schutze. Dimension of meaning. In Proceedings of Supercomputing'92. pp.787-796. IEEE Press. 1992.

[14] Y. Yang. Noise Reduction in a Statistical Approach to Text Categorization. In Proc. of 18th Annual ACM SIGIR Conference (SIGIR'95) 1995:256-263.

ERNEST ORLANDO LAWRENCE BERKELEY NATIONAL LABORATORY
ONE CYCLOTRON ROAD | BERKELEY, CALIFORNIA 94720

Prepared for the U.S. Department of Energy under Contract No. DE-AC03-76SF00098