A Probabilistic Model for Latent Semantic Indexing

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

Abstract

Dimension reduction methods, such as Latent Semantic Indexing (LSI), when applied to semantic space built upon text collections, improve information retrieval, information filtering and word sense disambiguation. A new dual probability model based on the similarity concepts is introduced to provide deeper understanding of LSI. Semantic associations can be quantitatively characterized by their statistical significance, the *likelihood*. Semantic dimensions containing redundant and noisy information can be separated out and should be ignored because their contribution to the overall statistical significance is negative. LSI is the optimal solution of the model. The peak in likelihood curve indicates the existence of an intrinsic semantic dimension. The importance of LSI dimensions follows the Zipf-distribution, indicating that LSI dimensions represent the latent concepts. Document frequency of words follow the Zipf distribution, and the number of distinct words follows log-normal distribution. Experiments on five standard document collections confirm and illustrate the analysis.

Keywords: Latent Semantic Indexing, intrinsic semantic subspace, dimension reduction, word-document duality, Zipf-distribution.

1 Introduction

As computers and Internet become part of our daily life, effective and automatic information retrieval and filtering methods become essential to deal with the explosive growth of accessible information. Many current systems, such as Internet search engines, retrieve information by exactly matching query keywords to words indexing the documents in the database. A well-known problem (Furnas et al., 1987) is the ambiguity in word choices. For example, one searches for "car" related items while missing items related to "auto" (synonyms problem). One looks for "capital" city, and gets venture "capital" instead (polysemy problem). Although both "land preserve" and "open space" express very similar ideas, Web search engines will retrieve two very different sets of webpages with little overlap. These kinds of problems are well-known.

One solution is to manually classify information into different categories by using human judgement. This categorized or filtered information, in essence, reduces the size of the relevant information space and is thus more convenient and useful. Hyperlinks between webpages and anchored words proved useful (Larson, 1996; Li, 1998; Chakrabarti et al., 1998).

A somewhat similar, but automatic approach is to use dimension reduction (data reduction) methods such as the Latent Semantic Indexing (LSI) method (Deerwester et al., 1990; Dumais, 1995; Berry et al., 1995). LSI computes a much smaller semantic subspace from the original text collection, which improves recall and precision in information retrieval (Deerwester et al., 1990; Bartell et al., 1995; Zha et al., 1998; Hofmann, 1999), (Husbands et al., 2004), information filtering or text classification (Dumais, 1995; Yang, 1999; Baker and McCallum, 1998), and word sense disambiguation (Schutze, 1998) (see §3).

The effectiveness of LSI in these empirical studies is often attributed to the reduction of noise, redundancy, and ambiguity. Synonyms and polysemy problems are somehow alleviated in the process. Several recent studies (Bartell et al., 1995; Papadimitriou et al., 1998; Hofmann, 1999; Zha et al., 1998; Dhillon, 2001) shed some lights on this direction (see §9 for detailed discussions).

A central question, however, remains unresolved. Since LSI is a pure "numerical" and automatic procedure, the noisy and redundant semantic information must be associated with a numerical quantity that was reduced or minimized in LSI. But how can one define a quantitative measure for semantic information? How can one verify that this quantitative measure of semantic information is actually reduced or minimized in LSI?

In this paper we address these questions by introducing a new probabilistic model based on document-document and word-word similarities, and show that LSI is the optimal solution of the model (see §5). Furthermore, we use the statistical significance, i.e., the *likelihood*, as a quantitative measure of the LSI semantic dimensions. We calculate likelihood curves of five standard document collections; as LSI subspace dimension k increases (Figure 1), all likelihood curves arise very sharply in the beginning, then gradually turn into a convex peak, and decrease steadily afterwards. This unambiguously demonstrates that the dimensions after the peak contain no *statistically meaningful* information — they represent noisy and redundant information. The existence of these *limited dimension* subspaces that contains the maximum statistically meaningful information, the "intrinsic semantic subspace", provides an explanation of the observed performance improvements for LSI (see §6).

Our model indicates that the statistical significance of LSI dimensions is related to the square of its singular values. For all the five document collections, the statistical significance of LSI dimensions is found to follow a Zipf-law, indicating LSI dimensions represent latent concepts in the same way as webpages, cities, and English words do (see §7).

We further study the word document frequency distribution, which helps to explain the Zipf-law characteristics of LSI dimensions. We also investigate the distribution of distinct

words, which reveals some internal structure of document collections (see §8). Overall, our results provide a statistical framework for understanding LSI type dimension reduction methods. Preliminary results of this work have been presented in conferences (Ding, 1999, 2000).

2 Semantic Vector Space

One of the fundamental relationships in human language is the dual relationship, the mutual interdependence, between words and concepts. Concepts are expressed by the choice of words, while the meanings of words are inferred by their usage in different contexts. Casting this relationship into mathematical form leads to the semantic vector space (Salton and McGill, 1983). A document (title plus abstract or first few paragraphs) is represented by a vector in a linear space indexed by d words (word-basis vector space). Similarly, a word (term) is represented by a vector in a linear space spanned by n document/contexts (document-basis vector space). These dual representations are best captured by the word-to-document association matrix X,

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

where each column $\mathbf{x}_i$ represents a document[1], and each row $\mathbf{t}^\alpha$ represents a word (term)[1]. The matrix entry $x_i^\alpha \equiv (\mathbf{x}_i)^\alpha \equiv (\mathbf{t}^\alpha)_i$. The matrix element x_i^α contains the term frequence (`tf`) of term α occurring in the document i, properly weighted by other factors (Salton and Buckley, 1988). For the common `tf.idf` weighting, $x_i^\alpha = \mathtt{tf}_i^\alpha \cdot \log(1 + n/\mathtt{df}^\alpha)$ where the document frequence $\mathtt{df}^\alpha$ is the number of documents in which the word α occurs.

3 Dimension Reduction: Latent Semantic Indexing

In the *initial* semantic vector space, the word-document relations contain redundancy, ambiguity, and noise — the subspace containing meaningful semantic associations is much smaller than the initial space. One method to obtain the subspace is to perform a dimension reduction (data reduction) to a semantic subspace that contains essential and meaningful associative relations.

LSI is one such dimension reduction method. It automatically computes a *subspace* which contains meaningful semantic associations and is much smaller than the initial space. This

[1]In this paper, capital letters refer to matrices, bold face lower-case letters to vectors: vectors with subscript represent documents and vectors with superscript represent terms; α, β sum over all d terms and i, j sum over all n documents.

is done through the singular value decomposition (SVD) [2] of the term-document matrix:

$$X = \sum_{k=1}^{r} \mathbf{u}_k \sigma_k \mathbf{v}^k = U_r \Sigma_r V_r^T, \tag{2}$$

where r is the rank of the matrix X, $\Sigma_r \equiv \text{diag}(\sigma_1 \cdots \sigma_r)$ are the singular values, $U_r \equiv (\mathbf{u}_1 \cdots \mathbf{u}_r)$ and $V_r \equiv (\mathbf{v}^1 \cdots \mathbf{v}^r)$ are left and right singular vectors. Typically the rank r is order $min(d, n)$, which is about 10,000. However, if we truncate the SVD and keep only the first k largest terms, the resulting $X \simeq U_k \Sigma_k V_k^T$ is a good approximation. Note here, $k \sim 200$ and $r \sim 10,000$, thus a substantial dimensionality reduction. (Good illustrative examples were given in (Deerwester et al., 1990; Berry et al., 1995).)

Document retrieval for a query $\mathbf{q}$ is typically handled by keyword matching, which is equivalent to a dot-product between the query and a document (variable document length is accounted for by normalizing document vectors to unit length). Relevance scores for each of the n documents form a row vector, which are calculated as $\mathbf{s} = \mathbf{q}^T X$. Documents are then sorted according to their relevance scores and returned to user. In a LSI k-dim subspace, a document $\mathbf{x}_i$ is represented by its projection in the subspace, $U_k^T \mathbf{x}_i$, and all n documents $(\mathbf{x}_1 \cdots \mathbf{x}_n)$ are represented as $U_k^T X = \Sigma_k V_k^T$. Queries and mapping vectors are transformed in the same way as documents. Therefore, the score vector in LSI subspace is evaluated as

$$\mathbf{s} = (U_k^T \mathbf{q})^T (U_k^T X) = (\mathbf{q}^T U_k)(\Sigma_k V_k^T). \tag{3}$$

Using these LSI dimensions in document retrieval, both recall and precision are improved, compared to the baseline keyword matching (Deerwester et al., 1990; Bartell et al., 1995; Zha et al., 1998; Hofmann, 1999; Husbands et al., 2004; Caron, 2000).

LSI has also been applied to information filtering (text categorization), such as classifying an incoming news item into predefined categories. An effective method is using the centroid vectors $(\mathbf{c}_1, \cdots, \mathbf{c}_m) \equiv C$ for the m categories (Dumais, 1995). Another method (Yang, 1999) view C as mapping vectors and obtain them by minimizing $||C^T X - B||^2$, where the $m \times n$ matrix B defines the known categories for each document and $||B||^2 = \sum_{i=1}^{n} \sum_{k=1}^{m} (B_i^k)^2$. Using LSI, the centroid matrix or mapping matrix is reduced from $d \times m$ to $k \times m$, reduces the complexity and noise (Dumais, 1995; Yang, 1999). LSI is also effective in Naive Bayes categorization (Baker and McCallum, 1998).

In word sense disambiguation (Schutze, 1998), the calculation involves a *collocation* matrix, where each column represent a context, words that collocate with the target word within a text window. These contexts are then clustered to find different senses of the target word. LSI dimension reduction is necessary to reduce the computational complexity in the clustering process and leads to better disambiguation results (Schutze, 1998).

The usefulness of LSI has been attributed to that the LSI subspace captures the essential associative semantic relationships better than the original document space, and thus par-

[2] Good textbooks on SVD and matrix algebra are (Golub and Loan, 1996; Strang, 1998)

tially resolves the word choice (synonyms) problem in information retrieval, and redundant semantic relationships in text categorization.

Mathematically, LSI with a truncated SVD is the best approximation of X in the reduced k-dim subspace in L_2 norm (Eckart-Young Theorem (Eckart and Young, 1936)). However, the *improved* results in information retrieval and filtering indicate that LSI seems to go beyond mathematical approximation.

From statistical point of view, LSI amounts to an effective dimensionality reduction, similar to the Principal Component Analysis (PCA) in statistics [3] Dimensions with small singular values are often viewed as representing semantic noise and thus are ignored. This generic argument, considering its fundamental importance, needs to be clarified. For example, how small do the singular values have to be in order for the dimensions to be considered noise? A small singular value only indicates that the corresponding dimension is not as important as those with large singular values, but "less important" in itself does not directly imply "noise" or "redundancy".

Thus the question becomes how to quantitatively characterize and measure the associative semantic relationship. If we have a quantitative measure, we can proceed to verify if dimensions with smaller singular values do indeed represent noise.

Directly assigning an appropriate numerical score to each associative relationship appears to be intangible. Instead, we approach the problem with a probabilistic model, and use statistical significance, the *likelihood*, as the quantitative measure for the semantic dimensions. The governing relationship in the probabilistic model is the *similarity* relationship between documents and between words, which we discuss next.

4 Similarity Matrices

It is generally accepted that the dot-product between two document vectors (normalized to 1 to account for different document lengths) is a good measure of the correlation or similarity of word usages in the two documents; therefore the similarity between two documents is defined as

$$\text{sim}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j = \sum_\alpha x_i^\alpha x_j^\alpha = (X^T X)_{ij}. \tag{4}$$

[3] PCA uses the eigenvectors of the covariance matrix $S = \tilde{X}\tilde{X}^T$, where $\tilde{X} = (\mathbf{x}_1 - \bar{\mathbf{x}}, \cdots, \mathbf{x}_n - \bar{\mathbf{x}})$ uses the centered data while LSI does not. Thus the difference between LSI and PCA are tiny: The first principal dimensions (singular vectors) of LSI are nearly identical to row and column means of X, which are subtracted out in PCA. The second principal dimensions of LSI are nearly identical to first principal dimensions of PCA; The third principal dimensions of LSI match the second principal dimensions of PCA; etc.

X^TX contains similarities between all pairs of documents, and is the document - document similarity matrix. Similarly, the dot-product between two word vectors

$$\text{sim}(\mathbf{t}^\alpha, \mathbf{t}^\beta) = \mathbf{t}^\alpha \cdot \mathbf{t}^\beta = \sum_i x_i^\alpha x_i^\beta = (XX^T)^{\alpha\beta} \tag{5}$$

measures their co-occurrences through all documents in the collection, and therefore their closeness or similarity. XX^T contains similarities between all pairs of words and is the word-word similarity matrix. If we assign binary weights to term-document matrix elements x_i^α, one can easily see that XX^T contains the word-word co-occurrence frequency when the context window size is set to the document length.

These similarity matrices define the semantic relationships, and are of fundamental importance in information retrieval (Salton and McGill, 1983). Note that document-document similarity is defined in the word-vector-space, while word-word similarity is defined in document-vector-space. This strong dual relationship between documents and words is a key feature of our model.

5 Dual Probability Model

In recent years, statistical techniques and probabilistic modeling are widely used in IR. Here we propose a probabilistic model to address some of the questions on LSI in §3. Traditional IR probabilistic models (van Rijsbergen, 1979; Fuhr, 1992) focus on relevance to queries. Our approach focuses on the data, the term-document association matrix X. Query-specific information is ignored at present, but may be included in future developments.

Documents are data entries in the d-dimensional word-vector-space; they are assumed to be distributed according to certain probability density function in the probabilistic approach. The form of the density function is motivated by the following considerations: (1) The probability distribution is governed by k characteristic (normalized) document vectors $\mathbf{c}_1 \cdots \mathbf{c}_k$ (collectively denoted as C_k), (2) The occurrence of a document $\mathbf{x}_i$ is proportional to its similarity to $\mathbf{c}_1 \cdots \mathbf{c}_k$. When projecting onto a dimension $\mathbf{c}_j$, $\pm$ signs are equivalent, thus we use $(\mathbf{c}_j \cdot \mathbf{x})^2$ instead of $\mathbf{c}_j \cdot \mathbf{x}$; (3) $\mathbf{c}_1 \cdots \mathbf{c}_k$ are statistically independent factors; (4) Their contribution to total probability for a document is additive. With these considerations, and further motivated by Gaussian distribution, we consider the following probability density function:

$$\Pr(\mathbf{x}_i | \mathbf{c}_1 \cdots \mathbf{c}_k) = e^{(\mathbf{x}_i \cdot \mathbf{c}_1)^2} \cdots e^{(\mathbf{x}_i \cdot \mathbf{c}_k)^2} / Z(C_k) \tag{6}$$

where $Z(C_k)$ is the normalization constant

$$Z(C_k) = \int \cdots \int e^{(\mathbf{x} \cdot \mathbf{c}_1)^2 + \cdots + (\mathbf{x} \cdot \mathbf{c}_k)^2} dx^1 \cdots dx^d. \tag{7}$$

This probabilistic model can be seen as a generalization of the loglinear model (Bookstein

et al., 1992). Consider the case of $k = 1$ and ignore the square of the projection, we have

$$\Pr(\mathbf{x}|\mathbf{c}) = \frac{1}{Z} e^{\mathbf{x}\cdot\mathbf{c}} = \frac{1}{Z} e^{x^1 c^1 + \cdots + x^d c^d} = \frac{1}{Z} {\gamma_1}^{x^1} \cdots {\gamma_d}^{x^d} \tag{8}$$

where $\gamma_\alpha = \exp(c^\alpha)$ is the weight for the word α. Log-linear model ignore the correlation between features (words), i.e., the occurence of each word is independent of other words. Furthermore, if we extend to two factors (characteristic vectors) by using $e^{\mathbf{x}\cdot\mathbf{c}_1 + \mathbf{x}\cdot\mathbf{c}_2} = e^{\mathbf{x}\cdot(\mathbf{c}_1+\mathbf{c}_2)}$, we end up with only one factor $\mathbf{c} = \mathbf{c}_1 + \mathbf{c}_2$. The squaring of $\mathbf{x}\cdot\mathbf{c}_1$ accounts for the correlation between different words; it also make it easy to extend to multiple factors.

Given the form of a probabilistic density function and the data (in the form of matrix X), the maximum likelihood estimation(MLE) is the most widely used method for obtaining the optimal values for the parameters in the density function [4]. (Another method to determine parameters is the *method of moments*, which is used in §8.2.) In our case, the parameters for the probability model are $\mathbf{c}_1 \cdots \mathbf{c}_k$. In the following, we use MLE to determine these parameters. First, we note that assumption (3) requires $\mathbf{c}_1 \cdots \mathbf{c}_k$ to be mutually orthogonal. Assuming $\mathbf{x}_i$ are independently, identically distributed, the (logarithm of) probability *likelihood* of the entirely data under this model is

$$\ell(C_k) \equiv \log \prod_{i=1}^{n} \Pr(\mathbf{x}_i|\mathbf{c}_1 \cdots \mathbf{c}_k) \tag{9}$$

which becomes

$$\ell(C_k) = \sum_{i=1}^{n} [\sum_{j=1}^{k} (\mathbf{x}_i \cdot \mathbf{c}_j)^2 - \log Z(C_k)] = \sum_{j=1}^{k} \mathbf{c}_j^T X X^T \mathbf{c}_j - n \log Z(C_k) \tag{10}$$

after some algebra and noting $\sum_{i=1}^{n} (\mathbf{x}_i \cdot \mathbf{c})^2 = \sum_i (\sum_\alpha x_i^\alpha c^\alpha)(\sum_\beta x_i^\beta c^\beta) = \sum_{\alpha,\beta=1}^{d} c^\alpha (XX^T)^{\alpha\beta} c^\beta$ for any given $\mathbf{c} = \mathbf{c}_j$. Note that in Eq.(10), it is the word-word similarity matrix XX^T (the word co-occurrence matrix) that arises here as a natural consequence of MLE, rather than the document-document similarity matrix that one might have expected. This is because of the dual relationship between documents and words. Rephrasing it differently, documents are data points which live in the index space (word vector space). XX^T measures the "correlation" between components of data points, i.e., correlation between words. When properly normalized, XX^T would not change much if more data points are included, thus serving a role similar to the covariance matrix in principal component analysis. Therefore, understanding document relationship is ultimately related to the understanding of word co-occurrence. Although this fact is known, it is interesting to see its mathematical demonstration as an result of our model.

In MLE, the optimal values of C_k are those that maximizes the log-likelihood $\ell(C_k)$. This usually involves a rather complex numerical procedure, particularly due to the analytically

[4] A good textbook on statistics is (Rice, 1995)

intractable $Z(C_k)$ as a high ($d = 10^3 - 10^5$) dimensional integral. Here we attempt to obtain an approximate optimal solution. This is based on the observation that $n\log Z(C_k)$ is a very slow changing function in comparison to $\sum_j \mathbf{c}_j^T XX^T \mathbf{c}_j$: (1) In essence, $\mathbf{c}_j$ is similar to the mean vector μ in Gaussian distribution, where the normalization constant is independent of μ. Thus $Z(C_k)$ should be nearly independent of $\mathbf{c}_j$. (2) The logarithm of a slow changing function changes even slower. Thus $n\log Z(C_k)$ can be regarded as fixed, and we concentrate on maximizing the first term in Eq.(10).

The symmetric positive-definite matrix XX^T has a spectral decomposition (eigenvector expansion): $XX^T = \sum_{i=1}^r \lambda_i \mathbf{u}_i \mathbf{u}_i^T$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$, here λ_i and $\mathbf{u}_i$ are the ith eigenvalue and eigenvector ($XX^T\mathbf{u}_i = \lambda_i \mathbf{u}_i$). Therefore the optimal solution for characteristic dimensions $\mathbf{c}_1 \cdots \mathbf{c}_k$ in maximizing $\sum_j \mathbf{c}_j^T XX^T \mathbf{c}_j$ are $\mathbf{u}_1 \cdots \mathbf{u}_k$ (for more details, see §4.2.2 in (Horn and Johnson, 1985)). They are precisely the left singular vectors $\mathbf{u}_1 \cdots \mathbf{u}_k$ in SVD of X used in LSI. Thus LSI is the optimal solution of our model, and we will refer to $\mathbf{u}_1 \cdots \mathbf{u}_k$ as LSI dimensions. The final maximal likelihood is

$$\ell(U_k) = \lambda_1 + \cdots + \lambda_k - n\log Z(U_k). \tag{11}$$

We can also model words as defined by their occurrences in all documents, the document vector space. In this space, data points are words and coordinates are documents. Words are represented as row vectors in the word-document matrix X. Consider k (normalized) row vectors $\mathbf{r}^1 \cdots \mathbf{r}^k$ (collectively denoted as R_k) representing k characteristic words. Using the word-word similarity, we assume the probability density function for the occurrence of word $\mathbf{t}^\alpha$ to be

$$\Pr(\mathbf{t}^\alpha | \mathbf{r}^1 \cdots \mathbf{r}^k) = e^{(\mathbf{t}^\alpha \cdot \mathbf{r}^1)^2} \cdots e^{(\mathbf{t}^\alpha \cdot \mathbf{r}^k)^2} / Z(R_k). \tag{12}$$

The log-likelihood becomes

$$\ell(R_k) \equiv \log \prod_{\alpha=1}^d \Pr(\mathbf{t}^\alpha | \mathbf{r}^1 \cdots \mathbf{r}^k) = \sum_{j=1}^k \mathbf{r}^{jT} X^T X \mathbf{r}^j - d\log Z(R_k), \tag{13}$$

after some algebra, and noting $\sum_{\alpha=1}^d t_i^\alpha t_j^\alpha = \sum_{\alpha=1}^d x_i^\alpha x_j^\alpha = (X^T X)_{ij}$. The document-document similarity matrix $X^T X$ arises here. To determine R_k, we maximize the log-likelihood Eq.(13). Following the same line of reasoning for $\ell(C_k)$ of Eq.(10), we see that the second term, $d\log Z(R_k)$, is a slow changing function; thus we only need to maximize the first term in Eq.(13). $X^T X$ has a spectral decomposition: $X^T X = \sum_{\alpha=1}^r \xi_\alpha (\mathbf{v}^\alpha)^T \mathbf{v}^\alpha$, $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_r \geq 0$, here ξ_α and $\mathbf{v}^\alpha$ are the αth eigenvalue and eigenvector, $X^T X \mathbf{v}^\alpha = \xi_\alpha \mathbf{v}^\alpha$. The optimal solution for characteristic words $\mathbf{r}^1 \cdots \mathbf{r}^k$ in maximizing $\sum_j \mathbf{r}^{jT} X^T X \mathbf{r}^j$ are $\mathbf{v}^1 \cdots \mathbf{v}^k$ (see §4.2.2 in (Horn and Johnson, 1985)). By construction, $\mathbf{v}^1 \cdots \mathbf{v}^k$ are precisely the right singular vectors of SVD of X. Therefore, $\mathbf{v}^1 \cdots \mathbf{v}^k$ of LSI are the optimal solution of the document-space model, and the maximal log-likelihood is

$$\ell(V_k) = \xi_1 + \cdots + \xi_k - n\log Z(V_k). \tag{14}$$

Eqs.(6,12) are dual probability representations of the LSI. This dual relation is further enhanced by the facts: (a) XX^T and X^TX have the same eigenvalues

$$\lambda_j = \xi_j = \sigma_j^2,\ j = 1, \cdots, k; \tag{15}$$

(b) left and right LSI vectors are related by

$$\mathbf{u}_j = (1/\sigma_j)X(\mathbf{v}^j)^T, \mathbf{v}_j = (1/\sigma_j)\mathbf{u}_j^T X. \tag{16}$$

Thus both probability representations have the same maximum log-likelihood

$$\ell_k = \sigma_1^2 + \cdots + \sigma_k^2 \tag{17}$$

up to a small and slowly changing normalization constant. This is the direct consequence of the dual relationship between words and documents. In particular, for statistical modeling of the observed word-text co-occurrence data, both probability models should be considered with the same number k, as is the case in the SVD.

Eq.(17) also suggests that the contribution (or the statistical significance) of each LSI dimension is approximately the square of its singular value. This quadratic dependence indicates that LSI dimensions with small singular values are much more insignificant than have been perceived earlier: previously it was generally thought that contributions of LSI dimensions are proportional to singular values linearly, since their singular values appear directly in SVD (cf. Eq.2). Suppose we have two LSI dimensions with singular values 10 and 1 respectively. Compared to the importance of the first dimension, the second dimension is only 1% (rather than 10%) as important. This result gives the first insight as to why one needs to keep only a small number of LSI dimensions and ignore the large number of dimensions with small singular values.

6 Intrinsic Semantic Subspace

The central theme in LSI is that the LSI subspace captures the essential meaningful semantic associations while reducing redundant and noisy semantic information. Our model provides a quantitative mechanism to verify this claim by studying the statistical significance of the semantic dimensions: If a few LSI 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. In other words, if the inclusion of a LSI dimension increases the model likelihood, this LSI dimension represent meaningful semantic relationships. We further conjecture that semantic dimensions 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 key to resolving this central question relies on the behavior of the log-likelihood as a function of k. In the word-space model it is given by Eq.(11). The analytically intractable $Z(U_k) = Z(\mathbf{u}_1 \cdots \mathbf{u}_k)$ can be evaluated numerically by statistical sampling. We generate uniform random numbers in the domain of the integration: on the unit sphere in d-dimensional space, restricted in the positive quadrant. This sampling method converges very quickly. It achieves an accuracy of 4 decimal places with merely 2000 points for $d = 2000 - 5000$ dimensions.

The log-likelihood of LSI word dimensions (defined in document-space model) (cf. Eq.14) can be calculated similarly. Due to the fact that column vectors of X are normalized to 1, the matrix norm of X is $||X||^2 = n$. Thus the normalization of terms, i.e., row vectors, should be

$$||\mathbf{t}^\alpha||^2 = \sum_{i=1}^{n} (x_i^\alpha)^2 = n/d, \ \alpha = 1, \cdots, d, \tag{18}$$

and the domain of integration is the positive quadrant on the sphere of radius $\sqrt{n/d}$ in n-dimensional document space.

Likelihood curves are calculated for five standard test document collections in IR: CRAN (1398 document abstracts on Aeronautics from Cranfield Institute of Technology), CACM (3204 abstracts of articles in *Communications of ACM*), MED (1033 abstracts from National Library of Medicine), CISI (1460 abstracts from Institute of Scientific Information). and NPL (11429 titles from National Physical Laborotry). In the term-document matrices we use the standard term frequency- inverse document frequency (tf.idf) weighting. The calculated likelihood curves are shown in Figures 1 and 2.

For all five collections, both word-space and document-space likelihoods grow rapidly and steadily as k increases from 1 up to $k_{\rm int}$, clearly indicating that the probability models provide better and better statistical descriptions of the data. They reach a peak at $k_{\rm int}$. However, starting from $k > k_{\rm int}$, the likelihood decreases steadily, indicating no meaningful statistical information is represented by those LSI dimensions with smaller eigenvalues. For all four collections, the intrinsic dimensions determined from document-space, $k_{\rm int}^{(u)}$, and from word-space, $k_{\rm int}^{(v)}$, are fairly close as indicated in Figure 1.

Note that the theoretical $k_{\rm int}$ from the likelihood curve for CACM is quite close to that experimentally determined for text classification (Yang, 1999). For Medline, however, $k_{\rm int}^{(v)}$ is larger than the experimentally determined value (Deerwester et al., 1990; Zha et al., 1998), based on the 11-point average precision for 30 standard queries. Since the model contains no information on the queries, these reasonable agreements indicate that the statistical model and the statistical significance-based arguments capture some essential relationships involved.

Overall, the general trend for the five collections is quite clear. These likelihood curves quantitatively and unambiguously demonstrate the existence of an intrinsic semantic subspace: dimensions with small eigenvalues do represent redundant or noisy information, and contributes negatively to the statistical significance. This is one of the main results of this

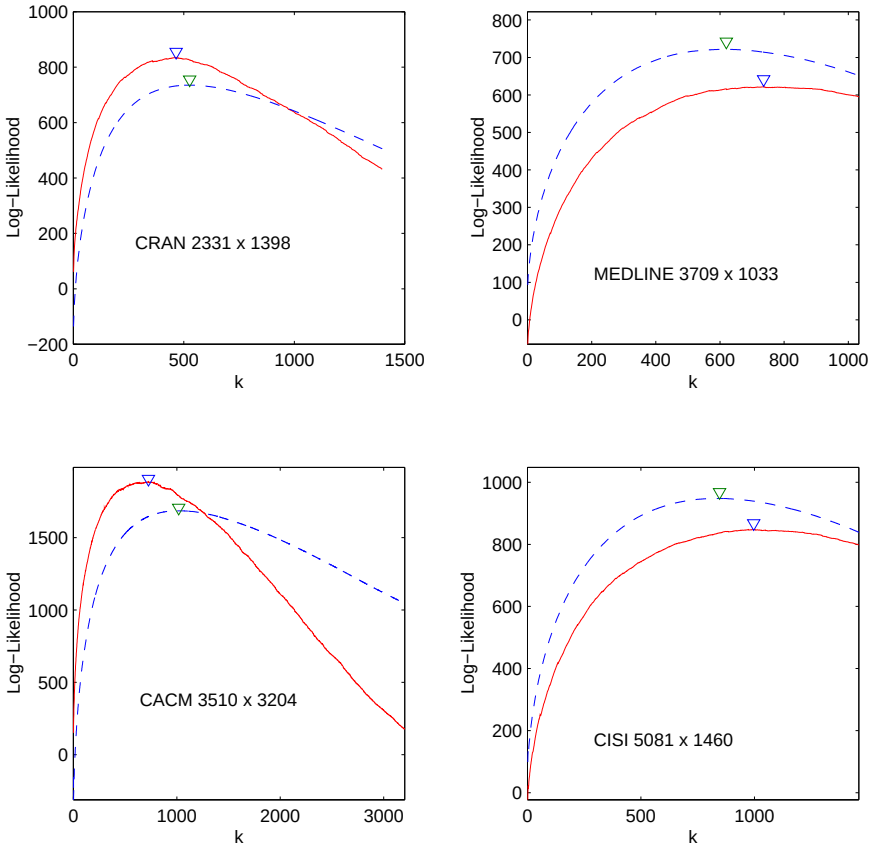

Figure 1: Log-likelihood curves for CRAN, Medline, CACM and CISI collections. Solid lines for moduling documents in word space, $\ell(U_k)$, and dashed lines for moduling words in document space, $\ell(V_k)$. The intrinsic semantic subspace dimension, $k_{\text{int}}^{(u)}$ for word space and $k_{\text{int}}^{(v)}$ for document space, are also indicated. Number of words d and number of documents n for each collection is given after the collection name.

work.

7 Do LSI Dimensions Represent Concepts?

LSI dimensions are optimal solutions for the characteristic document vectors introduced in the dual probability model. Note that the similarity relationship, i.e., the dot-product, can also be viewed as projection of document $\mathbf{x}_i$ onto characteristic vector $\mathbf{c}_j$ (see Eq.6). Thus LSI dimensions are actually projection directions, which is obvious from the SVD point of view.

Besides the projection directions, do these LSI dimensions represent something about the document collection? Or equivalently, do the projection directions mean something? As explained in the original paper (Deerwester et al., 1990), the exact meaning of those

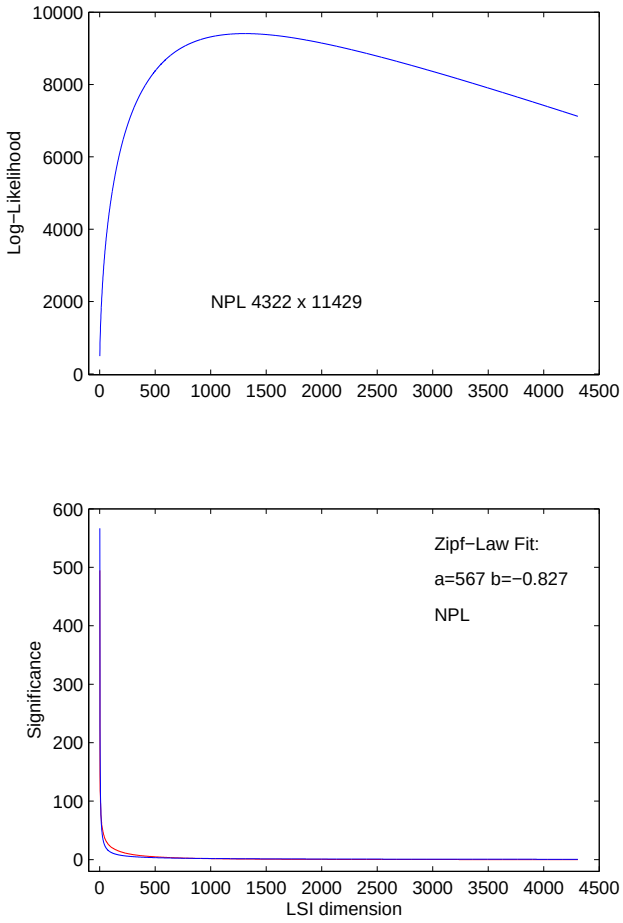

Figure 2: Log-likelihood (in word space) and statistical significance for NPL.

LSI dimensions are complex and can not directly inferred (thus named "latent" semantic indexing).

Our probability model provides additional insights to this issue. The statistical significance or importance of each LSI dimension directly relates to its singular value squared (cf. Eq.17). They are calculated for all five document collections and shown in Figures 2 and 3.

The statistical significance of LSI dimensions clearly follow a Zipf law, $\sigma_i^2 = a \cdot i^b$, with the exponent b very close to -1. These fits are very good: the data and the fits are almost indistinguishable for all 4 collections. In addition, Zipf-law also fits well for NPL and TREC6 collections (Husbands et al., 2004). We conjecture that the Zipf-law is obeyed by singular values squared of most if not all document collections.

Zipf-law (Zipf, 1949) is the observation that frequency of occurrence f, as a function of the rank i, is a power-law function

$$f_i = a \cdot i^{\ b} \tag{19}$$

with the exponent b close to -1. There are a wide range of social phenomena that obey Zipf-

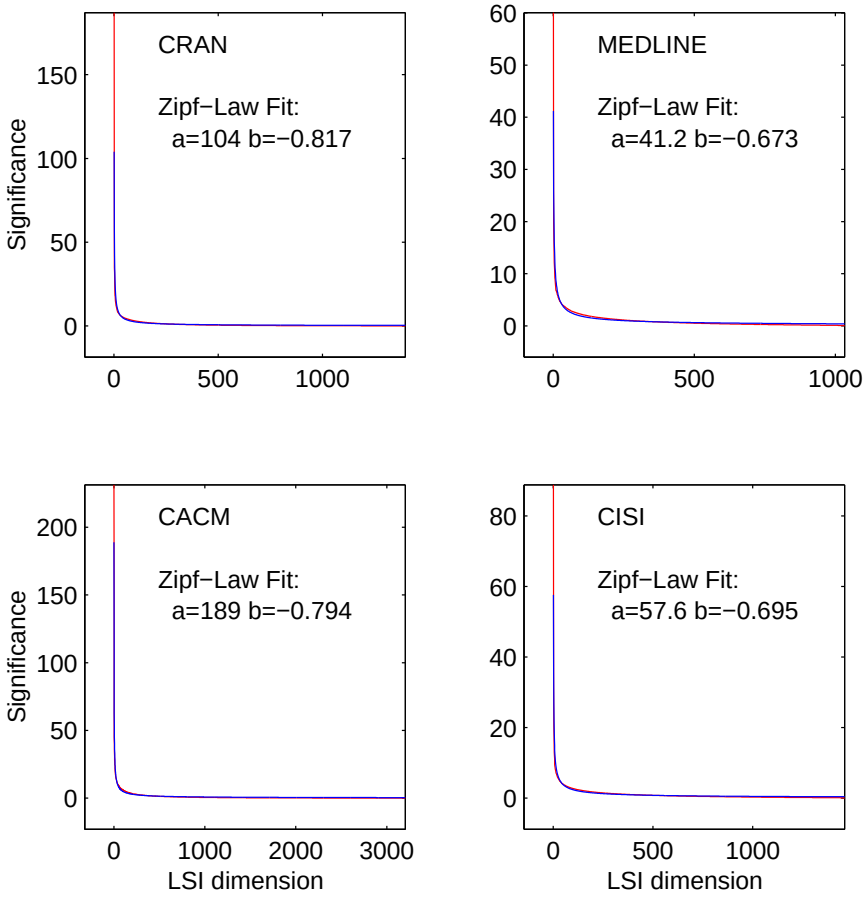

Figure 3: Statistical significance σ_i^2 of the LSI/SVD dimensions for CRAN, MEDLINE, CACM and CISI. The Zipf-law with only two parameters fits the data extremely well: the original data and fit are essentially indistinguishable.

law. The best known example is the frequency of word usage in English and other languages. Ranking all cities in the world according to their population, they also follow Zipf-law. Most recently on the Internet, if we rank the website by their popularity, the number of user visits, the webpage popularity also obey the Zipf-law (Glassman, 1994).

One common theme among English words, cities, webpages, etc, is that each one has distinct characters or identities. Since LSI dimensions on all document collections display very clear Zipf-distribution, we may infer that LSI dimensions represent some latent concepts or identities in a similar manner as English words, cities, or webpages do. However, the exact nature of LSI dimensions remains to be explored (see§10.4).

8 Characteristics of Document Collections

To provide a perspective on the statistical approach discussed above, we further investigate the characteristics of the four document collections. There are many studies on statistical distributions in the context of natural language processing (see (Manning and Schuetze, 1999) and references there). One of the emphasis there is the word frequency distribution in documents, ranging from the simple Poisson distribution to the K-mixture distribution (Katz, 1996). Here we studied the two distributions that have close relations to the probablistic model discussed above.

8.1 Document Frequency Distribution

We first study the document frequency (`df`) of each word, i.e., the number of document a word occurs in. In the term-document matrix X, this corresponds to the number of nonzero elements in each row.

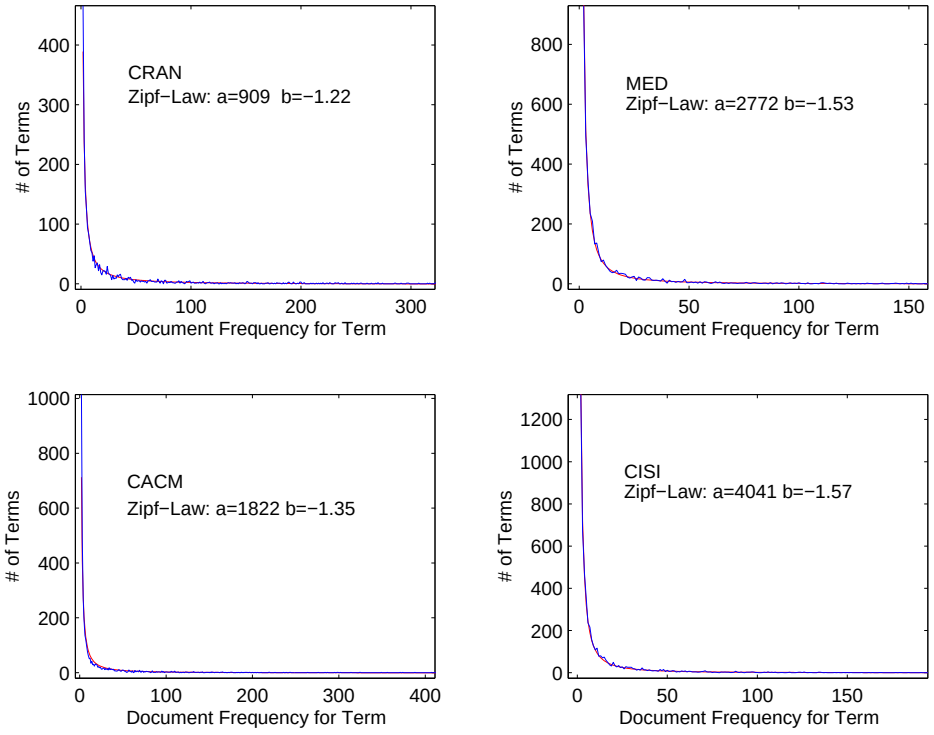

Figure 4: Distributions of document frequency for terms in the four collections. The Zipf-law fits are also shown.

In Figure 5, we show the distribution of document frequency for four document collections. Plotted are the number of words at a given document frequency, the histogram. For all four collections, there are large number of words which have small `df`. For example, for

CRAN, there are 466 words with df=2, 243 words with df=3, etc. On the other hand, the number of words with large df is small. There are only a total number of 186 words which has df $\geq$ 100, although df can reach as high as 860 for one word.

From earlier discussions, this kind of distribution can be described by the Zipf distribution (more precisely, the power law),

$$N(\texttt{df}) = a \cdot \texttt{df}^b$$

In fact, the Zipf-law fits data very well for all four collections, as shown in Figure 5. The exponents are generally close to -1. Zipf-law originally describe the frequency of word usage; Here we see that Zipf-law also governs the document frequence of *content* words.

The distribution of document frequency gives a better understanding of the LSI dimensions discussed in previous section. Since LSI dimensions are essentially linear combinations of the words, we may say that the Zipf-law behavior of the words directly implies that the statistical significance of the LSI dimensions also follow the Zipf-law. This analogy further strengthen our previous arguments that LSI dimensions represent latent concepts, in much the same way as the indexing words do.

In the literature, the average document frequency is often quoted as a numerical characterization of the document collection. For Gaussian type distributions, the mean (average) is the center of the bell-shaped curve and is a good characterization of the distribution; however, for scale-free Zipf type of distributions, the mean does not capture the essential features of the distribution. Whether df=1 words are included or not will change the mean quite significantly since they dominate the averaging process; but the Zipf-curve will not change much at all. For this reason, parameters a, b are better characteristic quantities for document frequences since they uniquely determine their distribution. a, b also have clear meaning: b is the exponent that governs the decay; a is the *expected* number of words with df=1 according to the Zipf-curve. The fact that we can know the expected number of df=1 words without actually counting df=1 words indicates the value of the analysis of document frequence distribution.

Since document frequency is very often used as the global weighting for document representation in the vector space (as in tf.idf weighting), knowing their distribution will help to understand the effects of weighting and to further improve the weighting.

8.2 Distribution of Distinct Words

Next we investigate the number of distinct words (terms) in each document. This is the number of nonzero elements in each column in the term-document matrix X. In Figure 5, we plot the distribution of this quantity, i.e., the number of documents for a given number of distinct words. For the Cranfield collection, the minimum number of distinct words for a document is 11 (document # 506), and the maximum is 155 (document # 797). The peak point (40, 39) in the histogram indicates there are 39 documents in the collection, each of

which has 40 distinct words.

This leads to a distribution very different from the Zipf distribution for document frequency above. The distribution appears to follow a log-normal distribution (Hull, 2000), which has the following probability density function

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right) \tag{20}$$

with mean and variance:

$$\bar{x} = e^{\mu+\sigma^2/2}, \;\; v = \langle (x - \bar{x})^2 \rangle = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1). \tag{21}$$

Calculating the mean $\bar{x}$ and variance v directly from the data, we can solve Eq.(21) to obtain the parameters μ, σ. The probability density function can be drawn (the smooth curves in Figure 4). This simple procedure provides a very good fit to the data.

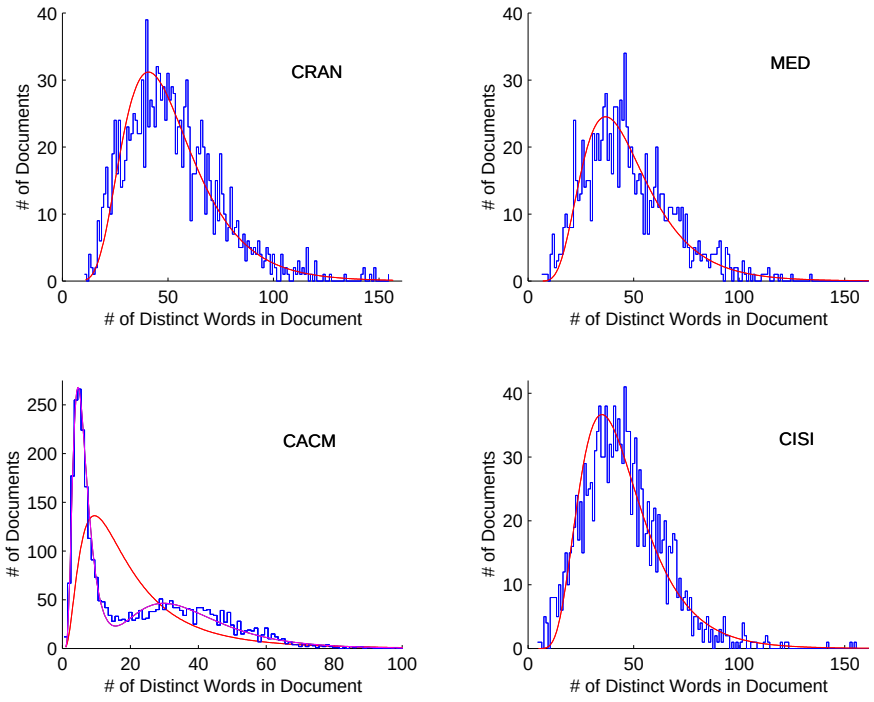

Figure 5: Distribution of distinct words for four document collections. Smooth curves are from the log-normal distribution (see text).

For three document collections, Cranfield, Medline, and CISI, the number of distinct words follow log-normal distributions. Normally, we expect this quantity to follow a normal (Gaussian) distribution. However, due to the fact that the quantity is a non-negative count variable, we expect the logarithm of the variable to follow a Gaussian distribution with mean μ and variance σ^2. This leads to the log-normal distribution for the original variable.

At first look, the histogram for CACM seems to deviate substantially from the log-normal distribution, as shown by the smooth single peak curve determined by the mean and variance of the entire data points in CACM. However, a careful examination of the two-peak pattern indicates that each of them can be fitted by a log-normal distribution; the second peak part is quite similar to those of all three other collections. In fact, a simple fit of the CACM histogram with two log-normal distributions $p(x) = w_1 p_1(x) + w_2 p_2(x)$ is also shown in Figure 4, with $\mu_1 = 1.75, \sigma_1 = 0.472$ and $\mu_2 = 3.57, \sigma_2 = 0.415$. The fit is quite good. From the weights w_1, w_2, we can calculate the number of documents in each log-normal distribution. The calculated values are 1633 docs for the left peak part, and 1571 docs for the right peak part.

The CACM collection consists of titles and abstract of articles in *Communications of ACM*. However, out of the 3204 CACM documents, 1617 documents contain titles only, and therefore have much less number of distinct words per document (around 5). The remaining 1587 documents contain both a title and an abstract, therefore have number of distinct words around 30-40. Our statistical analysis automatically picks up the substantial difference and gives very close estimates of documents in each category: 1633 vs 1617 for the title-only documents and 1571 vs 1587 for the title+abstract documents. This indicates the usefulness of statistical analysis of document collections.

9 Related Work

With the contexts and notations provided above, we give pertinent descriptions of related work on probabilistic interpretation of LSI. Traditional IR probabilistic models, such as the binary independence retrieval model (van Rijsbergen, 1979; Fuhr, 1992) 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 term-document matrix using a probabilistic generative model. This occurrence probability could also be used in the language modeling approach for IR (Ponte and Croft, 1999).

Similarity matrices XX^T and X^TX are key considerations of our model. X^TX is used as the primary goal in the multi-dimensional scaling interpretation[6] of LSI where it is shown that LSI is the best approximation to X^TX in the reduced k-dimensional subspace. There, the document-document similarity is also generalized to include arbitrary weighting, which improved the retrieval precision.

If the first k singular values of SVD are well separated from the rest, i.e., σ_i has a sharp drop near $i = k$, the k-dim subspace is proved to be stable against smaller perturbations in (Papadimitriou et al., 1998; Azar et al., 2001). A probabilistic corpus model built upon k topics is then introduced and is shown to be essentially the LSI subspace (Papadimitriou et al., 1998). Our calculations in Figure 2 show that singular values obey Zipf distribution which drop off steadily and graduately for all k.

Introducing a latent class variable in the joint probability for P(document,word), the resulting probability of the aspect model (Hofmann, 1999) follows the chain rule and can be written quite similarly as $U\Sigma V^T$ of the SVD. This probabilistic LSI formalism is further developed to handle queries and its effectiveness is shown. The latent class and the concept vector/index have many common features. This is further analyzed in (Azzopardi et al., 2003).

A subspace model using the low-rank-plus-shift structured is introduced in (Zha et al., 1998) and lead to a relation to determine optimal subspace dimension $k_{\rm int}$ from the singular values. The relation was originally developed for array signal processing using minimum description length principle.

Using a spherical K-means method for clustering documents (Dhillon, 2001) leads to the concept vectors (centroids of each clusters), which are compared to LSI vectors. The subspace spanned by the concept vectors are close to the LSI subspace. This method is further developed into concept indexing (Karypis and Han, 2000).

A "dimension equalization" of LSI is proposed in (Jiang and Littman, 2000) and developed into trans-lingual method document retrival. Another development is iterative scaling of LSI, which appears to improve the retrieval precision (Ando and Lee, 2001). The determination of the optimal dimension were examined in many of aboved mentioned studies and are also investigated in (Story, 1996; Efron, 2002; Dupret, 2003).

One advantage of LSI is the reduced storage kd on the database, by storing only a small number of k singular vectors, rather than the original term-document matrix. However, the original term-document matrices in information retrieval are usually very sparse (the fraction of nonzeros $f < 1\%$). Thus the breakeven point is $kd \leq fdN$ or $k \leq fN$. For $k \simeq fN$, there is no storage saving for LSI. Several recently developed methods further significantly reduce the storage of singular vectors by either using a discrete approximation (Kolda and O'Leary, 1998) or thresholding on values of the LSI/SVD vectors (Zhang et al., 2002).

10 Discussions

In this paper, we focus on term-document association X, and study four distributions, the distributions of the the columns (documents) and rows (words) of X, and the distributions of the number of nonzero elements in each row (document frequency) and the number of nonzero elements in each column (number of distinct words in each document). Here we point out several more features on these distributions.

10.1 Invariance Properties

LSI has several invariance properties. First, the model is invariant with respect to (w.r.t.) the order that words or documents are indexed, since they depend on the dot-product which is invariant w.r.t. the order. The singular vectors 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 similarity relations between documents and between words are preserved in the k-dim LSI subspace. In the LSI subspace, documents are represented as their projections, i.e., columns of $U_k^T X = \Sigma_k V_k^T$; words are represented as the rows of $XV_k = U_k\Sigma_k$. The document-document similarity matrix in the LSI subspace is

$$(\Sigma_k V_k^T)^T(\Sigma_k V_k^T) = (U_k\Sigma_k V_k^T)^T(U_k\Sigma_k V_k^T) \simeq X^TX, \tag{22}$$

up to the minor difference due to the truncation in SVD. Similarly, the term-term similarity matrix in LSI subspace is

$$(U_k\Sigma_k)(U_k\Sigma_k)^T = (U_k\Sigma_k V_k^T)(U_k\Sigma_k V_k^T)^T \simeq XX^T, \tag{23}$$

up to the minor difference due to the truncation in SVD.

Note that the self similarities, i.e., the diagonal elements in the similarity matrices, are the length of the vectors (L_2 norm). Thus, if document vectors are normalized in the original space, they remain approximately normalized in the LSI subspace. For this reason, we believe that documents should be normalized before LSI is applied, to provide a consistent view.

Third, the probabilistic model is invariant with respect to a scale parameter s, an average similarity, which could be incorporated in Eq.(6) as,

$$\Pr(\mathbf{x}_i|\mathbf{c}_1 \cdots \mathbf{c}_k) \propto e^{[(\mathbf{x}_i\cdot\mathbf{c}_1)^2+\cdots+(\mathbf{x}_i\cdot\mathbf{c}_k)^2]/s^2}, \tag{24}$$

similar to the standard deviation in Gaussian distributions. We can repeat the analysis in Section 5, and obtain the same LSI dimensions and same likelihood curves except that the vertical scale is enlarged or shrinked depending on $s > 1$ or $s < 1$.

10.2 Normalization Factors

In this paper, we started with documents (columns of X) that are normalized: $||\mathbf{x}_i|| = 1$. This implies that the cosine similarity is equivalent to the dot-product similarity between documents:

$$\text{sim}_{\cos}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2/||\mathbf{x}_1|| \cdot ||\mathbf{x}_2|| = \text{sim}_{\text{dot}}(\mathbf{x}_1, \mathbf{x}_2).$$

However, the normalization of columns does not imply that each term (rows of X) are normalized to $\sqrt{n/d}$ (see Eq.18), although they do so on average.

This implies that for words, the dot-product similarity is not equivalent to cosine similarity. This is not a serious problem in itself, since dot-product similarity is a well-defined

similarity measure. Furthermore, columns and rows can be normalized simultaneously, by alternatively normalizing rows and columns. We can prove that this process will converge to a unique final results, independent of whether we first normalize row or columns. Afterwards, for terms $\mathbf{t}^1, \mathbf{t}^2$ we have

$$\text{sim}_{\cos}(\mathbf{t}^1, \mathbf{t}^2) = \mathbf{t}^1 \cdot \mathbf{t}^2 / ||\mathbf{t}^1|| \cdot ||\mathbf{t}^2|| = (n/d) \cdot \mathbf{t}^1 \cdot \mathbf{t}^2 = (n/d) \cdot \text{sim}_{\text{dot}}(\mathbf{t}^1, \mathbf{t}^2),$$

showing that the dot-product similarity is the same as the cosine similarity (the proportional constant (n/d) will not change ranking and is thus irrelevant).

10.3 Separation of Term and Document Representations

In the LSI subspace, documents and words are represented by their projections $(\mathbf{u}_1, \cdots, \mathbf{u}_k)$ and $(\mathbf{v}^1, \cdots, \mathbf{v}^k)$. The dual relationship between them is no longer directly represented as rows and columns of the *same* matrix. Instead, they are related through a *filtered* procedure, $\mathbf{u}_j = (1/\sigma_j) X \mathbf{v}_j$. This filtering process can be regarded as a learning process: from several contexts, the meaning of a word is better described by a number of filtered contexts, instead of the original raw contexts.

10.4 Cluster Indicator Interpretation of LSI Dimensions

The meaning of LSI dimension are discussed at length in (Landauer and Dumais, 1997) and briefly in §7. A recent progress (Zha et al., 2002; Ding and He, 2003) on K-means clustering leads to a new interpretation of LSI dimensions. The widely adopted K-means clustering (Hartigan and Wang, 1979) minimizes the sum of squared errors,

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \mu_k)^2 = \sum_i \mathbf{x}_i^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \mathbf{x}_i^T \mathbf{x}_j, \tag{25}$$

where $\mu_k = \sum_{i \in C_k} \mathbf{x}_i / n_k$ is the centroid of cluster C_k and n_k is the number of documents in C_k. The solution of clustering can be represented by K cluster membership indicator vectors: $H_K = (\mathbf{h}_1, \cdots, \mathbf{h}_K)$, where

$$\mathbf{h}_k = (0, \cdots, 0, \overbrace{1, \cdots, 1}^{n_k}, 0, \cdots, 0)^T / n_k^{1/2} \tag{26}$$

In Eq.(25), $\sum_i \mathbf{x}_i^2$ is a constant. The second term can be written as $J_h = \mathbf{h}_1^T X^T X \mathbf{h}_1 + \cdots + \mathbf{h}_k^T X^T X \mathbf{h}_k$, which is to be maximized. Now we relax the restriction that $\mathbf{h}_k$ takes discrete values of $\{0, 1\}$ and let $\mathbf{h}_k$ take continuous values. The solution for the maximizing J_h is given by the principal eigenvectors of $X^T X$, according to a well-known Theorem (Fan, 1949). In other words, LSI dimensions (eigenvectors of $X^T X$) are the continuous solutions to the cluster membership indicators in K-means clustering problem. In unsupervised learning, cluster represent concepts — we may say that LSI dimensions represent concepts.

11 Summary

In this paper, we introduce a dual probabilistic generative model based on similarity measures. Similarity matrices arise naturally during the maximum likelihood estimation process, and LSI is the optimal solution of the model, via maximum likelihood estimation.

Semantic associations characterized by the LSI dimensions are measured by their statistical significance, the likelihood. Calculations on four standard document collections exhibit a maximum in likelihood curves, indicating the existence of an limited-dimension intrinsic semantic subspace. The importance (log-likelihood) of LSI dimensions follows a Zipf-like distribution.

The term-document matrix is the main focus of this study. The number of nonzero elements in each row of the matrix, the document frequency, follows the Zipf-distribution. This is the direct reason that the statistical significance of LSI dimensions follow the Zipf law. The number of nonzero elements in each column of the matrix, the number of distinct words, follows a log-normal distribution and gives useful insights to the structure of the document collection.

Besides automatic information retrieval, text classification, and word sense disambiguation, our model can apply to many other areas, such as image recognition and reconstruction, as long as the relevant structures are essentially characterized or defined by the dot-product similarity. Overall, the model provides a statistical framework upon which LSI and similar dimension reduction methods can be analyzed.

Beyond information retrieval and computational linguistics, LSI is used as the basis for a new theory of knowledge acquisition and representation (Landauer and Dumais, 1997) in cognitive science. Our results that LSI is an optimal procedure and that the intrinsic semantic subspace is much smaller than the initial semantic space lend better understanding and support to that theory.

Acknowledgements. The author thanks Hongyuan Zha for providing term-document matrices used in this study and for motivating this research, Parry Husbands for help computing SVDs of large matrices, Zhenyue Zhang, Osni Marques and Horst Simon for valuable discussions, Micheal Berry and Inderjit Dhillon for seminars given at NERSC/LBL that help motivated this work, and Dr. Susan Dumais for communications. He also thanks an anounymous referee for suggesting the connection to the log-linear model. This work is supported by Office of Science, Office of Laboratory Policy and Infrastructure, of the U.S. Department of Energy under contract DE-AC03-76SF00098.

References

Ando, R. and Lee, L. (2001). Iterative residual rescaling: An analysis and generalization of LSI. *Proc. ACM Conf. on Research and Develop. IR(SIGIR)*, pages 154–162.

Azar, Y., Fiat, A., Karlin, A., McSherry, F., and Saia, J. (2001). Spectral analysis for data mining. *Proc. ACM Symposium on Theory of Computing, Crete*, pages 619–626.

Azzopardi, L., Girolami, M., and van Risjbergen, K. (2003). Investigating the relationship between language model perplexity and ir precision-recall measures. *Proc. ACM Conf. Research and Develop. Info. Retrieval (SIGIR)*, pages 369–370.

Baker, L. and McCallum, A. (1998). Distributional clustering of words for text classification. *Proc. ACM Conf. on Research and Develop. Info. Retrieval (SIGIR)*.

Bartell, B., Cottrell, G., and Belew, R. (1995). Representing documents using an explicit model of their similarities. *J.Amer.Soc.Info.Sci,* **46***, 251-271, 1995*, pages 251–271.

Berry, M., Dumais, S., and O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573–595.

Bookstein, A., O'Neil, E., Dillon, M., and Stephens, D. (1992). Applications of loglinear models for informetric phenomena. *Information Processing and Management*, 28.

Caron, J. (2000). Experiments with lsa scoring: Optimal rank and basis. *Proc. SIAM Workshop on Computational Information Retrieval, ed. M. Berry*.

Chakrabarti, S., Dom, B. E., Raghavan, P., Rajagopalan, S., Gibson, D., and Kleinberg, J. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks and ISDN Systems*, 30:65–74.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and Harshman, R. (1990). Indexing by latent semantic analysis. *J. Amer. Soc. Info. Sci*, 41:391–407.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. *Proc. ACM Int'l Conf Knowledge Disc. Data Mining (KDD 2001)*.

Ding, C. (1999). A similarity-based probability model for latent semantic indexing. *Proc. 22nd ACM SIGIR Conference*, pages 59–65.

Ding, C. (Oct. 2000). A probabilistic model for latent semantic indexing in information retrieval and filtering. *Proc. SIAM Workshop on Computational Information Retrieval, ed. M. Berry*, pages 65–74.

Ding, C. and He, X. (2003). K-means clustering and principal component analysis. *LBNL Tech Report 52983*.

Dumais, S. (1995). Using lsi for information filtering: Trec-3 experiments. *Third Text REtrieval Conference (TREC3), D Harman, Ed, National Institute of Standards and Technology Special Publication*.

Dupret (2003). Latent concepts and the number orthogonal factors in latent semantic analysis. *Proc. ACM Conf. on Research and Develop. IR(SIGIR)*, pages 221–226.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1:183–187.

Efron, M. (2002). Amended parallel analysis for optimal dimensionality reduction in latent semantic indexing. *Univ. N. Carolina at Chapel Hill, Tech. Report TR-2002-03*.

Fan, K. (1949). On a theorem of Weyl concerning eigenvalues of linear transformations. *Proc. Natl. Acad. Sci. USA*, 35:652–655.

Fuhr, N. (1992). Probabilistic models in information retrieval. *Computer Journal*, 35:243=255.

Furnas, G., Landauer, T., Gomez, L., and Dumais, S. (1987). The vocabulary problem in human-system communications. *Communications of ACM*, 30:964–971.

Glassman, S. (1994). A caching relay for the world wide web. *Comput. Networks ISDN System*, 27:165–175.

Golub, G. and Loan, C. V. (1996). *Matrix Computations, 3rd edition*. Johns Hopkins, Baltimore.

Hartigan, J. and Wang, M. (1979). A K-means clustering algorithm. *Applied Statistics*, 28:100–108.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proc. ACM Conf. on Research and Develop. IR(SIGIR)*, pages 50–57.

Horn, R. and Johnson, C. (1985). *Matrix Analysis*. Cambridge University Press.

Hull, J. (2000). *Options, Futures, and other Derivatives*. Prentice Hall.

Husbands, P., Simon, H., and Ding, C. (2004). Term norm distribution and its effects on latent semantic indexing. *To appear in Information Processing and Management*.

Jiang, F. and Littman, M. (2000). Approximate dimension equalization in vector-based information retrieval. *Proc. Int'l Conf. Machine Learning*.

Karypis, G. and Han, E.-H. (2000). Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. *Proc. 9th Int'l Conf. Information and Knowledge Management (CIKM 2000)*.

Katz, S. (1996). Distribution of content words and phrases in text and language modeling. *Natural Language Engineering*, 2:15–60.

Kolda, T. and O'Leary, D. (1998). A semi-discrete matrix decomposition for latent semantic indexing i n information retrieval. *ACM Trans. Information Systems*, 16:322–346.

Landauer, T. and Dumais, S. (1997). A solution to plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240.

Larson, R. R. (1996). Bibliometrics of the world wide web: an exploratory analysis of the intellectual structures of cyberspace. *Proc. SIGIR'96*.

Li, Y. (1998). Towards a qualitative search engine. *IEEE Internet Computing*, 2:24–29.

Manning, C. and Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA.

Papadimitriou, C., Raghavan, P., Tamaki, H., and Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. In *Proceedings of the 17th ACM Symposium on Principles of Database Systems*.

Ponte, J. and Croft, W. (1999). A language modeling approach to information retrieval. *Proceedings of SIGIR-1999*, pages 275–281.

Rice, J. (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5).

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Schutze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24:97 –124.

Story, R. (1996). An explanation of the effectiveness of latent semantic indexing by means of a bayesian regression model. *Information Processing & Management*, 32:329 –344.

Strang, G. (1998). *Introduction to Linear Algebra*. Wellesley.

van Rijsbergen, C. (1979). *Information Retrieval*. Butterworths.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *J. Information Retrieval*, 1:67–88.

Zha, H., Ding, C., Gu, M., He, X., and Simon, H. (2002). Spectral relaxation for K-means clustering. *Advances in Neural Information Processing Systems 14*, pages 1057–1064.

Zha, H., Marques, O., and Simon, H. (1998). A subspace-based model for information retrieval with applications in latent semantic indexing. *Proc. Irregular '98, Lecture Notes in Computer Science, Vol. 1457. pp.29-42.*

Zhang, Z., Zha, H., and Simon, H. (2002). Low-rank approximations with sparse factors i: Basic algorithms and error analysis. *SIAM Journal of Matrix Analysis and Applications*, 23:706–727.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.