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. Email: chqding@lbl.gov

Abstract

Experiments show that information retrieval and filtering can be much improved by Latent Semantic Indexing (LSI) -type dimensionality reduction methods. A new dual probabilistic model based on similarity concepts is introduced and LSI is shown to be the optimal solution of the model. Likelihood curves calculated from four standard document collections indicate the existence of optimal semantic subspace. The calculated optimal subspace dimensions are close to experimental values. Statistical significance of each latent semantic vectors follows a Zipf-like distribution, indicating LSI vectors represent some hidden concepts. The theory answers several fundamental questions and provides a framework for further developments of LSI and similar methods.

Keywords: Information Retrieval, probabilistic model, latent semantic indexing, Zipf distribution, optimal semantic subspace.

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)[19]. Latent semantic indexing (LSI) [6, 7, 8, 3] 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 [1, 16, 23] have appeared, and these studies provide better understanding of LSI. However, many fundamental questions remain unresolved.

Here we outline a probabilistic model based on the fundamental document-document and word-word similarity concepts in information retrieval, and show LSI is the optimal solution of the model. Using the likelihood as a quantitative measure of the statistical significance of the semantic subspace, we calculated likelihood curves of four standard document collections and found they all exhibit the "optimal semantic subspace", a key concept in any effective dimensionality reduction methods. The calculated optimal subspace dimensions are close to experimental values. The statistical significance of the LSI vectors are found to be directly related to their eigenvalues which are found to follow a Zipf-like distribution, indicating LSI vectors represent some hidden concepts. This establishes a theoretical framework upon which LSI and similar methods can be analyzed and further improved.

1 Latent Semantic Indexing

We begin with a brief description of LSI. The word matching scheme in informal retrieval (IR) is typically implemented in a vector-space spanned by d indexing words [19, 18]. Each document $\mathbf{x}_i$ is represented by a vector, containing the frequences of index words occurring in the document, properly weighted by other factors [20]. Collecting all n document vectors into a word-document matrix

$X \equiv (\mathbf{x}_1 \cdots \mathbf{x}_n) \equiv (\mathbf{t}^1 \cdots \mathbf{t}^d)^T$, X then uniquely defines the words to documents associative relations. (Here subscripts denote column vectors, representing documents; superscripts denote row vectors, representing words.) Given a user query $\mathbf{q}$, consisting of a set of keywords, the retrieval system calculates a list of scores $s_i = \mathbf{q}^T \mathbf{x}_i$ as the relevance of each document to the query. (We assume both documents and query are normalized to 1 without loss of generality. Thus cosine similarity is equivalent to dot-product similarity.) The relevant documents are sorted according to the score and returned to user.

LSI re-represents both words and documents in a much smaller k-dim subspace. This is done through the truncated singular value decomposition [13], $X \simeq U_k \Sigma_k V_k^T$, where $\Sigma_k \equiv \mathrm{diag}(\sigma_1 \cdots \sigma_k)$ are the singular values, $U_k \equiv (\mathbf{u}_1 \cdots \mathbf{u}_k)$ and $V_k \equiv (\mathbf{v}^1 \cdots \mathbf{v}^k)$ are left and right singluar vectors, which span the k-dim subspace. The query is transformed to $\mathbf{q}^T U_k$, the documents are represented as $\Sigma_k V_k^T$, and the relevance score is computed as $\mathbf{s} = (\mathbf{q}^T U_k)(\Sigma_k V_k^T)$. Typically taking $k = 200$ (far less than original dimensions), LSI increases the precision for retrieval and accuracy for classification [6, 8, 1, 23, 22]. The success of LSI is attributed to that LSI subspace captures the essential associative semantic relationships better than the original document space, and thus partially resolves the word choice problem.

Clearly, a theoretical and quantitative understanding is important. Mathematically, the truncated SVD is the best approximation of X in the reduced k-dim subspace [13]. But are there reasons beyond mathematical approximation? How to characterize and measure the associative semantic relationships? Is there an optimal semantic subspace? If yes, could it be determined theoretically rather than by exhaustive experiments? Although several studies have appeared [1, 16, 23], these fundamental questions remain unresolved.

2 Dual Probability Model

We start with the matrices $X^T X$ and XX^T , since they determine the SVD and give arise to latent semantic vectors. It is widely accepted in IR that the cosine or dot-product between two documents vectors best measures the correlation or similarity of word usages in the two documents, therefore the similarity between the two documents. $X^T X$ contains similarities between all pairs of documents, and is the similarity matrix between documents. The dot-product between two word vectors measures their co-occurrences through all documents in the collection, therefore their closeness or similarity. XX^T contains similarities between all pairs of words and is the word-word similarity matrix. These similarity matrices are of fundamental importance in IR. Note that document-document similarity is defined in the word-space, while word-word similarity is defined in document-space. This strong dual relationship between documents and words is a key feature of the model.

If we view each document as a data entry in the d-dimensional word-space (index space), there are reasons to believe that documents do not occur entirely randomly. We assume they are generated according to certain probability distribution. To find the optimal distribution, we assume (1) The probability distribution is governed by k characteristic document vectors $\mathbf{c}_1 \cdots \mathbf{c}_k$ (collectively denoted as C_k), which will be later identified as the latent semantic index vectors in LSI. (2) The occurrence of a document $\mathbf{x}_i$ is proportional to its similarity to $\mathbf{c}_1 \cdots \mathbf{c}_k$. (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 assumptions, and further motivated by Gaussian distribution, we consider the following generative model:

$$p(\mathbf{x}_i | \mathbf{c}_1 \cdots \mathbf{c}_k) = e^{(\mathbf{x}_i \cdot \mathbf{c}_1)^2} \cdots e^{(\mathbf{x}_i \cdot \mathbf{c}_k)^2} / Z(C_k) \quad (1)$$

where $Z(C_k)$ is the normalization constant. To obtain $\mathbf{c}_1 \cdots \mathbf{c}_k$ as the optimal parameter for the probability model, we use the maximum likelihood estimation (MLE) that maximizes the log-likelihood function: $\ell(C_k) = \log \prod_{i=1}^{n} p(\mathbf{x}_i)$, assuming the documents are independently, identically distributed. After some algebra and using the relation,

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

we obtain

$$\ell(C_k) = \sum_{j=1}^{k} \mathbf{c}_j^T XX^T \mathbf{c}_j - n \log Z(C_k). \qquad (3)$$

The word-word similarity matrix XX^T arises as natural consequence of MLE. Note that due to the duality relationship, it is the word similarity matrix, rather than the document similarity matrix, that appears here. Since documents are data which live in the index space (word 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 principle component analysis.

In general, finding C_k that maximizes $\ell(C_k)$ involves rather complicated numerical procedure, primarily because $Z(C_k)$ is a high ($d = 10^3 - 10^5$) dimensional integral.

$$Z_k = \int \cdots \int e^{(\mathbf{x}\cdot\mathbf{c}_1)^2 + \cdots + (\mathbf{x}\cdot\mathbf{c}_k)^2} dx^1 \cdots dx^d. \qquad (4)$$

which is analytically intractable. However, note that $n \log Z(C_k)$ is a very slow changing function in comparing 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 logrithm of a slow changing function is even slower. Thus $n \log Z(C_k)$ can be regarded as fixed, and we concentrate on maximizing the first term.

The symmetric positive-definite matrix XX^T has a spectral decomposition (eigenvector expansion) with non-negative eigenvalues only. Therefore the optimal solution to $\sum_j \mathbf{c}_j^T XX^T \mathbf{c}_j$ is that $\mathbf{c}_1 \cdots \mathbf{c}_k$ are the first k eigenvectors of corresponding to k largest eigenvalues λ_i. They are exactly the left singular vectors $\mathbf{u}_1 \cdots \mathbf{u}_k$ in LSI/SVD (we refer them as LSI vectors). And the maximal likelihood is

$$\ell(U_k) = \lambda_1 + \cdots + \lambda_k - n \log Z(U_k). \qquad (5)$$

Thus LSI is the optimal solution of the model

The above analysis are carried out in word-space. We can also model words as defined by their occurrences in the document space. (This is similar to several automatic text categorizations, where words are often first clustered according their co-occurrences in documents). Here the data are the words, indexed by documents and 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 metric, we assume the probability for the occurrence of word $\mathbf{t}^{\alpha} (\alpha = 1, \cdots, d)$ to be

$$p(\mathbf{t}^{\alpha} | \mathbf{r}^1 \cdots \mathbf{r}^k) = e^{(\mathbf{t}^{\alpha}\cdot\mathbf{r}^1)^2} \cdots e^{(\mathbf{t}^{\alpha}\cdot\mathbf{r}^k)^2} / Z(R_k). \quad (6)$$

After some algebra and using the identity,

$$\sum_{\alpha=1}^{d} t_i^{\alpha} t_j^{\alpha} = (X^T X)_{ij}, \qquad (7)$$

the log-likelihood $\ell(R_k) = \log \prod_{\alpha=1}^{d} p(\mathbf{t}^{\alpha})$ becomes

$$\ell(R_k) = \sum_{j=1}^{k} \mathbf{r}^{jT} X^T X \mathbf{r}^j - d \log Z(R_k). \qquad (8)$$

The document-document similarity matrix $X^T X$ arise naturally here. Again, the second term, the normalization constant, is a slow changing function and we maximize the first term. The symmetric positive definite matrix $X^T X$ has a spectral decomposition with eigenvalues ξ_i and the optimal solution for $\mathbf{r}^1 \cdots \mathbf{r}^k$ are precisely the right singular vectors $\mathbf{v}^1 \cdots \mathbf{v}^k$ of LSI. The maximal log-likelihood is

$$\ell(V_k) = \xi_1 + \cdots + \xi_k - n \log Z(V_k). \qquad (9)$$

Eqs.(1,6) are dual probability representations of the LSI. This dual relation is further enhenced by the facts that XX^T and X^TX have the same eigenvalues ($\lambda_j = \xi_j = \sigma_j^2$); 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 representations have the same maximum log-likelihood

$$\ell_k = \sigma_1^2 + \cdots + \sigma_k^2$$

up to the normalization constants. This also suggests that the contribution of a small singular vector is σ_i^2, which is much smaller than previously thought: their statistical significance is proportional to the square of its singular values, instead of the singular values themselves as they occur in the SVD.

3 Optimal Semantic Subspace

From statistical point of view, LSI amounts to an effective dimensionality reduction. Dimensions with small singular values are often viewed as representing semantic noises and thus are ignored. This generic argument, considering its fundamental importance, needs to be clarified and quantified. 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 key is to calculate the log-likelihood for the k-dim latent semantic subspace. In word-space model it is Eq.5. The analytically intractable $Z(U_k) = Z(\mathbf{u}_1 \cdots \mathbf{u}_k)$ can be evaluated numerically by generating uniform random numbers in the domain of the integration: on the unit sphere with all components positive. We found this sampling method converges very quickly. It achieves accuracy of 4 decimal places with merely 2000 points for $d = 2000 - 5000$ dimensions. The log-likelihood in document-space model Eq.9 can be calculated similarly.

We calculated the likelihood curves for standard test document collections in IR [25]: CRAN

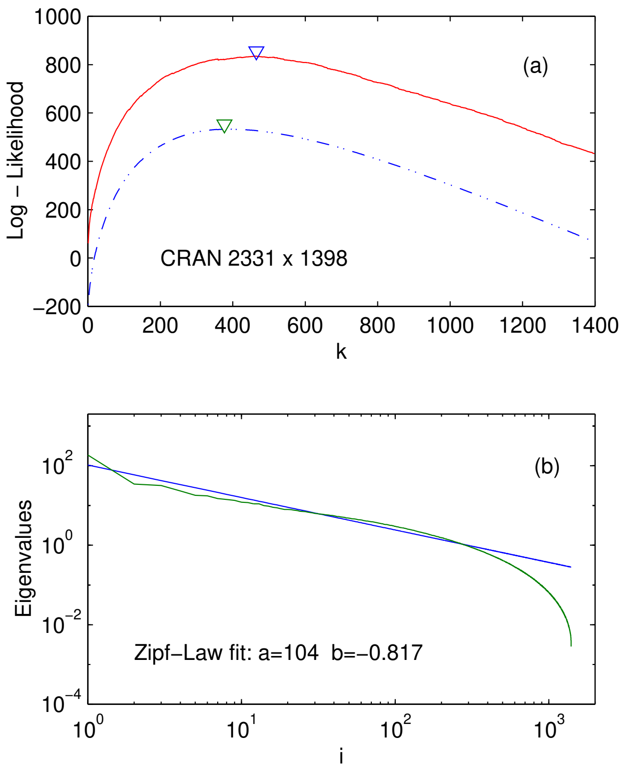

Figure 1: Results for Cranfield collection. (a) Log-likelihood: solid line for word-space model and dashed line for document-space model. Peak points are indicated. (b) Eigenvalues ($\lambda_i = \xi_i$) of LSI vectors plotted in log-log scale. The straight line is the Zipf-law fit. The two curves are indistinguishable if plotted in linear scales.

(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), and CISI (1460 abstracts from Institute of Scientific Information). Here we use the standard term frequency- inverse document frequence (TF-IDF) weighting. Each document (column vector $\mathbf{x}_i$) is normalized to 1, i.e., $\mathbf{x}_i \cdot \mathbf{x}_i = 1$. Results for the Cranfield collection are shown in Figure 1. Both word-space and document-space likelihoods grow steadily and rapidly as k increases from 1 upto k_{opt} = 377-465, clearly indicating that the probability models provide better and better statistical descriptions of the data. However, starting from k $>k_{\text{opt}}$, the likelihood actually

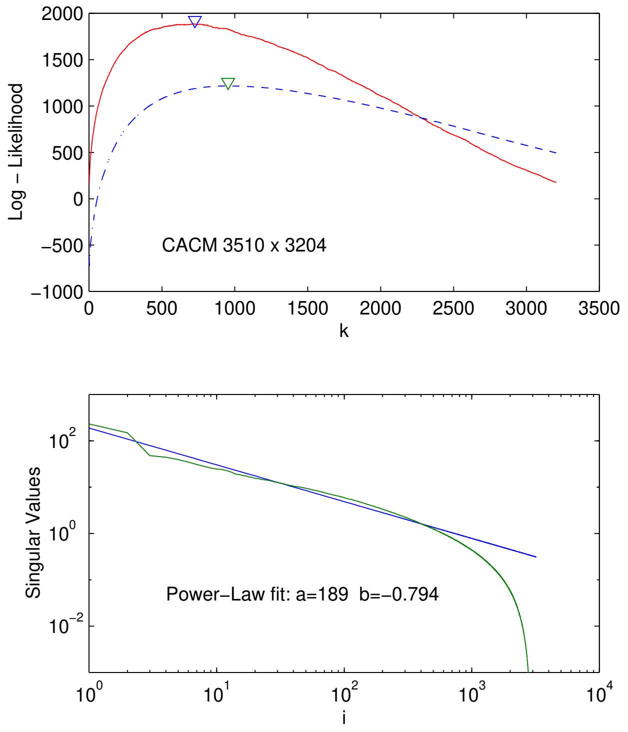

Figure 2: Results for CACM collection. Notations are same as in Figure 1.

decreases, indicating no meaningful statistical information are represented by those latent index vectors with smaller eigenvalues. Results for the CACM collection, with 3510 documents indexed by 3204 words are shown in Figure 2. The likelihood curves for CACM behave very similar to those of Cranfield. Likelihood curves for other collections are similar. Thus the existence of optimal semantic subspace is conclusively demonstrated here for the first time. If we pick the first peak as k_{opt} , then $k_{\text{opt}} = 377$ for Cranfield, 184 for Medline, 726 for CACM, and 274 for CISI. k_{opt} values for Medline and CACM are quite close to experimentally determined values [22, 23].

As explained before, the statistical significance of each LSI index vector relates to their eigenvalue. The eigenvalues of the latent semantic vectors for Cranfield collection are shown in Figure 1b. Within the range of meaningful latent semantic vectors, $1 \leq i \leq k_{\text{opt}}$, they follow a Zipf-like distribution

$$\lambda_i = a \cdot i^b$$

with the exponent $b = -0.82$, very close to $b = -1$ for word usage in English or webpage popularity on internet [26, 12]. The statistical significance of each LSI vectors for CACM (Figure 2) show very similar Zipf distribution. All other document collections display very similar Zipf-like distributions. This indicates that LSI vectors do represent some hidden concepts or identities in a similar manner as English words or webpages do.

4 Some Related Work on LSI

Traditional IR probabilistic models, such as the binary independence retrieval model [18, 10, 14] 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 word-document matrix X, using a generative model, i.e., the probability of occurrence of documents, which could be used as the prior probability in the language modeling approach for IR [17].

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 [1]. 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 improved the retrieval precision.

If the first k singular values of SVD are well separated from the rest, the k-dim subspace is proved to be stable against smaller perturbations [16]. A probabilistic corpus model built upon k topics is then introduced and is shown to be essentially the LSI subspace.

A subspace model using low-rank- plus- shift structured is introduced in [23] and lead to a relation to determine optimal subspace dimension k_{opt} from the singular values, which is originally developed for array signal processing using minimum description length principle.

Using a mixture model for document clustering [5], which is very close to the spherical K-means clustering, the concept vectors (the centroids of each clusters) are found be quite similar to the LSI index vectors. The subspace spanned by the concept vectors are also close to the LSI subspace.

Introducing a latent class variable in the joint probability for P(document,word), the resulting probability of the aspect model [9] can be written as $U\Sigma V^T$, similar to LSI formulation. This *direct* probabilistic formalism is further developed to handle queries and its effectiveness are demonstrated.

5 Conclusions

Using the similarity measures in information retrieval, we introduced a probabilistic generative model for dimension reduction type method of LSI. Similarity matrices then arise naturally during the maximum likelihood estimation of the model parameter, and latent semantic index vectors of LSI are the optimal solutions of the model.

Semantic associations characterized by the LSI index vectors are understood as the statistical significance and measured by the likelihood of the model. Calculations on four standard document collections indicate the existence of optimal semantic subspace and the calculated optimal subspace dimensions are close to experimental values. Importance of LSI indices follows a Zipf-like distribution, indicating LSI vectors represent some hidden concepts or identities. Overall, the model provides a statistical framework for understanding and developments of LSI in IR.

Besides automatic information retrieval and filtering, LSI have also been used in other areas, such as automatic word sense disambiguation (the word "capital" could mean "capital goods", "city where government seats", etc) in context space [21], etc. Our model applies there too, as long as the semantic structures as defined by the dot-product similarity remain the essential relationship. The existence of the optimal semantic subspace described here provide quantitative understanding and support to a new theory of knowledge acquisition and representation[15] in cognitive science.

Acknowledgement. I thank Hongyuan Zha for providing word-document matrices initiating LSI research here, Inderjit Dhillon and Michael Berry for illuminating seminars at NERSC/LBL, Horst Simon, Parry Husbands, Osni Marques and Zhenyue Zhang for valuable discussions. 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

[1] 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.

[2] N.J. Belkin and W.B. Croft. Information filtering and information retrieval: two sides of the same coin? Commun. ACM, **35**, 29 (1992).

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

[4] M.W. Berry and M. Browne. Understanding Search Engines: Mathematical Modeling and Text Retrieval. SIAM Press. June 1999.

[5] I. Dhillon and D. Modha. Concept Decomposition for Large Sparse Text Data Using Clustering. Talk given at 5th Int'l Symposium on Solving Irregularly Structured Problems in Parallel. Berkeley, CA. August 1998. To appear in *Machine Learning.*

[6] 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**, pp:391-407 (1990).

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

[8] S.T. Dumais. Using LSI for information filtering: TREC-3 experiments. Third Text REtrieval Conference (TREC3), D. Harman, Ed., National Institute of Standards and Technology Special Publication, 1995.

[9] T. Hofmann. Probabilistic Latent Semantic Indexing. Proc. of SIGIR'99 (ACM Press, 1999), pp.50-57.

[10] N. Fuhr. Probabilistic Models in Information Retrieval. Computer Journal, **35**, 243 (1992).

[11] G.W. Furnas, T.K. Landauer, L. Gomez, S.T. Dumais. The Vocabulary Problem in Human-system communications. Commun. ACM, **30**, pp:964-971. (1987).

[12] S. Glassman. A Caching Relay for the World Wide Web. Comput. Networks ISDN System, **27**, 165 (1994).

[13] G. Golub and C.V. Loan. Matrix Computation. Johns-Hopkins, Baltimore, 2nd ed. 1989.

[14] C.D. Manning and H. Schutze. Foundations of Statistical Natural Language Processing. MIT Press. Cambridge, MA. 1999.

[15] T.K. Landauer and S.T. Dumais. A Solution to Plato's Problem: the Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review, **104**, 211 (1997).

[16] C.H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala. Latent Semantic Indexing: A Probabilistic Analysis. Proc. of Symposium on Principles of Database Systems (PODS), Seattle, Washington, June 1998. ACM Press.

[17] J. Ponte and W.B. Croft. A Language Modeling Approach to Information Retrieval. Proc. of SIGIR'98 (ACM Press, 1999), pp.275-281.

[18] C.J. van Rijsbergen. Informational Retrieval. 2nd Ed. (Butterworths. 1979).

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

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

[21] H. Schutze. Dimension of meaning. Proc. of Supercomputing'92. (IEEE Press, 1992), pp.787-796.

[22] Y. Yang. Noise Reduction in a Statistical Approach to Text Categorization. Proc. of SIGIR'95 (ACM Press, 1995), pp.256-263.

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

[24] A list of LSI related papers, data, links, etc is on webpage: http://www.cs.utk.edu/ lsi/.

[25] These documents are publically available on internet: ftp://ftp.cs.cornell.edu/pub/smart/.

[26] G.K. Zipf. Human Behavior and the Principle of Least Effort. Addison-Wesley, 1949. A complete list of references and information on Zipf law is http://linkage.rockefeller.edu/wli/zipf/.