Journal of Computing and Information Technology - CIT 17, 2009, 3, 259–264
doi:10.2498/cit.1001268

On the Performance of Latent Semantic Indexing-based Information Retrieval

Ch. Aswani Kumar[1] and S. Srinivas[2]

[1]Intelligent Systems Division, School of Computing Sciences, VIT University, Vellore, India
[2]Fluid Dynamics Division, School of Science, VIT University, Vellore, India

Conventional vector-based Information Retrieval (IR) models: Vector Space Model (VSM) and Generalized Vector Space Model (GVSM) represents documents and queries as vectors in a multidimensional space. This high dimensional data places great demands on computing resources. To overcome these problems, Latent Semantic Indexing (LSI), a variant of VSM, projects the documents into a lower dimensional space. It is stated in IR literature that LSI model is 30% more effective than classical VSM models. However, statistical significance tests are required to evaluate the reliability of such comparisons. Focus of this paper is to address this issue. We discuss the tradeoffs of VSM, GVSM, LSI and evaluate the difference in performance on four testing document collections. Then we analyze the statistical significance of these performance differences.

Keywords: dimensionality reduction, generalized vector space model, latent semantic indexing, singular value decomposition, vector space model

1. Introduction

Information Retrieval (IR) deals with the representation, storage, organization of and access to information items. The models for text retrieval can be primarily divided into two categories: keyword oriented and matrix oriented [1]. Keyword based models uses certain data structures and searching algorithms. Matrix oriented models changes the keyword representation of documents into a matrix format. Vector Space Model (VSM) is a conventional IR model, which represents a document collection by a term-document matrix. VSM views documents as vectors in a high dimensional space with inter document similarity measured by the corresponding vector cosine [2]. Generalized Vector Space Model (GVSM) attempts to improve VSM by altering the axes of information space to account for inter-term correlation. With the correlation matrix, GVSM mitigates the error introduced to the VSM by assuming term independence [3]. Since term-document matrices are usually high dimensional and sparse, they are susceptible to noise. Dimensionality reduction is a way to overcome these problems. Latent Semantic Indexing (LSI) attempts to improve GVSM model of term correlations by means of dimensionality reduction. Features of the derived LSI space are orthogonal and convey most of the variance of observed data using relatively few dimensions.

It is stated in IR literature that LSI model for IR outperforms classical VSM models by an average of 30% [1, 4, 5]. However, best to our knowledge, no empirical results have been presented in the literature evaluating significance of the performance of these models. While evaluation has been a fundamental issue in IR for at least two decades, statistical tests for differences between retrieval models have not received nearly the same attention. These tests can be extremely useful as they provide information about whether observed differences in evaluation scores are really meaningful or simply due to chance. In this paper we discuss the tradeoffs of each model, evaluate the differences in performance of these models and analyze the statistical significance of their performance. The rest of this paper is organized as follows. Section 2 reviews VSM. GVSM is discussed in Section 3. Section 4 illustrates the dimensionality reduction and LSI. Section 5 presents the analysis of the experimental results followed by conclusions and references.

2. Vector Space Model (VSM)

VSM depends on the assumption that the meaning of a document can be derived from the document's constituent terms. A vector is used to represent each document in collection. Each component of vector reflects a term associated with the document. The value assigned to that component reflects the importance of the term in representing semantics of the document. Each document is represented by a weight vector $\overrightarrow{\mathbf{d_j}} = (w_{1j}, w_{2j}, ..., w_{tj})^T$ where $w_{zj}$is the weight or importance of the term z in representation of the document $\overrightarrow{\mathbf{d_j}}$, t is the size of the indexing term set. A collection of d documents is then represented by a term-document matrix with t rows and d columns. Query vector representation is given as, $\overrightarrow{\mathbf{q_i}} = (q_{1i}, q_{2i}, ..., q_{ti})^T$ where q_{zi} is the weight of term z in representation of the query $\overrightarrow{\mathbf{q_i}}$. A variety of models is available in the literature for weighting the document and query vector elements [6]. We measure the cosine similarity between a document and a query. Since the individual terms and keywords are not adequate discriminations of the semantic content of documents and queries, performance of the VSM suffers from two classical problems of synonymy and polysemy [2, 7]. The prevalence of synonymy tends to decrease the recall performance of retrieval systems. Polysemy is one factor for poor precision performance.

3. Generalized Vector Space Model (GVSM)

In VSM, the term-document matrix $\mathbf{A}$ is assumed to be the term occurrence frequency matrix obtained from automated indexing. However, it ignores term correlations. Use of a co-occurrence matrix can be justified only if the documents and term vectors are assumed to be orthogonal. GVSM proposed by Wong represent term vectors explicitly in terms of chosen set of orthonormal basis vectors [3]. GVSM modifies VSM by introducing some ad-hoc schemes for including the important effect of term correlation. The correlation matrix provides a model of the relationships that obtain among the corpus indexing terms. The correlation between any two index terms depends on the number of documents in which two terms appear together.

For a term-document matrix $\mathbf{A}$ of dimension tXd, GVSM calculates the term correlation matrix $\mathbf{R}$ of dimension tXt by multiplying $\mathbf{A}$ with its transpose $\mathbf{A^T}$ matrix. Then GVSM calculates similarity between a query vector and document collection as the dot product between query vector, correlation matrix and term-document matrix.

4. Latent Semantic Indexing (LSI)

Domains such as text have large amounts of redundancies and ambiguities among the attributes that result in considerable noise effects which leads to higher dimensionality. Even though the data are lying in a high dimensional space, it is beneficial to reduce the dimension of the data to improve efficiency and accuracy of data analysis [8, 9]. In IR applications dimensionality reduction is able to effectively improve the data representation by understanding the data in terms of concepts rather than words, where a concept is a linear combination of terms [10]. Latent Semantic Indexing (LSI) is a variant of VSM, which maps a high dimensional space into a low dimensional space. LSI tries to take advantage of the conceptual content of documents. Instead of searching on individual terms, a search is performed on concepts.

To approximate a source space with fewer dimensions, LSI uses a matrix algebra technique termed Singular Value Decomposition (SVD) [2, 11]. SVD takes advantage of the implicit higher order structure in association of terms within documents by largest singular vectors. Vectors representing the documents and queries are projected in new, low dimensional space obtained by truncated SVD. LSI starts with a term-document matrix $\mathbf{A}$ of dimension tXd and rank r and uses the SVD to factor it into three matrices $\mathbf{A} = \mathbf{USV^T}$, where $\mathbf{U}$ and $\mathbf{V}$ are matrices whose columns are left and right singular vectors of $\mathbf{A}$, $\mathbf{S}$ is a diagonal matrix whose diagonal elements are non-negative and arranged in decreasing order. The elements on the main diagonal of $\mathbf{S}$ are known as singular values of $\mathbf{A}$ and are the square roots of the eigenvalues of $\mathbf{A^TA}$ and $\mathbf{AA^T}$ [12, 13]. Computationally, a k-dimensional SVD of $\mathbf{A}$ returns $\mathbf{A_k} = \mathbf{U_kS_kV_k^T}$,

where $\mathbf{U_k}$, $\mathbf{V_k}$ are first k columns of $\mathbf{U}$ and $\mathbf{V}$. In this way the rank of $\mathbf{A}$ has been lowered from r to k. Using this low rank approximation, we project high dimensional documents and query vectors into a low dimensional space. In this new space it is reasoned that underlying structure of the collection is revealed, thus enhancing the retrieval performance. To date, several theoretical explanations and results have appeared in the literature and these studies have provided a better understanding of LSI [7, 10, 14].

5. Experimental Analysis

Table 1 summarizes some characteristics of the datasets we have used in our experiments. These datasets are widely used in IR and text mining research. All the document collections are preprocessed by removing stopwords and performing stemming using Porter stemming algorithm. Performance in IR systems is often summarized in two parameters precision and recall [15, 16]. Precision is the portion of relevant documents in the set returned to the user and recall is the portion of all relevant documents in the collection that are retrieved by the system. To evaluate ranked lists, precision can be plotted against recall after each retrieved document. The precision figures at 11 standard recall levels are interpolated by the rule which states that the interpolated precision at the j^{th} standard recall level is the maximum known precision at any recall level between the j^{th} recall level and the $(j+1)^{th}$ recall level.

The success or failure of VSM heavily depends upon the term weighting schemes. We have evaluated various term weighting schemes in the context of both VSM and LSI. Through these experiments we have identified the suitable weighting scheme for each of these collections. A detailed discussion about these weighting methods and their appropriateness for respective testing collection can be found in our recent work [6]. Also, we have explored the best rank approximation for these document collections. Based on these experiments, 100 dimensions for Medline, CACM, CISI and 300 dimensions for Cranfield are retained. Figure 1 presents the comparison of the performance of VSM, GVSM and LSI models on the above mentioned document collections. On Medline collection, all the three models produced similar results at lower recall levels. But at higher recall levels, LSI exhibited a marginal superiority over VSM and GVSM. Among VSM and GVSM, only at higher recall levels is VSM able to retrieve more relevant documents than GVSM. On Cranfield collection VSM exhibited better results than other models at recall levels between 10 and 70. However, at recall levels 10, 90 and 100 LSI exhibited its superiority. GVSM exhibited poor performance on this collection at all recall levels. On CACM collection, VSM performed fairly well over all recall levels. LSI has performed similar to VSM at lower and higher recall levels. GVSM performed equally to VSM and LSI only at higher recall levels. On CISI collection, VSM performed well better at lower recall levels only. But at higher recall levels both GVSM and VSM have performed better than LSI.

Curves for Cranfield, CACM and CISI collections display a fairly typical pattern that illustrates well known tradeoff between precision and recall: high precision at low levels of recall with a relatively rapid drop in precision as recall is increased. The curves for Medline collection illustrate an abnormally high level of performance with relatively high levels of precision maintained across wide range of recall levels. This appearance is likely a result of the way that the document set was created.

Identifier	Description	Terms	Documents	Queries
Medline	Medical Abstracts	5735	1033	30
Cranfield	Aeronautical Collection	4563	1398	225
CACM	Communications of ACM	5763	3204	52
CISI	Information Science abstracts	5544	1460	76

Table 1. Details of the document collections.

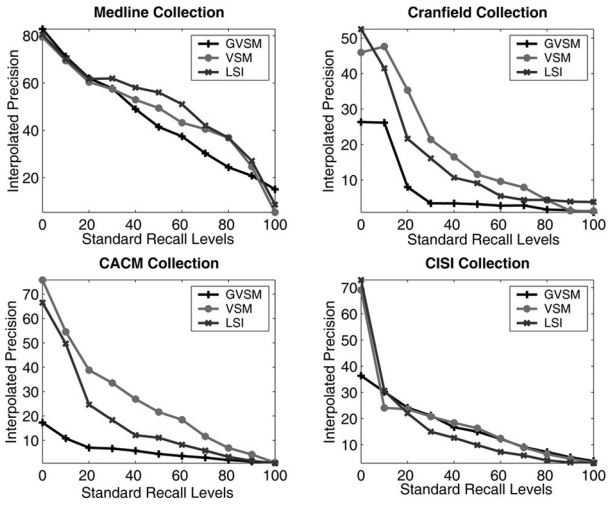
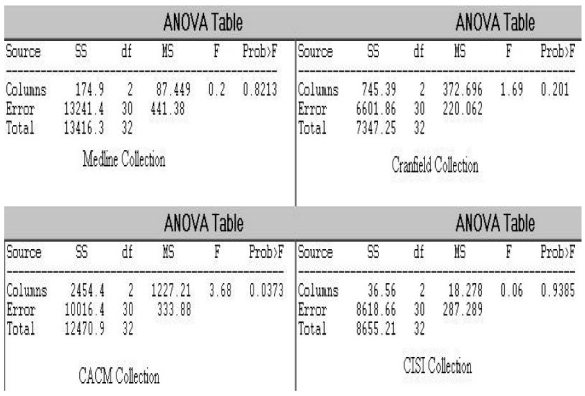


Figure 1. Interpolated precision analysis for VSM, GVM and LSI models.



ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	174.9	2	87.449	0.2	0.8213
Error	13241.4	30	441.38		
Total	13416.3	32			

Medline Collection

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	745.39	2	372.696	1.69	0.201
Error	6601.86	30	220.062		
Total	7347.25	32			

Cranfield Collection

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	2454.4	2	1227.21	3.68	0.0373
Error	10016.4	30	333.88		
Total	12470.9	32			

CACM Collection

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	36.56	2	18.278	0.06	0.9385
Error	8618.66	30	287.289		
Total	8655.21	32			

CISI Collection

Figure 2. ANOVA test for VSM, GVSM and LSI models.

5.1. Statistical Significance Tests

The next step for the evaluation is to analyze values of the interpolated precision obtained by different models. An important question is whether the difference in the precision values is really meaningful or by chance. In order to make such a distinction, it is necessary to apply a statistical test. One-way Analysis of Variance (ANOVA) performs comparison of two or more columns of data in the matrix where each column represents an independent sample containing mutually independent observations [17]. It returns a p-value for the null hypothesis that all samples in the data are drawn from same population or from different populations with same mean. The null hypothesis $\mathbf{H}_0$ will be that all retrieval models being tested are equivalent in terms of performance. The significance test will attempt to disprove this hypothesis by determining a p-value. If the p-value is near zero, ANOVA suggests that at least one sample mean is significantly different than other sample means. It is common to declare that a result is significant if the p-value is less than 0.05 [18]. ANOVA table displays the following information: Source of Variability, Sum of Squares (SS), Degrees of Freedom (DF), Mean Squares (MS) which is the ratio SS/DF, F-statistic which is ratio of the MS's and p-value is derived from F. Figure 2 presents the ANOVA results for the document collections using GVSM, VSM and LSI models of IR.

Very small p-value of CACM collection indicates that the differences between column means are significant. Probability of this outcome supports the alternate hypothesis that one or more of the samples are drawn from population with different means. Hence performance difference between GVSM, VSM and LSI models on this collection is statistically significant. However large p-values from ANOVA table of Medline, Cranfield and CISI collections indicate that differences in performances of GVSM, VSM and LSI are supporting the null hypothesis that differences in performances are statistically not significant.

A box plot is a graphical data analysis technique for determining if differences exist between various levels of ANOVA. A box plot produces a box and whisker plot for each column of X. The box has lines at the lower quartile, mean and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers. If there is no data outside the whisker, a dot is placed at the bottom whisker. Boxes at the same Y axis elevation for all categories indicate little difference among groups. Each box contains a horizontal red line indicating the mean.

Figure 3, presents the box plot of the ANOVA results on GVSM, VSM and LSI. Column 1 in each graph represents GVSM, column 2 represents VSM and column 3 represents LSI. For Medline, Cranfield and CISI collections box plots for VSM, GVSM and LSI are roughly at the same elevation on Y-axis indicating little difference among these models. Only on CACM,

different elevations on the Y-axis are indicating that differences in performances of GVSM, VSM and LSI are significant. A set of mild outliers are visible for GVSM and LSI models on Cranfield collection, CACM collection and set of extreme outliers for VSM and LSI models on CISI collection. Outliers on column 3 reveal that LSI model is able to retrieve more relevant documents at lower recall levels.

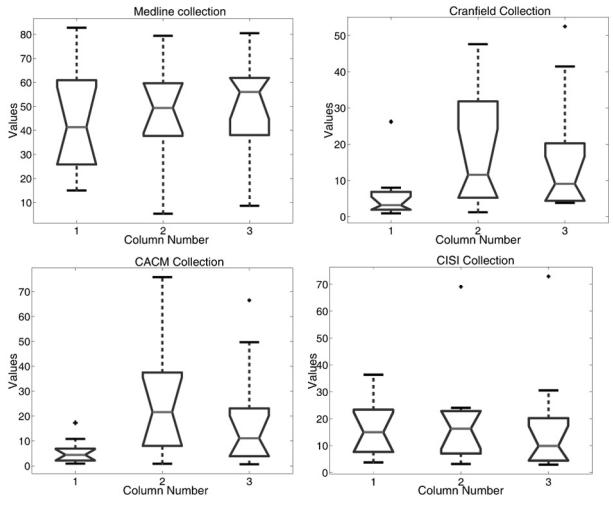


Figure 3. Box plots of ANOVA results.

A valuable point made by Keen [19] is that differences that are not statistically significant can still be important if they occur repeatedly in many different contexts. From these experiments we can conclude that LSI model can almost always be found that can match the performance of classical VSM models rather than outperform: a result which was not confirmed in the earlier literature. According to Deerwester [10], overcoming the lexical mismatch problems is the main benefit that LSI provides. More specifically they claim that the primary objective of LSI is to increase precision at higher levels of recall. This pattern is clearly visible on Medline, Cranfield and CISI collections. The reason behind this pattern is that VSM handles well the relevant documents with a high degree of term overlap, leaving little room for LSI to improve. Relevant documents for a query, that have little or no term overlap, can be improved by LSI. Also, ability of handling dimensionality curse by reducing the dimensionality through SVD makes LSI attractive for IR applications. The major concern in LSI computation is the number of dimensions to be retained [20]. Success of LSI heavily depends on it. Further research should focus on devising a methodology for selecting the number of dimensions to be retained. Other measures of retrieval effectiveness such as fallout ration, average search length etc. can also be applied. In recent literature, researchers have applied non-parametric tests: Wilcoxon signed rank test and Friedman test for comparison of various classifiers [21]. Comparisons of IR models performance can be made using these tests too. Many empirical studies show that good retrieval performance is closely related to the use of various retrieval heuristics [6]. It is expected that optimizing these heuristics will improve the performance of the LSI-based IR systems.

6. Conclusions

It is mentioned in the IR literature that LSI model of IR outperforms classical VSM models by an average of 30%. Primary concern of this paper is to evaluate the significance of the performance tradeoffs between LSI and classical VSM-based IR. Our analysis has revealed that performance improvement of LSI over classical VSM models is not statistically significant, a result which was not confirmed in the literature. However, unlike VSM and GVSM models, LSI model has the ability to handle the higher dimensionality.

7. Acknowledgment

Authors gratefully acknowledge the financial support from Dept of Science and Technology, Govt. of India under the grant number SR/S3/EECE/25/2005. Also, authors are thankful to the anonymous reviewers for their useful comments.

References

[1] E. R. JESSUP, J. H. MARTIN, Taking a New Look at the Latent Semantic Analysis Approach to Information Retrieval. *Computational Information Retrieval, SIAM Publishers*, 121–144, 2001.

[2] M. W. BERRY, Z. DRMAC, E. R. JESSUP, Matrices, Vector Spaces, and Information Retrieval. *SIAM Review*, 41, 335–362, 1991.

[3] S. K. M. WONG, W. ZIARKO, C. N. P. WONG, Generalized Vector Space Model in Information Retrieval, *Proceedings of 8th ACM SIGIR Conference on Research and Development in Information Retrieval*, 18–25, 1985.

[4] Y. HUA, Techniques for improved LSI text retrieval, *Ph.D Thesis, Wayne State University*, 2006.

[5] P. SUWANNAJAN, Predicting the performance of LSI compared to that of vector space model. *Ph.D Thesis, University of Colorado*, 2004.

[6] S. SRINIVAS, CH. ASWANI KUMAR, Optimising the Heuristics in Latent Semantic Indexing for Effective Information Retrieval. *Journal of Information and Knowledge Management*, 5, 97–105, 2006.

[7] A. KONTOSTHATHIS, W. M. POTTENGER, A Framework for Understanding Latent Semantic Indexing Performance. *Journal of Information Processing and Management*, 42, 56–73, 2006.

[8] CH. ASWANI KUMAR, A. GUPTA, S. TREHAN, M. BATOOL, Latent Semantic Indexing-based Intelligent Information Retrieval System for Digital Libraries. *Journal of Computing and Information Technology*, 14, 191–196, 2006.

[9] L. XIAO, J. SUN, S. BOYD, A Duality View of Spectral Methods for Dimensionality Reduction. *Proceedings of* 23^{rd} *International Conference on Machine Learning*, 1041–1048, 2006.

[10] S. DEERWESTER, S. T. DUMAIS, G. W. FURNAS, T. K. LANDAUER, R. HARSHMAN, Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391–407, 1990.

[11] CH. ASWANI KUMAR, S. SRINIVAS, Latent Semantic Analysis Using Eigenvalue Analysis for Efficient Information Retrieval. *International Journal of Applied Mathematics and Computer Science*, 16, 551–558, 2006.

[12] G. GOLUB, C. F. VAN LOAN, Matrix Computations. *The John Hopkins University Press*, 1996.

[13] J. E. TOUGAS, R. J. SPITERI, Updating the partial singular value decomposition in latent semantic indexing. *Computational Statistics and Data Analysis*, 52, 174–183, 2007.

[14] J. C. VALLE-LISBOA, E. MIZZAJI, The uncovering of hidden structures by latent semantic analysis. *Information Sciences*, 177, 4122–4147, 2007.

[15] M. SANDERSON, J. ZOBEL, Information Retrieval System Evaluation: Effort, Sensitivity and Reliability. *Proceedings of* 28^{th} *ACM SIGIR Conference on Research and Development in Information Retrieval*, 62–169, 2005.

[16] R. B. YATES, B. R. NETO, Modern Information Retrieval. *Pearson Education*, New Delhi, 1999.

[17] D. HULL, Using Statistical Testing in the Evaluation of Retrieval Experiments. *Proceedings of* 16^{th} *ACM SIGIR Conference on Research and Development in Information Retrieval*, 329–338, 1993.

[18] J. SAVOY, Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing and Management*, 33, 495–512, 1997.

[19] M. E. KEEN, Presenting Results of Experimental Retrieval Comparisons. *Information Processing and Management*, 28, 491–502, 1992.

[20] CH. ASWANI KUMAR, S. SRINIVAS, Identifying the Number of Dimensions for Dimensionality Selection in Latent Semantic Indexing. *Proceedings of* 2^{nd} *International Conference on Information Processing*, 64–70, 2008.

[21] J. DEMSAR, Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning*, 7, 1–30, 2006.

Received: April, 2008
Accepted: December, 2008

Contact address:
Ch. Aswani Kumar
Intelligent Systems Division
School of Computing Sciences
VIT University
Vellore
India
e-mail: aswanis@gmail.com

CH. ASWANI KUMAR is Assistant Professor (Selection Grade) in Intelligent Systems Division, School of Computing Sciences, VIT University, Vellore, India. He obtained his Masters Degree in Computer Science from Nagarjuna University, India and doctorate degree from VIT University, India. He has research interests in information retrieval, text mining and machine intelligence. He has published 24 refereed research papers in various national and international journals and conferences. He is Principal Investigator to a major research project sponsored by the Department of Science and Technology, Government of India. He is a member of various professional socities including ACM, CSI, ISTE.

S. SRINIVAS obtained his Doctoral Degree from the National Institute of Technology Warangal, India. He has about 20 years of teaching and research experience. He has research interests in fluid dynamics and information retrieval. He has published 50 research papers in national and international journals. He is Principal Investigator to a major research project in the field of fluid dynamics sponsored by DRDO, Govt. of India and co-investigator to two other major research projects. Presently he is Senior Professor and division leader of Fluid Dynamics Division at School of Science, VIT University, India.