BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **9**, No 1

Sofia • 2009

A Note on the Effect of Term Weighting on Selecting Intrinsic Dimensionality of Data

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

[1] *Intelligent Systems Division, School of Computing Sciences*
[2] *Applied Mathematics Division, School of Science and Humanities*
VIT University, Vellore – 632014, India
E-mail: aswanis@gmail.com

Abstract: *The effect of term weighting on selecting intrinsic dimensionality of data is discussed. Experiments are conducted, using different term weighting and dimensionality selection methods, on four testing document collections (namely Medline, Cranfield, CACM and CISI). The results point that transforming the data matrix using a term weighting scheme plays a vital role in identifying the intrinsic dimensionality.*

Keywords: *Dimensionality selection, Latent semantic indexing,Ssingular value decomposition, Term weighting.*

1. Introduction

Vector Space Model (VSM) is a popular Information Retrieval (IR) model which represents documents as vectors in a multidimensional term space. The performance of VSM is influenced by the usage of heuristics: term weighting, similarity measure, etc. VSM models the given document collection in the form of a term-document matrix. Usually this matrix is of high dimensional, sparse in nature and susceptible to noise in the form of synonymy and polysemy. Latent Semantic Indexing (LSI) aims to overcome these problems by uncovering the latent semantic relations of the terms using dimensionality reduction. In addition to the above heuristics as a variant of VSM, LSI depends on the dimensionality reduction technique and intrinsic dimensionality number. The influence of term weighting and

dimensionality reduction on the performance of LSI is well studied in the literature [1, 2]. It is proved that optimizing these heuristics will certainly produce better LSI performance [3]. However the effect of term weighting in selecting the intrinsic dimensionality number is not studied so far. In this work, an attempt is made to understand the effect of term weighting on intrinsic dimensionality selection. This note is organized as follows. Section 2 describes LSI and term weighting. Section 3 discusses the dimensionality selection methods. Section 4 analyzes our experiments on four document collections using different term weighting and dimensionality selection methods.

2. Latent semantic indexing

Latent Semantic Indexing (LSI) analyzes correlations among the terms by identifying usage patterns of terms in document. To capture the major associative patterns in documents, LSI applies truncated Singular Value Decomposition (SVD) and reduces the dimensionality. The process of LSI is well illustrated in [1, 4, 5]. Many studies have shown that good retrieval performance is closely related to the use of various heuristics especially term weighting [6]. Before applying truncated SVD, LSI implements term weighting for the words in the documents. The term weight is given by $L_{ij}G_iN_j$, where L_{ij} is the local weight for term i in document j, G_i is the global weight for term i and N_j is the normalization factor for document j. Local weight represents the importance of a term in the document, global weight represents its importance in the entire document collection and normalization factor compensates the discrepancies in length of the documents. Several term weighting formulae are available in literature ([3, 7] and references therein).

3. Intrinsic dimensionality selection

A heuristic which influences performance of LSI is the number of intrinsic dimensions retained during the dimensionality reduction. The choice should be large enough to characterize the entire dataset and small enough that unimportant features do not fit in the data. Researchers have proposed few methods to select the number of dimensions. Z h u and G h o d s i [8] have proposed to construct a model explicitly for all eigenvalues of the data matrix and estimate position of the gap by maximizing a Profile Likelihood (PL) function. Then we identify the index at which the PL reaches maximum as intrinsic dimensionality number. Parallel Analysis (PA) is a resampling procedure in which the eigenvalues from the research data are compared with those from a random matrix of identical dimensionality to the original data [9]. Amended Minimum Description Length (AMDL) verifies the closeness of eigenvalues by checking the ratio of their arithmetic and quadratic mean and identifies the intrinsic dimensionality [10].The index point that achieves the minimum of AMDL function values is considered as the number of intrinsic dimensions.

4. Experimental results and discussion

To verify the effect of term weighting on selecting the dimensionality we have conducted experiments on Medline, Cranfield, CACM and CISI document collections. Table 1 shows the types of term weighting methods we have applied over the original term document matrix.

Table 1. Local and global weights used

Method	Local weight	Global weight
TF-IDF	$f_{i,j}$	$\log \frac{N}{n_i}$
Log-IDF	$1+\log f_{i,j}$ if $f_{i,j} > 0$ 0 if $f_{i,j} = 0$	$\log \frac{N}{n_i}$
GF-IDF	$\frac{gf_i}{n_i}$	$\log \frac{N}{n_i}$
No-weight	$f_{i,j}$	None

We have applied PA, PL and AMDL over the term-document matrix of each collection. In Table 1, f_{ij} stands for the frequency of term i in document j. N is the number of documents in the collection, n_i is the number of documents in which term i appears and gf_i is the total number of times that term i occurs in the whole collection. For comparison, we have considered the term frequencies as local weights and without any global term weighting under the notion "no-weight".

Based on strategies on which it is designed, a term weighting method either increases or decreases the importance of terms [7, 11]. For e.g. the tf-idf scheme assigns the highest weight to those terms that appear frequently in a small number of documents in the entire document set. For illustration Table 2 shows the values assigned by different term weight methods on the term "Bacillus" in Medline collection.

Table 2. Term weight results on a term in Medline collection

Document No	No-weight	GF-IDF	LOG-IDF	TF-IDF
22	1	0.4730	3.8957	3.8957
144	3	0.9461	8.1756	11.6871
145	1	0.4730	3.8957	3.8957
146	2	0.4730	3.8957	3.8957
147	2	0.7498	6.5960	7.7914
194	4	0.7498	6.5960	7.7914
195	3	1.0984	9.2963	15.5828
196	1	0.9461	8.1756	11.6871
197	2	0.4730	3.8957	3.8957
198	2	0.7498	6.5960	7.7914
199	1	0.7498	6.5960	7.7914
473	1	0.4730	3.8957	3.8957
483	1	0.4730	3.8957	3.8957

Figs. 1, 2 and 3 show the dimensionality selection function values generated, using PL, AMDL and PA respectively for each collection. Each term weighting method on a term-document matrix resulted in different transformations and ultimately produced different intrinsic dimensionality numbers. Estimates of intrinsic dimensionality numbers with different term weighting and dimensionality selection methods can be seen in Tables 3, 4 and 5.

Our results prove that term weighting has significant influence on the number of intrinsic dimensionality. Since the number of unique terms and terms per document vary for each document collection, the performance of various weighting schemes also varies. Further on we have analyzed how this effect of term weighting on selecting the intrinsic dimensionality is influencing the quality of the retrieval. To verify it, we have conducted retrieval experiments on the document collections and measured the interpolated precision at standard recall levels [12]. Fig. 4 presents interpolated precision curves obtained on Medline collection. For each term weighting method, graph displays the interpolated precision curves obtained by LSI model with the intrinsic dimensionality chosen from different dimensionality estimation methods.

Table 3. Intrinsic dimensionality numbers using PL with different weights

Method	MED	CRAN	CACM	CISI
No-weight	165	158	430	195
TF-IDF	238	296	739	398
GF-IDF	132	105	415	151
LOG-IDF	327	406	927	501

Table 4. Intrinsic dimensionality numbers using AMDL with different weights

Method	MED	CRAN	CACM	CISI
No-weight	103	105	88	96
TF-IDF	69	54	48	40
GF-IDF	103	118	93	98
LOG-IDF	27	28	27	19

Table 5. Intrinsic dimensionality numbers using PA with different weights

Method	MED	CRAN	CACM	CISI
No-weight	106	118	134	135
TF-IDF	124	164	244	223
GF-IDF	102	105	181	125
LOG-IDF	424	197	292	253

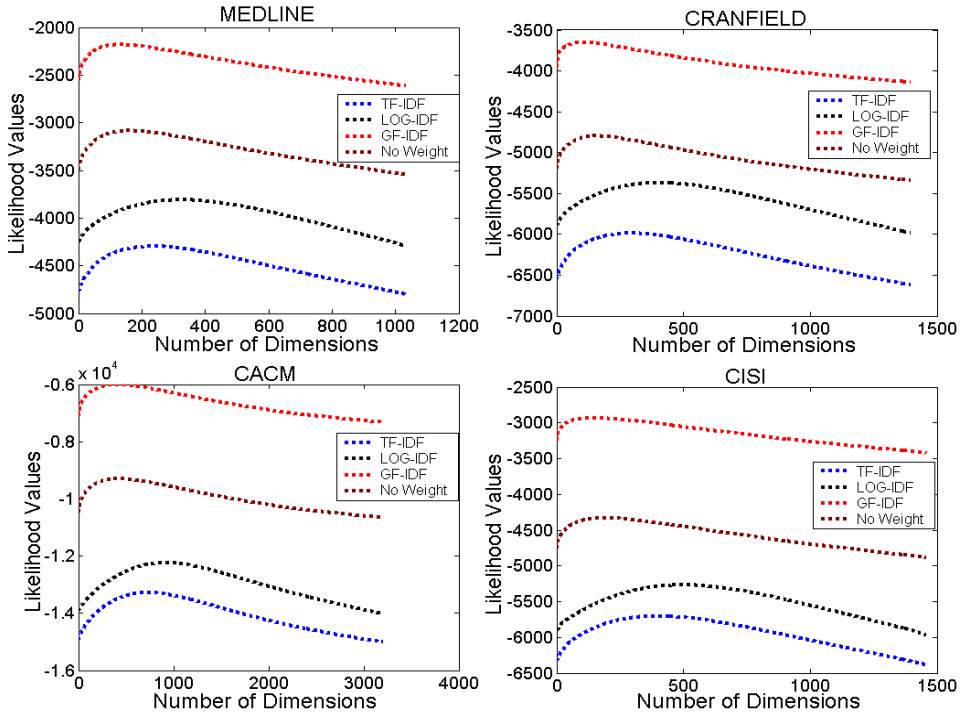

Fig. 1. Distribution of profile likelihood values with different term weights

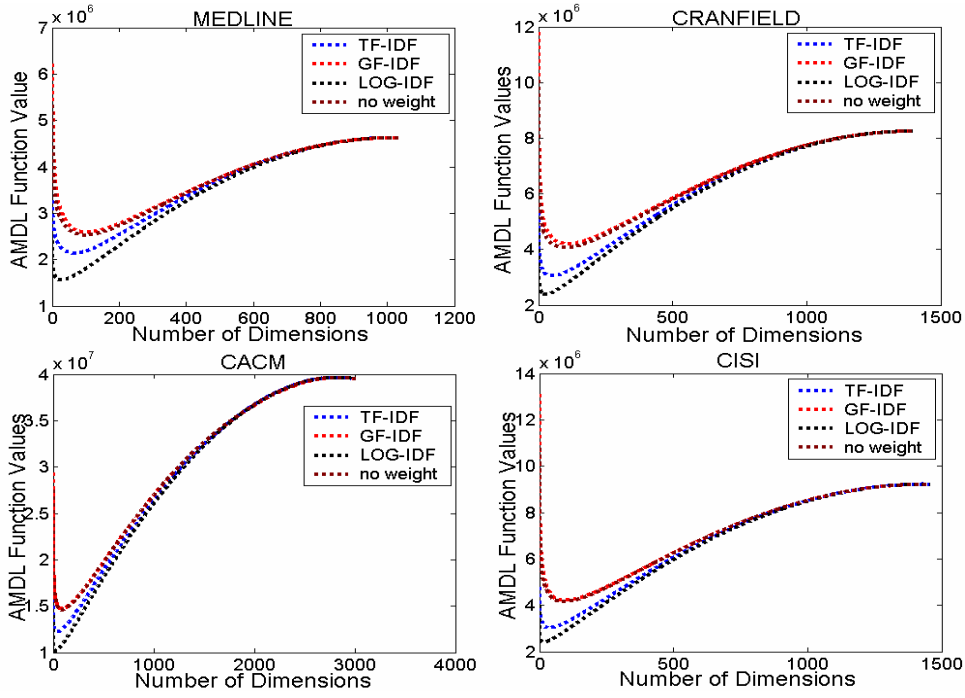

Fig. 2. Distribution of AMDL values with different weights

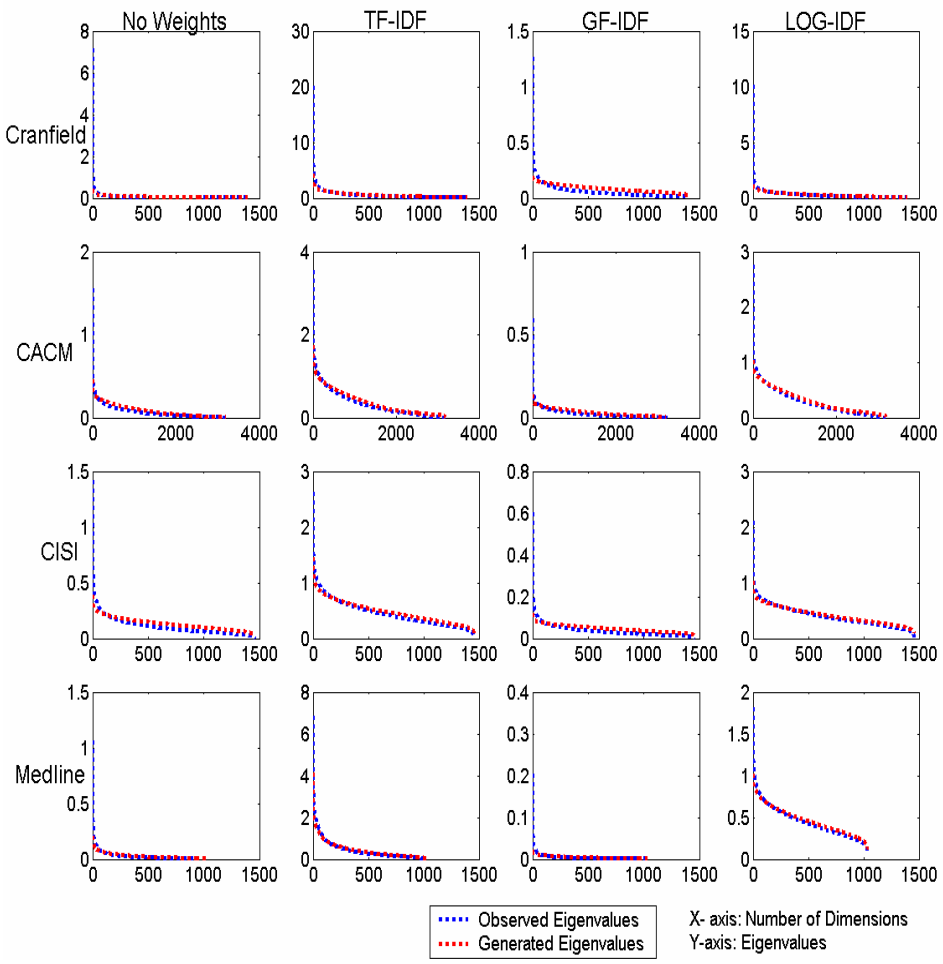

Fig. 3. Distribution of parallel analysis values with different term weights

When no-weight method is applied, LSI with dimensions chosen from AMDL method has presented better retrieval performance and LSI with dimensions chosen from profile likelihood has presented poor interpolated precision at all recall levels. When TF-IDF method is applied, LSI with dimensions chosen from parallel analysis method has produced better interpolated precision results at all recall levels. When GF-IDF method is applied, both parallel analysis and AMDL have selected the intrinsic dimensions 102 and 103 respectively. Hence LSI model performance in both cases is similar at all recall levels. When LOG-IDF method is applied, LSI with intrinsic dimensionality chosen from profile likelihood has produced better precision results at all recall levels and LSI with dimensionality chosen from AMDL has performed poorly.

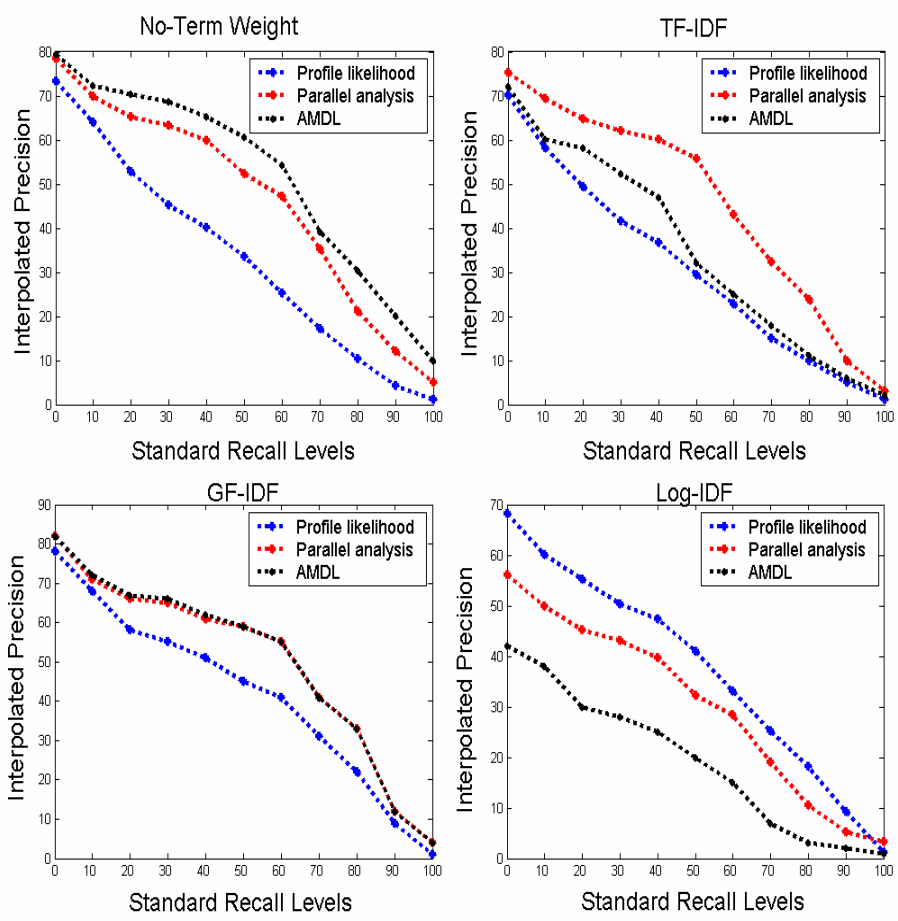

Fig. 4. Performance of LSI with different term-weights and dimensions

5. Conclusion

This note has provided insights into the application of term weighting on document collection and its effect on selecting the intrinsic dimensionality of the corpus in IR models like LSI. It is identified that a dimensionality selection method selects different number of intrinsic dimensions for different term weights. Further, the analysis has proved that this effect influences the performance of the retrieval.

Acknowledgment. The authors acknowledge the financial support from Dept of Science and Technology, Government of India under grant number SR/S3/EECE/25/2005. Also authors thank the anonymous reviewer for the helpful suggestions.

References

1. K u m a r, C h. A s w a n i, S. S r i n i v a s. Latent Semantic Indexing Using Eigenvalue Analysis for Efficient Information Retrieval. – International Journal of Applied Mathematics and Computer Science, Vol. **16**, 2006, No 4, 551-558.
2. H u s b a n d s, P., C. D i n g. Term Norm Distribution and its Effects on Latent Semantic Indexing. – Information Processing and Management, Vol. **41**, 2005, No 4, 777-787.
3. S r i n i v a s, S, C h. A s w a n i K u m a r. Optimizing the Heuristics in Latent Semantic Indexing for Effective Information Retrieval. – Journal of Information and Knowledge Management, Vol. **5**, 2006, No 2, 97-105.
4. D e e r w e s t e r, S. et al. Indexing by Latent Semantic Analysis. – Journal of the American Society for Information Science, Vol. **41**, 1990, No 6, 391-407.
5. B e r r y, M. W. et al. Matrices, Vector Spaces, and Information Retrieval. – SIAM Review, Vol. **41**, 1999, No 2, 335-362.
6. F a n g, H, T. T a o, C. H. Z a i. A Formal Study of Information Retrieval Heuristics. – In: Proc. of 27th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, 49-56.
7. E r i c, C., T. G. K o l d a. New Term Weighting Formulas for the Vector Space Method in Information Retrieval. – Technical Report, ORNL/TM-13756, Oak Ridge National Laboratory, 1999.
8. Z h u, M., A. G h o d s i. Automatic Dimensionality Selection from the Screen Plot via the Use of Profile Likelihood. – Computational Statistics and Data Analysis, Vol. **51**, 2006, No.2, 918-930.
9. E f r o n, M. Eigenvalue Based Model Selection during Latent Semantic Indexing. – Journal of American Society for Information Science and Technology, Vol. **56**, 2005, No.9, 969-988.
10. K u m a r, C h. A s w a n i, S. S r i n i v a s. Identifying the Number of Dimensions for Dimensionality Reduction in Latent Semantic Indexing. – In: Proc. of 2nd International Conference on Information Processing. Bangalore, India, 2008, 64-70.
11. D e b o l e, F., F. S e b a s t i a n i. Supervised Term Weighting for Automated Text Categorization. – In: Text Mining and its Applications. S. Sirmakessis (Ed.), Heidelberg, Physica-Verlag, DE, 2004, 81-98.
12. K u m a r, C h. A s w a n i, S. S r i n i v a s. On the Performance of Latent Semantic Indexing Based Information Retrieval. – Journal of Computing and Information Technology, 2009 (accepted).