Amended Parallel Analysis for Optimal Dimensionality Reduction in Latent Semantic Indexing

Miles Efron
School of Information and Library Science
UNC, Chapel Hill
efrom@ils.unc.edu

16th December 2002

Abstract

This study describes amended parallel analysis (APA), a novel method for dimensionality estimation in unsupervised learning problems such as information retrieval (IR). At issue is the selection of k, the number of dimensions retained under latent semantic indexing (LSI). APA is an elaboration of Horn's parallel analysis, which advocates retaining eigenvalues larger than the values we would expect under term independence. APA operates by deriving confidence intervals on these "null eigenvalues." The technique amounts to a series of non-parametric hypothesis tests on the correlation matrix eigenvalues. In the study, APA is tested along with five previous dimensionality estimators on four standard IR test collections. These estimates are evaluated with regard to two standard IR performance metrics. APA appears to perform well, predicting the best values of k on three of eight observations, and never offering the worst estimate of optimal dimensionality.

1 Introduction

Latent Semantic Indexing (LSI) uses factor-analytic techniques to improve the inter-object similarity function of a vector space model (VSM) IR system [3]. Given an $n \times p$ term-document matrix $\mathbf{A}$ of rank r, LSI projects the n documents and p terms into the space spanned by the first k eigenvectors of $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$, where $k \ll r$. Proponents of LSI argue that this dimensionality reduction removes overfitted information from the system's similarity model. By discarding spurious inferences, dimensionality reduction leads to better predictions of inter-document similarity. Although empirical studies differ in the degree of improvement over keyword retrieval afforded by LSI, they do suggest that dimensionality reduction entails an important elaboration on the standard vector space model (cf. [6, 10]).

However, LSI's benefits depend on the severity of its dimensionality truncation. According to Deerwester *et al.,* choosing k, the number of retained dimensions, is "crucial" to the method's success. Yet in most applications of LSI, this choice is informed by *ad hoc* criteria. The current study introduces amended parallel analysis (APA), a novel method for dimensionality estimation under LSI. Elaborating on earlier work by Horn [9], we argue that an LSI system ought to retain those dimensions whose corresponding eigenvalues are significantly greater than the eigenvalues expected if the variates (i.e. terms) of $\mathbf{A}$ were statistically independent.

To pursue this argument we first describe LSI and show how eigenvalues relate to its dimensionality reduction. Next we introduce the motivation and mathematics behind amended parallel analysis. In section 3 we apply APA to four standard IR test collections, comparing our method's dimensionality estimations to estimates based on four standard eigen-

value analysis methods.

1.1 Latent Semantic Indexing

Dimensionality reduction under LSI is motivated by the idea that an observed term-document matrix $\mathbf{A}$ contains redundant information. Such redundancy introduces error at the hands of the cosine similarity metric (cf. [16]), which assumes that the system's terms are orthogonal. To mitigate this error, LSI derives a low-rank approximation of $\mathbf{A}$ by a standard orthogonal projection. Given the $n \times p$ matrix $\mathbf{A}$ of rank r, LSI begins by taking the singular value decomposition (SVD) of $\mathbf{A}$:

$$\mathbf{A} = \mathbf{T}\mathbf{\Sigma}\mathbf{D}' \tag{1}$$

where $\mathbf{T}$ is an $n \times r$ orthogonal matrix, $\mathbf{\Sigma}$ is an $r \times r$ diagonal matrix, and $\mathbf{D}$ is an $r \times r$ orthogonal matrix. Matrices $\mathbf{T}$ and $\mathbf{D}$ are the left and right singular vectors of $\mathbf{A}$, and the diagonal elements of $\mathbf{\Sigma}$ are the singular values. It can be shown (cf. [8]) that if the columns of $\mathbf{A}$ are centered and standardized to unit length, the singular vectors are the principal components of the co-occurrence matrices (and by virtue of standardization, the correlation matrices) $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$, while the singular values comprise the positive squares roots of the co-occurrence matrix eigenvalues. Thus the diagonal elements of $\mathbf{\Sigma}$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ show the amount of variance captured by each principal component. By choosing to retain only the first k principal components, where $k < r$ and setting the remaining $r - k$ singular values to zero, by matrix multiplication LSI derives $\widehat{\mathbf{A}}_k$, the best rank-k approximation of $\mathbf{A}$, in the least-squares sense.

Advocates of LSI argue that $\widehat{\mathbf{A}}_k$ provides a better model of term-document associations than the full-rank matrix can. Reducing the dimensionality of the model lessens the influence of random, idiosyncratic word choice in our similarity judgements. Thus LSI is capable of overcoming problems of *synonymy* and *polysemy,* allowing an IR system to infer query-document similarity even in the absence of any shared indexing terms.

1.2 Estimating the Intrinsic Dimensionality

It is a mainstay of LSI research that dimensionality truncation entails a noise reduction procedure. According to Berry and Dumais, "the truncated SVD...captures most of the important underlying structure in the association of terms and documents, yet at the same time removes the noise or variability in word usage that plagues word-based retrieval methods" [2]. As Chris Ding notes, this argument begs an important question: which singular vectors are meaningful, and which comprise noise [4, 5]? That is, what value should developers choose for k, the representational dimensionality of an LSI system? Landauer and Dumais ascribe great importance to the choice of k [12], arguing that if k is too small, the similarity model will lack sufficient power to collocate similar documents. On the other hand, as k approaches r, the model becomes overfitted, inferring spurious term-document relationships.

The intuition behind LSI's dimensionality reduction lies in the argument that small singular values imply weak evidence for retaining singular vectors. In Ding's dual-similarity model of LSI, this intuition gains theoretical motivation insofar as the co-occurrence matrix eigenvalues describe each dimension's contribution to the overall likelihood of the LSI model. According to Ding, we should retain those singular vectors whose associated eigenvalues increase the model likelihood [4, 5]. Using this theory, Ding identifies an "optimal semantic subspace" for several IR test collections. By analyzing the eigenvalues, Ding derives an estimate of a corpus' intrinsic dimensionality which correlates strongly with good performance.

2 Amended Parallel Analysis (APA)

Like Ding, we suggest that an analysis of the eigenvalues provides good evidence for estimating the optimal dimensionality of an LSI system. However, we offer another motivation for reducing the number of dimensions. Instead of removing noise, we argue that

dimensionality truncation improves retrieval by removing error from the VSM similarity function. The proposed method, APA, counsels us to retain those dimensions whose corresponding eigenvalues are significantly greater than the eigenvalues expected if the columns of $\mathbf{A}$ were orthogonal. In other words, we argue that LSI's dimensionality reduction is merited to the extent that the observed data violate the assumption of term orthogonality inherent in the vector space model (cf. [17]).

To see that this is the case, let $\mathbf{S}$ be the $p \times p$ covariance matrix of $\mathbf{A}$, such that $\mathbf{S} = (\mathbf{A} - \boldsymbol{\mu})(\mathbf{A} - \boldsymbol{\mu})'$, where the p-vector $\boldsymbol{\mu}$ is the column-wise means of $\mathbf{A}$. Also let $\mathbf{D}$ be the $p \times p$ diagonal matrix containing the square roots of the diagonal elements of $\mathbf{S}$. Then $\mathbf{R}$, the correlation matrix of $\mathbf{A}$, is given by $\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}$. If $\mathbf{R}$ is diagonal then the p columns of $\mathbf{A}$ *are* the principal components and eigenvalues are all equal. Consider the matrix $\mathbf{R}$ below:

$$\mathbf{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{2}$$

with the characteristic equation:

$$|\mathbf{S} - \boldsymbol{\lambda}\mathbf{I}| = \begin{vmatrix} 1-\lambda & 0 \\ 0 & 1-\lambda \end{vmatrix} = (1-\lambda)(1-\lambda) = 0 \tag{3}$$

which gives the eigenvalues $\boldsymbol{\lambda}' = \begin{pmatrix} 1 & 1 \end{pmatrix}$ and eigenvectors $\mathbf{R}$. Thus if $\mathbf{A}$ were orthogonal (as is assumed under the VSM), with $\mathbf{R} = \mathbf{I}_p$ principal component analysis yields no benefit, and dimensionality reduction is not appropriate because each principal component describes equal variance.

2.1 Implementation of APA

APA operates by retaining principal components with eigenvalues that are significantly larger than the eigenvalues expected if the columns of $\mathbf{A}$ were orthogonal. Assuming multivariate normality of terms, the technique uses a statistical simulation to test the null hypothesis that each observed eigenvalue λ_k is equal to the corresponding eigenvalue, λ_{0k}, expected under term independence. We thus reject the k components whose eigenvalues λ_k are significantly smaller than λ_{0k}.

To estimate the eigenvalues under the null hypothesis, let $\mathbf{A}_0^*$ be an $n \times p$ matrix drawn from the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{S}_0 = \mathbf{I}_p$. From $\mathbf{A}_0^*$ we calculate the principal components, with eigenvalues $\boldsymbol{\lambda}_0^*$. By generating a large number, B, replications of $\boldsymbol{\lambda}_0^*$ and finding their average, $\overline{\boldsymbol{\lambda}}_0^*$, we may derive a point estimate of the true $\boldsymbol{\lambda}_0$.

Under Horn's parallel analysis, we retain those principal components where $\lambda_k > \overline{\lambda}_{0k}^*$. However, this approach is somewhat unsatisfying insofar as it takes no account of the standard error of $\overline{\boldsymbol{\lambda}}_0^*$. Thus amended parallel analysis supplements the point estimate of $\overline{\boldsymbol{\lambda}}_0^*$ with an estimate of its standard error to derive a confidence interval upon which we base each hypothesis test, $H_0 : \lambda_k = \lambda_{0k}$.

Without knowledge of the distribution of $\boldsymbol{\lambda}_0$ we derive a $1 - \alpha\%$ confidence interval (CI) for its elements by recourse to non-parametric methods. For each $\overline{\lambda}_{0k}^*$ we use the bootstrap-t method described in [7] to generate our CI. Let the standard error of $\overline{\lambda}_{0k}^*$ be given by Equation 4:

$$\widehat{se}_k = \{\sum_{b=1}^{B} [\lambda_{0k}^*(b) - \overline{\lambda}_{0k}^*]^2/(B-1)\}^2 \tag{4}$$

where $\lambda_{0k}^*(b)$ is k^{th} eigenvalue of the b^{th} draw of $\mathbf{A}_0^*$. Using this estimate of the standard error we calculate $Z^*(b)$, a non-parametric estimate of the likelihood of seeing the bth observation of λ_{0k}^*:

$$Z^*(b) = \frac{\lambda_{0k}^*(b) - \overline{\lambda}_{0k}^*}{\widehat{se}_k} \tag{5}$$

We thus find the α^{th} percentile of $Z^*(b)$ by the value $\widehat{t}^{(\alpha)}$ such that

$$\#\{Z^*(b) \leq \widehat{t}^{(\alpha)}\}/B = \alpha.$$

In other words if we have $B = 100$ bootstrap iterations, the estimate of the fifth percentile point is the fifth largest value of $Z^*(b)$ and the 95th percentile is given by the 95th largest $Z^*(b)$. This approach essentially allows us to construct a pseudo-probability table, tailored to the distribution of the observed data. Thus we observe the variability of our test statistic

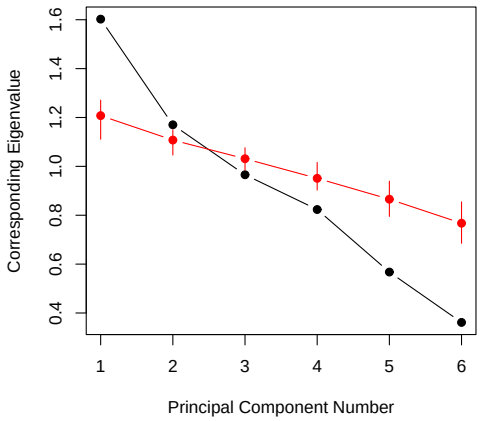

Figure 1: APA applied to physiology data

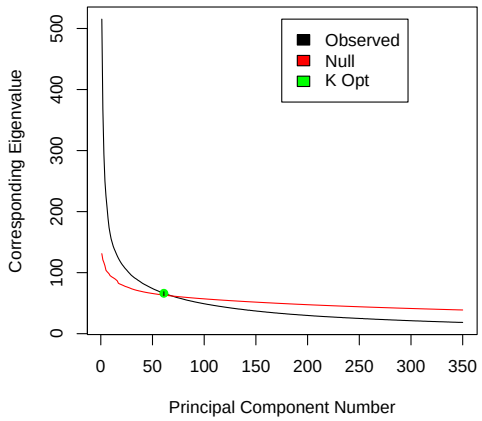

Figure 2: APA applied to the Medline test collection

over a wide number of iterations, generating $Z^*(b)$ for each of our $B = b$ samples. Based on these calculations we derive probability estimates. Having used our pseudo-table of $Z^*(b)$ values to derive an appropriate $\widehat{t}^{(\alpha)}$, our bootstrap-t confidence interval is given by Equation6.

$$(\overline{\lambda}^*_{0k} - \widehat{t}^{(\alpha)}, \overline{\lambda}^*_{0k} - \widehat{t}^{(1-\alpha)}) \tag{6}$$

So with probability $1 - \alpha$ we state that if given infinite data from the same distribution that gives $\mathbf{A}_0$, the *kth* eigenvalue λ_{0k} would lie within the interval specified by Equation 6.

Under APA we reject the last $p-k$ singular vectors whose corresponding eigenvalues, λ_k, are *significantly smaller* than the corresponding λ_{0k}. We define the optimal value of k to be the lowest positive integer such that λ_k is less than the lower bound of the $1-\alpha$ CI for λ_{0k}.

2.2 An Example

Figure 1 shows an application of APA to a small data set concerning human physiology, where each of 60 observations contains 6 measurements. The Figure shows 95% confidence intervals for the null eigenvalues, generated after $B = 100$ simulations. Under the APA method we would retain the first three principal components because only components 4-6 lie below the null line's confidence interval.

Figure 2 shows APA performed on the Medline test collection. Due to the scale of the plot, the confidence intervals are not shown. However, the green dot shows the dimensionality estimation under APA. As shown in the following Section, this prediction appears to agree with standard IR performance metrics.

3 Tests of APA's Estimates

To test the suitability of APA for dimensionality estimation in IR, we compared its predictions against several standard methods from the statistical literature [14]. These methods are summarized in Table 1. The PA and APA methods were described above. The Eigenvalue-one criterion suggests that we should retain those eigenvalues greater than the average eigenvalue. This is similar in spirit to the 70% variance rule, which retains enough eigenvalues to account for 70% of the total variance. Finally, Bartlett's test of isotrapy is a χ^2-based hypothesis

Abbreviation	*Name*
APA	Ammended Parallel Analysis
PA	Horn's Parallel Analysis
EV1	Eigenvalue-One Rule
70% var	70% Variance Rule
Bartlett's	Bartlett's test of Isotrapy

Table 1: Standard Dimensionality Estimators

	CACM	*CISI*	*CRAN*	*MED*
Docs	3204	1460	1398	1033
Terms	5831	5615	1033	3204
k_{opt} *(ASL)*	271	751	121	91
var at k_{ASL}	0.4	0.73	0.19	0.16
overfit (ASL)	0.63	0.63	0.93	0.92
k_{opt} *(pr)*	1936	1276	661	151
var at k_{pr}	1	0.96	0.71	0.25
overfit (pr)	0.71	-0.66	-0.77	-0.95

Table 2: Corpus Stats

test, under which we retain all eigenvalues such that the null hypothesis $H_0 : \lambda_k > \lambda_{k+1}$. These techniques are discussed in [11] and [15].

We tested each of these dimensionality estimators on four standard IR test collections: the CACM data, the CISI collection, Cleverdon's Cranfield set, and the Medline Corpus. To evaluate the quality of our dimensionality predictions, we compared them with two standard performance measures: average precision (cf. [1]) and average search length ([13]). A summary of this information appears in Table 2. For each collection the Table shows the number of documents and terms that it contains (after removing stop-words), the optimal dimensionality according to the ASL measure and according to average precision, the percent of variance accounted for by the model defined by each performance metric's k_{opt}. In addition, we report the correlation between k and each performance metric as k increases from k_{opt} to k_{max}. This measure is intended to show the strength of an overfitting effect encountered by adding too many singular vectors.

Focusing on the Medline data set, Figure 3 plots mean precision against model dimensionality. The

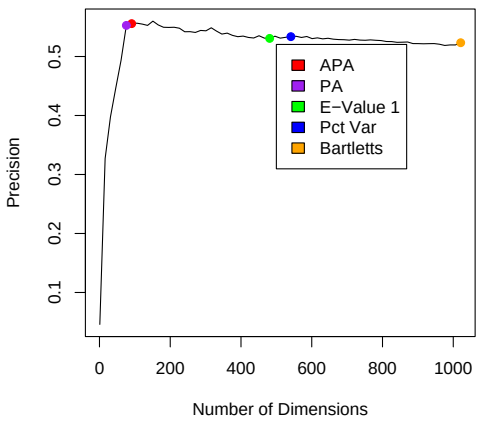

Figure 3: Dimensionality Predictions for Medline (Pr)

	CACM	*CISI*	*CRAN*	*MED*
APA	0.143	-0.461	**-0.019**	**-0.004**
PA	0.142	-0.466	-0.029	-0.016
EV1	0.34	-0.095	0.352	0.372
70% Var	**0.127**	**-0.07**	0.378	0.441
Bartlett's	0.915	0.487	0.913	0.92

Table 3: Dimensionality Predictions (ASL)

Figure shows a region of optimality near $k = 150$. The Figure also shows the predictions yielded by each of our dimensionality estimators. Figure 4 shows ASL performance as a function of dimensionality for the Cranfield data. In both Figures, the methods of PA and APA appear to yield superior dimensionality estimates to the other dimensionality estimators, with APA consistently outperforming standard PA, though by only a small margin.

Our experimental results are summarized in Tables 3 and 4. The cells of these tables contain the distance of each prediction from k_{opt}^{ASL} and k_{opt}^{pr}, respectively, divided by k_{max}, the rank of the data. Thus values near zero indicate good predictions with respect to precision. As the Tables show, APA performed as

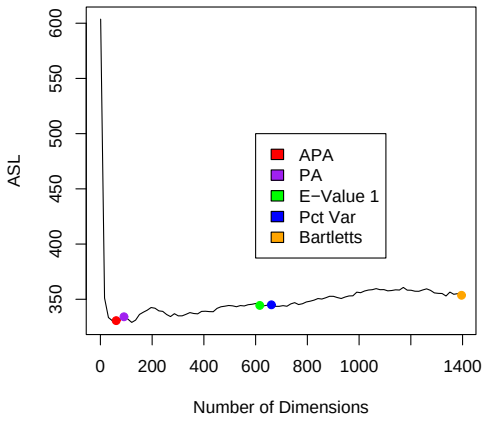

Figure 4: Dimensionality Predictions for Cranfield (ASL)

well as or better than PA for all data sets according to both performance measures. For the Medline data, APA performed dramatically better than the other methods, with similar results on the Cranfield data with respect to ASL. APA's performance on the CACM data measured by ASL also appears to be very good.

The simpler EV1 and *70% Var* methods appear to outperform APA on the CISI data, and on the CACM and Cranfield data for the precision metric. However, it should be noted that this phenomenon may be an artifact of the query-specific Cranfield method of performance evaluation. In both of these cases, the observed k_{opt}^{ASL} and k_{opt}^{pr} were widely disparate. Moreover, the strength of an overfitting phenomenon for these collections varied across performance metrics. Thus the ability of the supplied queries to demonstrate these collections' intrinsic dimensionalities is subject to debate. In other words, without strong evidence for an overfitted model at k_{max}, the consistently higher estimates yielded by However, *EV1* and *70% Var* outperformed APA simply by virtue of their inherent bias toward a model of high dimensionality.

	CACM	*CISI*	*CRAN*	*MED*
APA	-0.414	-0.821	-0.406	**-0.063**
PA	-0.414	-0.827	-0.415	-0.074
EV1	**-0.217**	-0.455	-0.34	0.313
70% Var	-0.429	-0.421	**-0.009**	0.382
Bartlett's	0.396	**0.126**	0.527	0.861

Table 4: Dimensionality Predictions (pr)

4 Conclusions

Amended parallel analysis appears to give good estimations of model dimensionalities that lead to optimal performance under LSI. On three of our eight performance observations, APA outperformed all five previous methods of dimensionality estimation. In the remaining five observations, APA was never the worst performer. These results suggest that the technique merits future work. For example, we hope to experiment on applications of APA that rely on wider confidence intervals. Because we chose $\alpha = 0.05$ for these experiments, our CI's were very narrow, yielding predictions close to those of Horn's parallel analysis. It appears that given the complex models native to IR, with their corresponding legions of eigenvalues, a looser definition of *significant departure from independence* would improve predictions.

Most importantly, however, the observed lack of agreement between our IR performance metrics–ASL and precision–with respect to k_{opt} demands more research. Insofar as most *ad hoc* approaches to optimizing k have relied on such metrics for retrospective model selection, these results argue for the importance of query-independent dimensionality estimators, such as APA. However, the complexity of our findings suggests that comparing the quality of eigenvalue estimation methods may also need to include query-independent metrics such as cross-validation. In particular in upcoming work we will test APA and other dimensionality estimators on a series of simulated data sets. Such approaches will be useful in future work on larger corpora, which we also plan to undertake.

Nonetheless, APA's performance appears encouraging. Not only does the method yield good dimen-

sionality estimates for IR, but it also puts LSI's dimensionality truncation on firmer theoretical ground. The success of APA suggests that dimensionality reduction is merited to the extent that a corpus' indexing features depart from orthogonality. Rejecting eigenvalues smaller than those expected under term independence implies that LSI improves retrieval by removing error from the cosine similarity function that is native to the vector space model of IR.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, 1999.

[2] M. W. Berry, Susan T. Dumais, and G. W. OBrien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. of the American Society for Information Science*, 41(6):391–407, 1990.

[4] C. H. Q. Ding. A similarity-based probability model for latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[5] Chris H.Q. Ding. A dual probabilistic model for latent semantic indexing in information retrieval and filtering.

[6] Susan T. Dumais. LSI meets TREC: A status report. In *Text REtrieval Conference*, pages 137–152, 1992.

[7] B. Efron. *An Introduction to the Bootstrap.* Chapman and Hall, 1993.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, New York, 2001.

[9] J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30:179–186, 1965.

[10] Fan Jiang and Michael L. Littman. Approximate dimension equalization in vector-based information retrieval. In *Proc. 17th International Conf. on Machine Learning*, pages 423–430. Morgan Kaufmann, San Francisco, CA, 2000.

[11] I. T. Jolliffe. *Principal Component Analysis.* Springer, 2 edition, 2002.

[12] 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(2):211–240, 1997.

[13] R. M. Losee. *Text Retrieval and Filtering: Analytic Models of Performance.* Kluwer, Boston, 1998.

[14] K. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis.* Academic Press, 1979.

[15] A. C. Rencher. *Methods of Multivariate Analysis.* Wiley-Interscience, 1995.

[16] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.

[17] S. K. Michael Wong, Wojciech Ziarko, Vijay V. Raghavan, and P. C. N. Wong. On modeling of information retrieval concepts in vector space. *TODS*, 12(2):299–321, 1987.