THE DOCUMENT REPRESENTATION PROBLEM: AN ANALYSIS OF LSI AND ITERATIVE RESIDUAL RESCALING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Rie Ando

August 2001

© 2001 Rie Ando
ALL RIGHTS RESERVED

THE DOCUMENT REPRESENTATION PROBLEM: AN ANALYSIS OF LSI AND ITERATIVE RESIDUAL RESCALING

Rie Ando, Ph.D.
Cornell University 2001

Important text analysis problems in information retrieval and natural language processing, such as document clustering and automatic text summarization, require accurate measurement of inter-document similarity. The goal of this work is to find methods for automatically creating document representations in which inter-document similarity measurements correspond to human judgment.

We present a new model for the task of creating document representations. From this model, we derive a new analysis of Latent Semantic Indexing (LSI), which is one of the successful approaches that has been studied extensively. In particular, we show a precise relationship between LSI's performance and the uniformity of the underlying distribution of documents over topics.

As a consequence, we propose a novel alternative method called Iterative Residual Rescaling (IRR), that, crucially, compensates for distributional non-uniformity. Experiments over a variety of practically-encountered settings and with several evaluation metrics validate our theoretical prediction and confirm the effectiveness of IRR in comparison to LSI.

We also propose several extensions including a new document sampling method to scale IRR up to large document collections. Comparison with random sampling provides further empirical evidence that performance can be improved by counteracting non-uniformity.

Finally, we present a system for multi-document summarization based on IRR, which demonstrates that IRR can be immediately useful in applications. We show that IRR works as a framework to find a tightly connected (and therefore interpretable) set of coherent texts, and effectively present them to the user.

Biographical Sketch

Rie Kubota Ando was born in Japan. She graduated from University of Tokyo with a B.S. degree in Computer Science. After three years in Ithaca, while on educational leave from IBM Japan, she is ready to return to industry – IBM Research, U.S.

Acknowledgements

Looking back over these three years, I feel that I have been having a tremendous amount of good luck. First of all, I found the best, truly the best(!) advisor for me, Lillian Lee. I do not know how to express how wonderful the time I spent with her was. She is the one from whom I learned the pleasure (and pain) of research work. Jon Kleinberg (my secondary advisor) always gave me inspiration and warm encouragement. I appreciated Naoki Sakai's (my minor advisor) insight into natural languages. I would like to greatly thank my advisors above and Clair Cardie for their advice, encouragement, ..., and everything. Also, I would like to thank everyone in the department including all the faculty members, staff members, and my former and current officemates. I would like to thank Hisao Tamaki and Santosh Vempala for their correspondences regarding their work.

My thesis work started while I was visiting IBM T.J. Watson Research Center, where I spent two summers and one semester. (Before then, my work was on word segmentation (Ando and Lee, 2000), which does not appear in this thesis.) I would like to thank Roy Byrd, Branimir Boguraev, and Mary Neff for joint work on summarization and also for reviewing my manuscripts (especially my SIGIR'00 paper) patiently. I would like to thank many IBMers (of course, including the three above) for having helped me get through these tough years: Eric Brown, Herb Chong, Anni Coden, Alan Marwick, Tetsuya Nasukawa, John Prager, Yael Ravin, Edward So, Takao Suzuki, and more. Besides helpful discussions, their warm support included reminding me to eat lunch, offering cozy office space, bringing good stuff from Japan, and so on.

The work described in this thesis was supported in part by a grant from the GE Foundation, a McMullen fellowship from Cornell University, and the National Science Foundation under ITR/IM grant IIS-0081334. Any opinions, findings, and conclusions or recommendations expressed below are those of the author and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

1 Introduction **1**

1.1 Research Contributions 3

1.2 Bibliographic Notes 4

2 Background: Introduction to SVD and LSI **5**

2.1 Mathematical Preliminaries 5

2.1.1 Vectors and matrices 5

2.1.2 Vector length, cosine 6

2.1.3 Eigenvectors and eigenvalues 6

2.1.4 Subspaces, column spaces, the orthogonal complement of a subspace 6

2.1.5 Orthogonal projection 7

2.1.6 Frobenius norm and matrix 2-norm 7

2.2 Singular value decomposition (SVD) 8

2.2.1 Singular value decomposition, singular values, singular vectors 8

2.2.2 Geometrical view of SVD, relations to matrix norms 9

2.2.3 Rank-k approximation by SVD 9

2.2.4 Geometrical interpretation of left singular vectors 10

2.2.5 Computation of SVD 11

2.3 Latent Semantic Indexing (LSI) 11

2.3.1 SVD and Latent Semantic Indexing (LSI) 11

2.3.2 Applications of and Issues with LSI 12
2.4 Preview 14

3 An Analysis of LSI **15**

3.1 Topic-based similarities 15
3.2 The optimum subspace 16
3.3 Non-uniformity and LSI 17
3.4 The dimensionality of the LSI subspace 21
3.5 Related Work: Theoretical Analyses of LSI 23
3.5.1 Generative models 23
3.5.2 Dimensionality selection 25
3.5.3 Other analyses of LSI 27

4 IRR: Overcoming non-uniform topic-document distributions **29**

4.1 IRR algorithm 29
4.1.1 Scaling factor selection: The AUTO-SCALE method 31
4.1.2 Dimensionality selection 32
4.2 Evaluation metrics 33
4.3 Controlled-distribution experiments: validation of theorems 34
4.3.1 Experimental setting 34
4.3.2 Controlled-distribution results 35
4.3.3 Non-uniformity and scaling factor, the best dimensionality 36
4.4 Evaluation on unrestricted distributions 42
4.4.1 Data 42
4.4.2 Kappa average precision 43
4.4.3 Clustering results 43
4.4.4 Discussion 45

5 Scaling IRR up **46**

5.1 Implementation and computation time for IRR 46

5.2 Random-IRR 48

5.3 Sampled IRR (SP-IRR) 49

5.4 Experiments of SP-IRR on controlled distributions 51

5.4.1 Settings 51

5.4.2 SP-IRR: Controlled-distribution results 51

5.5 Evaluation of SP-IRR on unrestricted distributions 54

5.5.1 Data 54

5.5.2 Kappa average precision 56

5.5.3 Clustering results 57

5.5.4 Discussion 57

6 Other ways to compensate for non-uniformity **63**

6.1 Pseudo-document SVD 63

6.2 PDSVD experiments 66

6.3 Discussion 66

7 An Application of IRR: Multi-document Summarization **69**

7.1 Multi-document summarization as an enabling technology for IR 69

7.2 Semantic space 71

7.3 Visual presentation of a semantic space: combining text and graphics 72

7.4 Mapping a document collection into semantic space 75

7.4.1 Term extraction and vector creation 75

7.4.2 Identifying topics 76

7.4.3 Associations between topics and linguistic objects 77

7.5 Further work 78

8 Related work: Document Representation **80**

8.1 Methods regarding document representation 80

8.1.1 Training by prior knowledge of inter-document similarity 80

8.1.2 Factor analysis 81

8.1.3 Probabilistic Latent Semantic Indexing 82

8.1.4 Vector Space-based Methods 83

8.2 Extensions of LSI 84

8.2.1 Document sampling 84

8.2.2 Random projection 84

8.2.3 Other extensions of LSI 85

9 Conclusion **87**

9.1 Main results 87

9.2 Future directions 88

A Theorems and Definitions from Previous Studies **89**

A.1 Perturbation of singular values for symmetric matrices 89

A.2 Canonical angle between subspaces 89

A.3 Tangent theorem 90

A.4 Relationship between tangent and orthonormal bases 90

B Proofs **92**

B.1 Definitions and notational conventions 92

B.2 Proof of Theorem 3.3.1 93

B.3 Proof of Theorem 3.3.2 93

B.4 Proof of Theorem 3.3.3 94

B.5 Proof of Theorem 3.3.4 95

B.6 Proof of Theorem 3.3.5 97
B.7 Proof of Theorem 3.4.1 99
B.8 Proof of Theorem 6.1.1 101

References **104**

List of Figures

2.1 SVD and rank-k approximation by SVD. 10

3.1 Framework of analyzing LSI 16

3.2 Definitions and notational conventions for analyzing LSI 18

3.3 LSI error bounds and dimensionality 22

4.1 Lengths of residuals inversely reflect topic dominances. 30

4.2 Effect of non-uniformity on LSI, and how IRR compensates. 30

4.3 High-level pseudocode for IRR. 31

4.4 Evaluation metric: Sample contingency table 34

4.5 IRR: kappa average precision results, two topics 36

4.6 IRR: floor and ceiling clustering results, two topics 37

4.7 IRR performance, three topics and four topics 38

4.8 IRR performance, five topics 39

4.9 Scaling factor and non-uniformity 40

4.10 Best dimensionality 41

4.11 LSI: Non-uniformity and the best dimensionality on two-topic data 42

4.12 LSI: Non-uniformity and the best dimensionality on five-topic data 42

4.13 IRR: evaluation settings for unrestricted distributions 42

4.14 IRR: absolute improvement in pair-wise kappa average precision over VSM, unrestricted distributions 43

4.15 IRR: p-values from the paired t-test on the kappa average precision results 43

4.16 IRR: document clustering performances, unrestricted distributions 44

4.17 IRR: p-values from the paired t-test on the clustering performance results 45

5.1 Example implementation of IRR. 47

5.2 Schematic illustration of SP-IRR's residual vector selection. 50

5.3 High-level pseudocode for SP-IRR. 50

5.4 Comparison of algorithms: random-LSI, random-IRR, SP-LSI, SP-IRR 51

5.5 SP- and random-IRR: kappa average precision results on controlled distribution data 52

5.6 Kappa average precision of the randomized methods; the best, worst, and average over five runs with different 'seeds'. 53

5.7 SP- and random-IRR: ceiling clustering performance results on controlled distribution data 54

5.8 SP- and random-IRR: floor clustering performance results on controlled distribution data 55

5.9 SP- and random-IRR: kappa average precision results, dimensionality trained . . 56

5.10 SP- and random-IRR: kappa average precision results, dimensionality set to k . . 57

5.11 SP- and random-IRR: ceiling and floor clustering performance results, dimensionality trained 58

5.12 SP- and random-IRR: ceiling and floor clustering performance results, dimensionality set to k 58

5.13 SP- and random-IRR: CPU time in seconds and kappa average precision with respect to IRR 59

5.14 SP- and random-IRR: kappa average precision results with respect to VSM 60

5.15 SP- and random-IRR: ceiling clustering performance results with respect to VSM 61

5.16 SP- and random-IRR: floor clustering performance results excluding single-link with respect to VSM 62

6.1 The effect of changing p 64

6.2 Iteration to find the vector which produces an extreme value of $\sum_{j=1}^{n}(\mathbf{x}^T\mathbf{r}_j^{(i)})^p$ subject to $||\mathbf{x}||_2 = 1$. 65

6.3 PDSVD: kappa average precision and clustering performance, controlled distributions over two topics 66

6.4 PDSVD: absolute improvement in pair-wise kappa average precision over VSM, unrestricted distributions 67

6.5 PDSVD: document clustering performances, unrestricted distributions 67

6.6 Comparison of IRR and PDSVD performance results on unrestricted data by the paired t-test 68

7.1 Topical summary of a multi-document set: initial screen display. 72

7.2 Topical summary of a multi-document set: dynamics of the display when the mouse rolls over a document proxy. 74

7.3 Multi-document summarization: overview of the process 76

7.4 Topic vector creation. 77

7.5 Document vector division procedure. 77

7.6 Multi-document summarization: schematic illustration of data flow 79

Chapter 1

Introduction

The rapid increase in the availability of electronic documents has created a great demand for automated text analysis technologies such as document clustering and summarization. Representations enabling accurate measurement of semantic similarities between texts would greatly facilitate such technologies. For instance, to perform document clustering by standard clustering algorithms such as k-means, similarities (or closeness) between documents need to be measured. Also, similarity measures are commonly used to analyze connectivity in text for automatically generating a summary. (see e.g., the introduction of Mani and Maybury (1999)).

In this thesis we focus on representations in which vector directionality (from the origin) represents the semantics, or constituent concepts, of the corresponding entity. The goal is a general method for constructing *semantic spaces* in which entities that humans would judge to be semantically related are represented by vectors pointing in similar directions. This goal is to be accomplished without access to concept labels, since they are typically not available in many applications.

The vector space model (VSM) (Salton and McGill, 1983) is a widely-used classic method for constructing vector representations for documents. It encodes a document collection by a *term-document matrix* whose $[i, j]$th element indicates the association between the ith term and jth document. In typical applications of VSM, a term is a word, and a document is an article. However, it is possible to use different types of text units. For instance, phrases or word/character n-grams can be used as 'terms', and 'documents' can be paragraphs, sequences of n consecutive characters, or sentences. The essence of VSM is that it represents one type of text unit (*documents*) by its association with the other type of text unit (*terms*) where the association is measured by explicit evidence based on term occurrences in the documents. A geometric view of a term-document matrix is as a set of document vectors occupying a vector space spanned by terms; we call this vector space *VSM space*. The similarity between documents is typically measured by the cosine or inner product (defined later) between the corresponding vectors, which increases as more terms are shared.

A common criticism of VSM is that it does not take account of relations between terms. For instance, having "automobile" in one document and "car" in another document does not contribute to the similarity measure between these two documents. Furthermore, VSM space is not

ideal for representing terms and documents simultaneously, as terms (or documents consisting of a single term) are always mapped into orthogonal vectors producing zero similarity even when the terms are clearly semantically related.

Latent Semantic Indexing (LSI) (Deerwester et al., 1990; Dumais, 1991) attempts to overcome this shortcoming by choosing linear combinations of terms as dimensions of the representation space. LSI is often empirically successful, and has been applied to information retrieval and many language analysis tasks (e.g., Dumais and Nielsen (1992); Foltz and Dumais (1992); Berry et al. (1995a); Foltz et al. (1998b); Wolfe et al. (1998)), prompting theoretical studies to explain its effectiveness (Bartell et al., 1992; Story, 1996; Ding, 1999; Papadimitriou et al., 2000; Azar et al., 2001).

In typical usage of VSM, a vector in a VSM space is essentially "a bag of words" with weights. We observe that when the words in a "bag" are mutually related, they complement each other and may convey some sense, at least, more than a single word. For example, compare the bags "bank, river, rock, boat", "bank, investment, money, interest", and "bank". On the other hand, when the words are not mutually related (in a common sense), it is hard to obtain a sense from it. What is "bank, moon, tone, rice" about? A VSM space, spanned by m terms, is a set of all possible m-dimensional vectors where any term combination is allowed, whether it makes sense or not. Now, *what if we could constrain the representation space, so that only sensible combinations of terms can exist and any senseless ones are precluded?*

A *subspace* is a subset of a vector space under certain constraints. LSI represents a document by the projection of the VSM document vector onto an LSI subspace. According to the mathematical formulation of LSI, the term combinations which are less frequently occurring in the given document collection tend to be precluded from the LSI subspace. This fact together with our examples above suggests that one could argue that LSI does 'noise reduction' *if* it was true that less frequently co-occurring terms are less mutually-related, and therefore less sensible. However, consider a document collection in which most of the documents discuss "bank, investment, money, interest" while a few documents discuss "bank, river, rock, boat". Is "bank, river, rock, boat" less sensible than "bank, investment, money, interest"? We would say no. Still, one may conjecture that minority documents should be ignored in applications for analyzing overall trends in a collection. However, note that if some document is mostly precluded from a subspace, it is mis-represented rather than being ignored or removed, which would result in, for instance, degrading document cluster purity or precision of document retrieval.

In this thesis, we formally analyze LSI starting with modeling subspace-based approaches to the task of creating document representation spaces. Our model is centered around the notion of hidden correlations between topics (or themes) and documents. Based on this framework, we provide a new theoretical analysis of LSI to quantify the factors that affect its performance. In particular, we show *a precise relationship between the performance of LSI and the uniformity of the underlying distribution of documents over topics.*

As a consequence, we propose a novel alternative method, which we call *Iterative Residual Rescaling* (IRR), that attempts to compensate for non-uniformity in the topic-document distribution. IRR does this without prior knowledge of topics by repeatedly rescaling vectors to amplify the presence of documents predicted to be on less dominant topics. Furthermore, we propose several

extensions of IRR including a new document-sampling method, *Sampled* IRR (SP-IRR), which speeds up IRR while attempting to compensate for non-uniformity by strategic sampling.

To support our theoretical results, we present performance measurements both on document sets in which the topic-document distributions are carefully controlled, and on unrestricted datasets as would be found in application settings. The results validate our theoretical predictions, and the experiments as a whole provide strong evidence for the usefulness of our model in general and the effectiveness of IRR and SP-IRR.

Furthermore, we present a multi-document summarization system that uses the representation space derived by IRR as a framework. Our starting point is the assumption that a tightly connected (and therefore intuitively interpretable) set of coherent text units would act as a 'prompting' device when presented to the user in an appropriate context. We show that the vector space derived by IRR works as a framework to find such coherent texts, and effectively present them to the user.

As a summary, the results presented in this thesis are three-fold: first, a formal model and theoretical analysis of LSI, secondly, a novel method IRR (and extensions) and its performance evaluation in a variety of settings in comparison with LSI, and finally, a multi-document summarization system which demonstrates that IRR is immediately useful in applications.

The rest of the thesis is organized as follows. Section 1.1 summarizes the research contributions. Chapter 2 provides background materials: mathematical notions, in particular introduction to singular value decomposition which underlies LSI, and applications of and issues with LSI. Chapter 3 presents an analysis of LSI. Previous analyses of LSI are discussed in Section 3.5. The IRR algorithm and evaluation experiments are described in Chapter 4. In Chapter 5, we propose a novel document sampling method to speed IRR up for larger document collections, and report the experimental results. Chapter 6 discusses other ways to compensate for non-uniformity. Chapter 7 shows that IRR is immediately useful in applications by describing a multi-document summarization system based on it. Finally, previous studies related to document representation and LSI are discussed in Chapter 8. We conclude in Chapter 9. Appendix A provides the theorems and definitions from previous studies which we refer to, and the proofs of our theorems may be found in Appendix B.

1.1 Research Contributions

The main research contributions of this work are:

- **New analysis of LSI**
  We provide a new theoretical analysis of LSI to quantify the factors that affect its performance. In particular, we show a precise relationship between the performance of LSI and the uniformity of the underlying distribution of documents over topics.

- **IRR: New method of creating document representation spaces**
  Based on our analysis of LSI, we propose an alternative method, Iterative Residual Rescaling (IRR). It attempts to compensate for non-uniformity in the topic-document distribution

by repeatedly rescaling vectors to amplify the presence of documents predicted to be on less dominant topics without access to topic labels. We note that in comparison to LSI, IRR achieved up to 10.1% higher *kappa average precision* and enabled up to 8.7% better document clustering performance. The experiments as a whole provide strong evidence for the usefulness of our model in general and the effectiveness of IRR.

- **SP-IRR: IRR with Document Sampling**
  We present a new document sampling method, called Sampled IRR (SP-IRR), which speeds up IRR on relatively large document collections. For instance, SP-IRR achieved kappa average precision performance that rivals IRR (0.8% of degradation) while reducing computation time by 65%.

- **An application of IRR: Multi-document summarization**
  We present our multi-document summarization system, which relies on the vector space derived by IRR.

1.2 Bibliographic Notes

Portions of this thesis have appeared elsewhere.

Portions of Chapter 3 and portions of Chapter 4 are based on work described in "Iterative Residual Rescaling: An Analysis and Generalization of LSI" with Lillian Lee (Ando and Lee, 2001). This paper will appear in the proceedings of SIGIR'2001.

Portions of Chapter 4 are based on work described in "Latent Semantic Space: Iterative Scaling Improves Inter-document Similarity Measurement" (Ando, 2000), which appeared in the proceedings of SIGIR'2000.

Chapter 7 is adapted from the paper "Multi-document summarization by visualizing topical content" with Branimir Boguraev, Roy Byrd, and Mary Neff (Ando et al., 2000), which appeared in the proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization.

Chapter 2

Background: Introduction to SVD and LSI

This chapter introduces the background for our work: mathematical preliminaries including our notational conventions; the singular value decomposition, which is the mathematical basis of LSI; and LSI including applications and issues.

2.1 Mathematical Preliminaries

In this section, we introduce the mathematical materials on which vector space-based methods, including VSM, LSI, and our new method IRR, have their basis: the notions regarding matrices, and subspaces which are based on a geometrical perspective on matrices. More detail may be found in Golub and Van Loan (1996) and Stewart and Sun (1990).

2.1.1 Vectors and matrices

A bold lowercase letter (e.g., $\mathbf{y}$) denotes a vector. A vector is equivalent to a matrix having a single column. The ith entry of vector $\mathbf{y}$ is denoted by $\mathbf{y}_{[i]}$. A bold uppercase letter (e.g., $\mathbf{X}$) denotes a matrix; the corresponding bold lowercase letter with subscript i (e.g., $\mathbf{x}_i$) denotes the matrix's ith column vector. The $[i,j]$th entry of matrix $\mathbf{X}$ is denoted by $\mathbf{X}_{[i,j]}$. We write $\mathbf{X} \in \Re^{m \times n}$ when matrix $\mathbf{X}$ has m rows and n columns whose entries are real numbers.

A *diagonal matrix* $\mathbf{X} \in \Re^{n \times n}$ has zeroes in its non-diagonal entries, and is denoted by $\mathbf{X} = \text{diag}(\mathbf{X}_{[1,1]}, \mathbf{X}_{[2,2]}, \cdots, \mathbf{X}_{[n,n]})$. An *identity matrix* is a diagonal matrix whose diagonal entries are all one. We denote the identity matrix in $\Re^{m \times m}$ by $\mathbf{I}_m$. For any $\mathbf{X} \in \Re^{m \times n}$, $\mathbf{X}\mathbf{I}_n = \mathbf{I}_m\mathbf{X} = \mathbf{X}$. We omit the subscript when the dimensionality is clear from the context.

The *transpose* of matrix $\mathbf{X}$ is a matrix whose rows are the columns of $\mathbf{X}$, and is denoted by $\mathbf{X}^T$, i.e., $\mathbf{X}_{[i,j]} = (\mathbf{X}^T)_{[j,i]}$. The columns of $\mathbf{X}$ are *orthonormal* if $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. A matrix $\mathbf{X}$ is *orthogonal* if $\mathbf{X}^T\mathbf{X} = \mathbf{X}\mathbf{X}^T = \mathbf{I}$. Note that to be orthogonal (i.e., for both columns and rows to be orthonormal), a matrix must be square.

2.1.2 Vector length, cosine

The *vector 2-norm* of $\mathbf{x} \in \Re^m$ is defined by $||\mathbf{x}||_2 = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{\sum_{i=1}^m \left(\mathbf{x}_{[i]}\right)^2}$. We call it the *(vector) length* of $\mathbf{x}$. The *inner product* of $\mathbf{x}$ and $\mathbf{y}$ is $\mathbf{x}^T\mathbf{y}$. The *cosine* of $\mathbf{x}$ and $\mathbf{y}$ is the length-normalized inner product, defined by

$$\cos(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \frac{\mathbf{x}^T\mathbf{y}}{||\mathbf{x}||_2 \cdot ||\mathbf{y}||_2},$$

for $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$; note that $\cos(\mathbf{x}, \mathbf{y}) \in [-1, 1]$. A larger cosine value indicates that geometrically $\mathbf{x}$ and $\mathbf{y}$ point in similar directions. In particular, if $\mathbf{x} = \mathbf{y}$ then $\cos(\mathbf{x}, \mathbf{y}) = 1$, and $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* if and only if $\cos(\mathbf{x}, \mathbf{y}) = 0$.

2.1.3 Eigenvectors and eigenvalues

When $\mathbf{X}\mathbf{y} = \lambda\mathbf{y}$, $\mathbf{y}$ is an *eigenvector* of $\mathbf{X}$, λ is an *eigenvalue* associated with $\mathbf{y}$, and $(\mathbf{y}, \lambda)$ is called an *eigenpair* of $\mathbf{X}$. A matrix may have more than one eigenpair. Conventionally, we put the eigenpairs in non-ascending order of the eigenvalues, and refer to each value by its index, such as "the ith eigenvalue" and "the ith eigenvector".

2.1.4 Subspaces, column spaces, the orthogonal complement of a subspace

A *subspace* $\mathcal{X}$ of $\Re^m$ is a subset of $\Re^m$ closed under linear operations, i.e., if $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ then $a\mathbf{x} + b\mathbf{y} \in \mathcal{X}$ for any $a, b \in \Re$. We use a calligraphic uppercase letter (e.g., $\mathcal{X}$) to denote a subspace. The dimensionality of a subspace $\mathcal{X}$ is denoted by $\dim(\mathcal{X})$.

We let range($\mathbf{X}$) denote $\mathbf{X}$'s range (column space) $\{\mathbf{z} \mid \exists \mathbf{y} \text{ such that } \mathbf{z} = \mathbf{X}\mathbf{y}\}$. Note that $\mathbf{X}\mathbf{y} = \sum_i (\mathbf{y}_{[i]})\mathbf{x}_i$, so any vector in range($\mathbf{X}$) can be expressed as a linear combination of the columns of $\mathbf{X}$; when $\mathbf{X}$ has m rows, range($\mathbf{X}$) is a subspace of $\Re^m$. The *rank* rank($\mathbf{X}$) of $\mathbf{X}$ is defined as the dimensionality of range($\mathbf{X}$). If the columns of $\mathbf{B}$ are orthonormal and $\mathcal{X} = \text{range}(\mathbf{B})$, we say that the columns of $\mathbf{B}$ form an *orthonormal basis* of $\mathcal{X}$. An orthonormal basis defines a subspace uniquely, but a subspace has more than one orthonormal basis. If the columns of $\mathbf{B}$ form an orthonormal basis of $\mathcal{X}$, and if $\mathbf{R}$ is orthogonal, then the columns of $\mathbf{B}\mathbf{R}$ also form an orthonormal basis of $\mathcal{X}$. In a sense, the orthogonal matrix $\mathbf{R}$ rotates the columns of $\mathbf{B}$ without changing their relative positions, and they remain an orthonormal basis after being rotated.

We say that subspace $\mathcal{X}_1$ and $\mathcal{X}_2$ are *orthogonal* if $\mathbf{x}_1^T\mathbf{x}_2 = 0$ for any $\mathbf{x}_1 \in \mathcal{X}_1$ and $\mathbf{x}_2 \in \mathcal{X}_2$.

The *orthogonal complement* of $\mathcal{X} \subset \Re^m$ is $\mathcal{X}^\perp = \{\mathbf{z} \in \Re^m \mid \mathbf{z}^T\mathbf{x} = 0 \text{ for } \forall \mathbf{x} \in \mathcal{X}\}$. $\mathcal{X}^\perp$ is the complement of $\mathcal{X}$ in the sense that $\mathcal{X} \cap \mathcal{X}^\perp = \{\mathbf{0}\}$, and that any $\mathbf{y} \in \Re^m$ can be uniquely decomposed into $\mathbf{y} = \mathbf{x} + \mathbf{x}^\perp$ so that $\mathbf{x} \in \mathcal{X}$ and $\mathbf{x}^\perp \in \mathcal{X}^\perp$.

Dimensionality of a subspace Suppose that $\mathcal{X} \subset \Re^m$ and $\dim(\mathcal{X}) = h$. This means that $\mathcal{X}$ is spanned by exactly h mutually orthogonal vectors of m dimensions. The dimensionality of $\mathcal{X}$ is h, but note that the vectors in $\mathcal{X}$ have m entries (not h!). Now let the columns of $\mathbf{B} \in \Re^{m \times h}$

and $\mathbf{B}_\perp \in \Re^{m\times(m-h)}$ form orthonormal bases of $\mathcal{X}$ and $\mathcal{X}^\perp$, respectively. (Note that m and h exactly determine the dimensions of $\mathbf{B}$ and $\mathbf{B}_\perp$.) Then, for any $\mathbf{x} \in \mathcal{X}$, we have

$$\begin{bmatrix}\mathbf{B} & \mathbf{B}_\perp\end{bmatrix}^T \mathbf{x} = \begin{bmatrix}\mathbf{B}^T\mathbf{x} \\ \mathbf{B}_\perp^T\mathbf{x}\end{bmatrix} = \begin{bmatrix}\mathbf{B}^T\mathbf{x} \\ \mathbf{0}\end{bmatrix},$$

and $\mathbf{B}^T\mathbf{x}$ is of h dimensions. In fact, $[\mathbf{B} \quad \mathbf{B}_\perp]$ is orthogonal by construction and its columns (and rows) form an orthonormal basis of $\Re^m$, so $[\mathbf{B} \quad \mathbf{B}_\perp]^T\mathbf{x}$ is another representation of $\mathbf{x}$ with respect to the basis formed by the columns of $[\mathbf{B} \quad \mathbf{B}_\perp]$ (*basis conversion*). As we see, for any h-dimensional subspace $\mathcal{X}$, there exists a basis conversion by which all the vectors in $\mathcal{X}$ are converted to the vectors having zeroes in all but the first h entries. In a sense, $\mathbf{B}^T\mathbf{x}$ of h dimensions is more compact representation for vector $\mathbf{x} \in \mathcal{X}$.

2.1.5 Orthogonal projection

Suppose that $\mathcal{X} \subseteq \Re^m$, and let the columns of $\mathbf{B}$ form an orthonormal basis of $\mathcal{X}$. Then, the orthogonal projection operator onto $\mathcal{X}$ is defined for any $\mathbf{x} \in \Re^m$ by

$$\mathbf{P}_\mathcal{X}(\mathbf{x}) = \mathbf{B}\mathbf{B}^T\mathbf{x}.$$

Note that the projection operator is the same independent of choice of basis for $\mathcal{X}$.

$\mathbf{P}_\mathcal{X}$ leaves the vectors in $\mathcal{X}$ as they are (e.g., for $\mathbf{x} \in \mathcal{X}$, $\mathbf{P}_\mathcal{X}(\mathbf{x}) = \mathbf{x}$), and factors out those orthogonal to $\mathcal{X}$ (e.g., for $\mathbf{x} \in \mathcal{X}^\perp$, $\mathbf{P}_\mathcal{X}(\mathbf{x}) = \mathbf{0}$).

The projection operator works for both vectors and matrices. A matrix with m rows is projected column by column as follows:

$$\mathbf{P}_\mathcal{X}(\mathbf{X}) = [\mathbf{P}_\mathcal{X}(\mathbf{x}_1) \quad \mathbf{P}_\mathcal{X}(\mathbf{x}_2) \quad \cdots].$$

Sometimes it is convenient to express the projection by an orthonormal basis of a subspace. Let $\{\mathbf{b}_1, \ldots, \mathbf{b}_r\}$ form an orthonormal basis of subspace $\mathcal{X}$. When $\mathbf{P}_\mathcal{X}(\mathbf{X})$ is defined, we let $\mathrm{proj}(\mathbf{X}, \{\mathbf{b}_1, \ldots, \mathbf{b}_r\})$ denote the projection of $\mathbf{X}$ onto subspace $\mathcal{X}$.

2.1.6 Frobenius norm and matrix 2-norm

Let $\mathbf{X} \in \Re^{m\times n}$, and let $h = \mathrm{rank}(\mathbf{X})$. The *Frobenius norm (F-norm)* is the square root of the sum of squares of the entries:

$$||\mathbf{X}||_\mathrm{F} \stackrel{\mathrm{def}}{=} \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mathbf{X}_{[i,j]}\right)^2}.$$

The matrix *2-norm* is defined by the vector 2-norm as

$$||\mathbf{X}||_2 \stackrel{\mathrm{def}}{=} \max_{||\mathbf{y}||_2=1} ||\mathbf{X}\mathbf{y}||_2.$$

The F-norm, 2-norm, and the maximum entry are related by

$$||\mathbf{X}||_2 \leq ||\mathbf{X}||_{\mathrm{F}} \leq \sqrt{h}||\mathbf{X}||_2\,,$$
$$\max_{i,j}|\mathbf{X}_{[i,j]}| \leq ||\mathbf{X}||_2 \leq \sqrt{mn}\max_{i,j}|\mathbf{X}_{[i,j]}|\,.$$

In particular, for any rank-one matrix, the 2-norm and F-norm are equal. We relate the F- and 2-norm to the singular value decomposition in the next section.

2.2 Singular value decomposition (SVD)

Singular value decomposition (SVD) is the mathematical technique underlying LSI. In this section, we introduce important properties of SVD related to our work.

2.2.1 Singular value decomposition, singular values, singular vectors

The *singular value decomposition* factors a rank-h matrix $\mathbf{D} \in \Re^{m\times n}$ into the product:

$$\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\,,$$

such that the columns of $\mathbf{U} \in \Re^{m\times h}$ and $\mathbf{V} \in \Re^{n\times h}$ are orthonormal and $\mathbf{\Sigma} \in \Re^{h\times h}$ is diagonal. Each diagonal entry of $\mathbf{\Sigma}$, which we denote by $\sigma_i = \mathbf{\Sigma}_{[i,i]}$, is positive. Following convention, we assume that $\sigma_1 \geq \cdots \geq \sigma_h$. We call σ_i the ith *singular value* of $\mathbf{D}$ and denote it by $\sigma_i[\mathbf{D}]$. Note that $\sigma_i[\mathbf{D}^T\mathbf{D}] = \sigma_i[\mathbf{D}\mathbf{D}^T] = \sigma_i[\mathbf{D}]^2$ for any i. The columns of $\mathbf{U}$ (and $\mathbf{V}$) are called the *left (right) singular vectors* of $\mathbf{D}$, and related by $\sigma_i[\mathbf{D}]\ \mathbf{u}_i = \mathbf{D}\mathbf{v}_i$. It is easy to show that $(\mathbf{u}_i, \sigma_i[\mathbf{D}]^2)$ is an eigenpair of $\mathbf{D}\mathbf{D}^T$, and that $(\mathbf{v}_i, \sigma_i[\mathbf{D}]^2)$ is an eigenpair of $\mathbf{D}^T\mathbf{D}$.

Letting $\mathbf{U}' = [\,\mathbf{U} \quad \mathbf{U}_o\,] \in \Re^{m\times m}$ and $\mathbf{V}' = [\,\mathbf{V} \quad \mathbf{V}_o\,] \in \Re^{n\times n}$ so that $\mathbf{U}'$ and $\mathbf{V}'$ are orthogonal, $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ can be rewritten as

$$\mathbf{D} = \mathbf{U}'\begin{bmatrix}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{V}'^T\,.$$

This is another notation for SVD used in some work. The columns of $\mathbf{U}_o$ (and $\mathbf{V}_o$) can be regarded as left (right) singular vectors associated with *zero singular values*. We use the notation omitting all the zero singular values unless we specify otherwise.

The SVD is symmetric, so that the right singular vectors of $\mathbf{D}$ are the left singular vectors of $\mathbf{D}^T$. In the following sections we assume $n < m$, and we focus on the properties of the left singular vectors.

For convenience, when we refer to $\sigma_i[\mathbf{D}]$ for $i > \mathrm{rank}(\mathbf{D})$, it is understood to be zero.

2.2.2 Geometrical view of SVD, relations to matrix norms

To interpret SVD geometrically, it helps to see that the singular values $\sigma_i[\mathbf{D}]$ and the left singular vectors $\mathbf{u}_i$ of $\mathbf{D}$ form a recursive sequence as follows:

$$\mathbf{u}_i = \arg\max_{||\mathbf{x}||_2=1} ||\mathrm{proj}(\mathbf{R}^{(i)}, \{\mathbf{x}\})||_2\,, \quad \sigma_i[\mathbf{D}] = \max_{||\mathbf{x}||_2=1} ||\mathrm{proj}(\mathbf{R}^{(i)}, \{\mathbf{x}\})||_2\,,$$

where

$$\begin{aligned} \mathbf{R}^{(1)} &= \mathbf{D}\,, \\ \mathbf{R}^{(i+1)} &= \mathbf{R}^{(i)} - \mathrm{proj}(\mathbf{R}^{(i)}, \{\mathbf{u}_i\})\,. \end{aligned}$$

The matrix $\mathbf{R}^{(i)}$ stands for a *residual*, initially $\mathbf{D}$. The ith left *singular pair* $(\mathbf{u}_i, \sigma_i[\mathbf{D}])$ is obtained by searching for the one-dimensional subspace onto which the 2-norm[1] of the projection of $\mathbf{R}^{(i)}$ is maximized. The maximum projection thus found is subtracted from $\mathbf{R}^{(i)}$, which defines $\mathbf{R}^{(i+1)}$ and consequently the next singular pair $(\mathbf{u}_{i+1}, \sigma_{i+1}[\mathbf{D}])$. Noting that $\mathrm{proj}(\mathbf{R}^{(i)}, \{\mathbf{u}_i\}) = \mathrm{proj}(\mathbf{D}, \{\mathbf{u}_i\})$ (because of mutual orthogonality of the $\mathbf{u}_i$'s), $\mathbf{D}$ can be expressed as a sum of the projections:

$$\begin{aligned} \mathbf{D} &= \mathrm{proj}(\mathbf{D}, \{\mathbf{u}_1\}) + \mathrm{proj}(\mathbf{D}, \{\mathbf{u}_2\}) + \cdots + \mathrm{proj}(\mathbf{D}, \{\mathbf{u}_h\})\,, \\ \sigma_i[\mathbf{D}] &= ||\mathrm{proj}(\mathbf{D}, \{\mathbf{u}_i\})||_2\,. \end{aligned}$$

(Recall that $h = \mathrm{rank}(\mathbf{D})$, so $\sigma_i[\mathbf{D}] = 0$ for $i > h$.) This view of the SVD is related to important properties of the F-norm and 2-norm:

$$\begin{aligned} ||\mathbf{D}||_\mathrm{F} &= \sqrt{\sigma_1[\mathbf{D}]^2 + \cdots + \sigma_h[\mathbf{D}]^2} \\ ||\mathbf{D}||_2 &= \max_{1\le i\le h} \sigma_i[\mathbf{D}] = \sigma_1[\mathbf{D}]\,. \end{aligned}$$

In a sense, the F-norm measures a total amount of $\mathbf{D}$'s one-dimensional projections, and the 2-norm measures the largest one.

2.2.3 Rank-k approximation by SVD

A matrix can be optimally approximated by computing its SVD. Set $\mathbf{\Sigma}'_k \in \Re^{h\times h}$ for given $k < h$, so that

$$\mathbf{\Sigma}'_k = \mathrm{diag}(\sigma_1[\mathbf{D}], \cdots, \sigma_k[\mathbf{D}], 0, \cdots, 0)\,.$$

Then,

$$\mathbf{D}_k = \mathbf{U}\mathbf{\Sigma}'_k\mathbf{V}^T = \mathrm{proj}(\mathbf{D}, \{\mathbf{u}_1\}) + \mathrm{proj}(\mathbf{D}, \{\mathbf{u}_2\}) + \cdots + \mathrm{proj}(\mathbf{D}, \{\mathbf{u}_k\})\,,$$

which yields the optimum rank-k approximation of $\mathbf{D}$ in the sense that

$$||\mathbf{D} - \mathbf{D}_k||_2 = \min_{\mathrm{rank}(\mathbf{X})=k} ||\mathbf{D} - \mathbf{X}||_2 = \sigma_{k+1}[\mathbf{D}]\,,$$

[1] In fact, $||\mathrm{proj}(\mathbf{R}^{(i)}, \{\mathbf{x}\})||_2 = ||\mathrm{proj}(\mathbf{R}^{(i)}, \{\mathbf{x}\})||_\mathrm{F}$, as the 2-norm and F-norm of a rank-one matrix are equal. For clarity, we write only the 2-norm in the equation.

$$||\mathbf{D}-\mathbf{D}_k||_{\mathrm{F}} = \min_{\mathrm{rank}(\mathbf{X})=k} ||\mathbf{D}-\mathbf{X}||_{\mathrm{F}} = \sqrt{\sigma_{k+1}[\mathbf{D}]^2+\cdots+\sigma_n[\mathbf{D}]^2}\,.$$

Note that because of the orthogonality of $\mathbf{u}_i$'s,

$$\begin{aligned}\mathbf{D}_k &= \mathrm{proj}(\mathbf{D},\{\mathbf{u}_1\})+\mathrm{proj}(\mathbf{D},\{\mathbf{u}_2\})+\cdots+\mathrm{proj}(\mathbf{D},\{\mathbf{u}_k\})\\ &= \mathrm{proj}(\mathbf{D},\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_k\})\,.\end{aligned}$$

That is, the subspace spanned by the first k left singular vectors of $\mathbf{D}$ maximally preserves $\mathbf{D}$ for a given k.

$\mathbf{D}_k$ can also be written as $\mathbf{D}_k = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$ where $\mathbf{U}_k$ and $\mathbf{V}_k$ are the first k left and right singular vectors of $\mathbf{D}$, respectively, and $\mathbf{\Sigma}_k = \mathbf{\Sigma}_{[1:k,1:k]}$, as in Figure 2.1 (b).

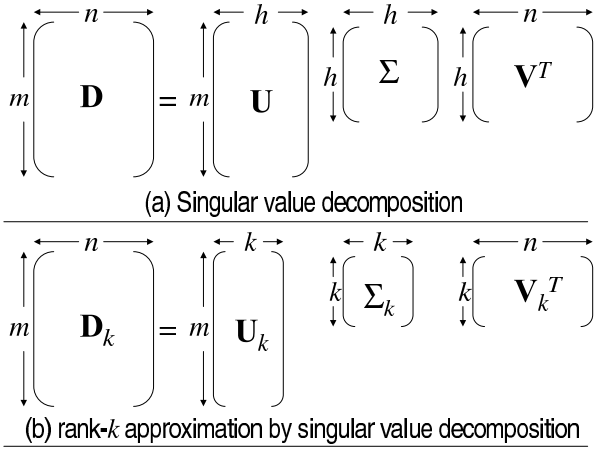

Figure 2.1: SVD and rank-k approximation by SVD.

2.2.4 Geometrical interpretation of left singular vectors

Let us investigate a left singular vector $\mathbf{u}_i$ in more detail, as it will help to introduce our new algorithm IRR later. Recall from Section 2.2.2 that $\mathbf{u}_i$ satisfies

$$\mathbf{u}_i = \arg\max_{||\mathbf{x}||_2=1} ||\mathrm{proj}(\mathbf{R}^{(i)},\{\mathbf{x}\})||_2 = \arg\max_{||\mathbf{x}||_2=1} ||\mathrm{proj}(\mathbf{R}^{(i)},\{\mathbf{x}\})||_{\mathrm{F}}\,. \tag{2.1}$$

As we have (for unit vector $\mathbf{x}$)

$$\begin{aligned}||\mathrm{proj}(\mathbf{R}^{(i)},\{\mathbf{x}\})||_{\mathrm{F}} &= \sqrt{\sum_{j=1}^{n}||\mathrm{proj}(\mathbf{r}_j^{(i)},\{\mathbf{x}\})||_2^2} = \sqrt{\sum_{j=1}^{n}||\mathbf{x}\mathbf{x}^T\mathbf{r}_j^{(i)}||_2^2}\\ &= \sqrt{\sum_{j=1}^{n}\left(||\mathbf{r}_j^{(i)}||_2\cos(\mathbf{x},\mathbf{r}_j^{(i)})\right)^2},\end{aligned}$$

(2.1) can be rewritten by

$$\mathbf{u}_i = \arg \max_{||\mathbf{x}||_2=1} \sum_{j=1}^{n} \left(||\mathbf{r}_j^{(i)}||_2 \cos(\mathbf{x}, \mathbf{r}_j^{(i)}) \right)^2 . \tag{2.2}$$

Recall that the cosine is larger for closer vectors. In a sense, $\mathbf{u}_i$ is a weighted average of *residual vectors* $\mathbf{r}_j^{(i)}$, where longer residuals receive greater weight. We will use equation (2.2) later in Section 4.1.

2.2.5 Computation of SVD

According to (Golub and Van Loan, 1996), the estimation of flop counts to obtain $\mathbf{U}$ and $\mathbf{\Sigma}$ when a matrix is in $\Re^{m \times n}$ and $m > n$ is $14m^2n - 8mn^2$ by the Golub-Reinsch SVD (Golub and Reinsch, 1970), and $4m^2n + 13n^3$ by R-SVD (Chan, 1982). These estimations are for the case that all the singular values and vectors are computed.

As we see in the next section, LSI needs to compute several (typically 100 to 500 for information retrieval) singular vectors associated with the largest singular values of a sparse matrix. Lanczos method, which originates from a method attributed to Lanczos (1950), is popular when only several largest eigenvalues (and associated eigenvectors) need to be computed as in LSI; recall from Section 2.2.1 that the SVD computation can be reduced to the eigenvector computation. A rough sketch of Lanczos method is as follows. Suppose that we need to compute the first k eigenvectors of matrix $\mathbf{X} \in \Re^{n \times n}$. The *Lanczos iteration* generates tridiagonal matrices $\mathbf{T}_i \in \Re^{i \times i}$ whose eigenvalues approximates $\mathbf{X}$'s first several eigenvalues with progressively better accuracies as i increases. $\mathbf{X}$'s eigenvectors are computed from the eigenvectors of $\mathbf{T}'_k$ for $k' \geq k$ and *Lanczos vectors*, which are obtained in the computation of $\mathbf{T}'_k$ simultaneously. Thus, the eigenvector computation of a $n \times n$ matrix is reduced to that of a $k' \times k'$ matrix for $k' < n$. Still, the computation of LSI is known to be relatively expensive when a term-document matrix is large.

Berry (1992) studied the CPU time of several methods specifically in the setting of LSI, and concluded that the single-vector Lanczos algorithm via an $\mathbf{A}^T\mathbf{A}$ eigensystem is the fastest to achieve low to moderate accuracy in the computing environment he used (the Cray-2S/4-128 and Alliant FX/80). SVDPACK (Berry et al., 1993) is a software package to compute the SVD of large sparse matrices that is publicly available via internet.

2.3 Latent Semantic Indexing (LSI)

We introduce LSI, which in this thesis we formally analyze and refer to as baseline in the evaluation of our new method. We describe LSI in relation to SVD (Section 2.3.1), and summarize its applications and research issues (Section 2.3.2).

2.3.1 SVD and Latent Semantic Indexing (LSI)

One of the first studies that introduced Latent Semantic Indexing (LSI) is Dumais et al. (1988). It reported promising evaluation results on an information retrieval task. LSI applies SVD to

a term-document matrix to construct a subspace, called *LSI subspace*, spanned by the first few left singular vectors. New document vectors (and query vectors) are obtained by orthogonally projecting the corresponding vectors in a VSM space (spanned by terms) onto the LSI subspace. As shown in Section 2.2.3, the projection of the term-document matrix onto the k-dimensional LSI subspace is the optimum rank-k approximation to minimize the 2- and F-norm of a residual matrix.

In the literature, this method is typically called "Latent Semantic Indexing (LSI)" when the application is information retrieval, and also often called "*Latent Semantic Analysis (LSA)*" when used for other types of applications. However, it seems that these two names are not strictly distinguished, presumably because the essence of the method is to *analyze* texts via *indexing* by latent semantics. For clarity, we consistently use LSI for the framework that measures similarities between the text units in the subspace obtained by computing the SVD of a term-document matrix. As we discussed in Chapter 1, we take a broader definition of term-document matrices, which includes word-sentence matrices, phrase-paragraph matrices, and so on.

The fact that VSM produces zero similarity between text units which share no terms is an issue, especially in the information retrieval task of measuring the relevance between documents and a query submitted by a user (*user query*). Typically, a user query is short and does not cover all the vocabulary for the target concept. Using VSM, "car" in a query and "automobile" in a document do not contribute to retrieving this document (*synonym problem*). The illustrative examples in (Deerwester et al., 1990; Landauer et al., 1998a) show that LSI may solve this synonym problem by producing positive similarity between related documents sharing no terms.

As the LSI subspace captures the most significant factors (associated with the largest singular values) of a term-document matrix, it is expected to capture the relations of the most frequently co-occurring terms[2]. In this sense, LSI can be regarded as a corpus-based statistical method. However, the relations among terms are not modeled explicitly in the computation of LSI subspace, which might make it hard to understand LSI in general. Although the fact that an LSI subspace provides the best low rank approximation of the term-document matrix is often referred to, this does not imply that the LSI subspace approximates the 'true' semantics of documents. Generally, LSI is explained as a way of 'noise' reduction without a precise definition of noise.

2.3.2 Applications of and Issues with LSI

There have been a number of studies in applying LSI to a variety of text related tasks: the traditional IR task such as the ad hoc track of TREC (Deerwester et al., 1990; Dumais, 1991; Dumais, 1993; Dumais, 1994; Dumais, 1995); cross-language information retrieval (Landauer and Littman, 1990); information filtering (Foltz, 1990; Foltz and Dumais, 1992; Dumais, 1995); assigning the submitted papers to the reviewers (Dumais and Nielsen, 1992); analyzing the essays to trace the source of knowledge (Foltz et al., 1996; Foltz, 1996); grading essays (Foltz, 1996; Landauer et al., 1997; Landauer et al., 1998b); grading students' answers (Wiemer-Hastings et al., 1998; Wiemer-Hastings et al., 1999; Wiemer-Hastings, 1999; Wiemer-Hastings, 2000); measuring

[2]This fact is understood when we realize that the SVD factors a term-document matrix into the largest one-dimensional projections of the document vectors as shown in Section 2.2, and that each of the document vectors can be regarded as a linear combination of terms.

the coherence and comprehensibility of texts by the mutual relatedness of the sentences (Foltz, 1996; Foltz et al., 1998b); matching the background level of readers and the difficulty level of texts (Wolfe et al., 1998); document clustering (Schütze and Silverstein, 1997; Soboroff et al., 1998; Kurimo, 2000); a text mining-type task (Jiang et al., 1999b); morphological analysis (Schone and Jurafsky, 2000a; Schone and Jurafsky, 2000b). The text units compared in the LSI subspace are either those from which the LSI subspace was constructed or 'unseen' text units: document clustering is the former, and the latter includes the traditional IR task to measure query-document relevance. A survey of the LSI applications (up to 1995) may be found in Berry et al. (1995a) and Berry et al. (1995c).

Several studies have reported that the performance of LSI rivals those of (non-expert) humans. For instance, Landauer and Dumais (1997) show that the synonym section of the Test of English as a Foreign Language were answered 64.4% correctly by using the inter-word similarities measured by LSI, while the average of the U. S. college students from non-English speaking countries was 65.5%. In the Autotutor experiments, LSI rates student answers by comparing them with the ideal answers. When the parameters are chosen appropriately, the correlation of ratings between LSI and humans approaches that of intermediate-level domain experts (Wiemer-Hastings, 1999). Nevertheless, negative results have been also reported. For instance, in Dumais et al. (1988)'s experiments, LSI did not improve the information retrieval performance over VSM on one corpus, while it achieved significant improvement on another corpus.

The implementation of LSI has been empirically studied. Dumais (1991) investigated the effects of several term weighting schemes to instantiate the input term-document matrix. Evaluation was based on the precision-recall curves on the retrieval tasks with the dimensionality of the LSI subspace being fixed. Several term-weighting schemes, which combine global weights (statistics in the collection) and local weights (statistics within each document), were investigated. *LogEntropy*, which is a combination of a local log weight and a global entropy weight, showed superiority over the combinations of the local term frequency and global weighting schemes or no global weighting. Two well-known global weightings (GfIdf and Normal) produced the performance worse than no global weighting.

Importantly, several studies have shown that the performance of LSI significantly varies over the dimensionalities of the LSI subspace (Deerwester et al., 1990; Dumais, 1991; Dumais, 1995). In practice, the dimensionality is determined experimentally, or 'blindly' picked by following the previous work. A systematic way to select dimensionality has been regarded as an important open issue (Zha et al., 1998). To approach this issue, essentially we need to understand the meaning of dimensions and formalize the notion of 'noise' or what should be factored out. This leads to a more fundamental question of why and when LSI is effective if it is effective. There have been several attempts to formally analyze LSI (Bartell et al., 1992; Story, 1996; Ding, 1999; Papadimitriou et al., 2000; Azar et al., 2001), which we will discuss later in Section 3.5.

Another practical issue of LSI is that computing SVD is expensive, which makes LSI practically infeasible for some applications (Budzik and Hammond, 1999). It has prompted the study in extending LSI for faster computation (Kolda and O'Leary, 1996a; Kolda and O'Leary, 1996b; Kolda and O'Leary, 1998; Jiang et al., 1999a), or partial computation at the arrival of new documents instead of recomputing SVD for all (Berry et al., 1995b; Zha and Simon, 1999; Witter and Berry, 1998).

2.4 Preview

We have described the background materials underlying this thesis. Starting from the next chapter, we will show our results: a formal analysis of LSI based on our model of subspace-based methods, a new alternative method which derives from this analysis, several extensions of our method including a document sampling method, and an application of our method to a multi-document summarization task.

Chapter 3

An Analysis of LSI

We present a framework for the task of creating document representation spaces using subspace projection to uncover document similarities with respect to their hidden topical content. Based on this framework, we provide a new theoretical analysis showing a precise relationship between the performance of LSI and the uniformity of the underlying distribution of documents over topics.

Notational conventions When a document collection has been specified, the symbols n, m, and k refer to the number of documents, the number of terms, and the number of topics the collection contains, respectively. We assume that $k \leq n \leq m$, as in typical situations.

3.1 Topic-based similarities

Our new model frames the question of how to select a good document representation given initial term-document relevance information.

Fix an n-document collection and corresponding term-document matrix $\widetilde{\mathbf{A}} \in \Re^{m \times n}$. We assume that the collection encompasses a total of k underlying topics. The collection and the topics are *hidden* from the method. Let $\mathrm{rel}(t, d)$ denote the (real-valued) degree of relevance between the tth topic and the dth document. We assume that for each document d, $\sqrt{\sum_{t=1}^{k} |\mathrm{rel}(t, d)|^2} = 1$, so that all the documents have equal total relevance weight. We then define the hidden *topic-based similarity* as:

$$\mathrm{sim}(d, d') = \sum_{t=1}^{k} \mathrm{rel}(t, d)\mathrm{rel}(t, d') \,,$$

which measures the similarity between the dth and d'th documents with respect to their topical content. It is convenient to summarize these similarities in a single matrix $\mathbf{S} \in \Re^{n \times n}$, where $\mathbf{S}_{[d,d']} = \mathrm{sim}(d, d')$.

The topic-based similarity is the hidden *true similarity* that a method seeks to uncover. Note that such a definition of true similarity ensures that we seek semantic similarity instead of, say, similarity in the writing style. When the model is applied to specific instances, the topics and

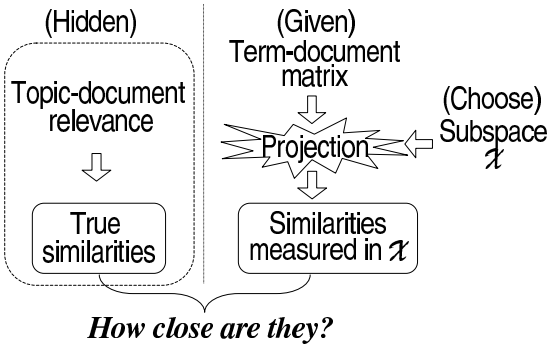

Figure 3.1: Framework for the analysis of LSI.

topic-document relevance should come from human judgments; however, their origin does not affect the analysis. Also note that although we assume the existence of underlying topics as the basis for the true document similarities, in contrast to other analyses (e.g., Ding (1999), Papadimitriou et al. (2000), Azar et al. (2001)) we do *not* assume that there is an underlying generative or probabilistic model that *creates* the term-document matrix $\widetilde{\mathbf{A}}$.

3.2 The optimum subspace

We formulate the ultimate goal of subspace-based algorithms, such as LSI, as choosing some subspace such that projecting a given $\widetilde{\mathbf{A}}$ onto this subspace creates new document vectors whose measured similarities (i.e., cosines) more closely correspond to the true topic-based similarities.

More formally, as in Section 2.1 let $\mathbf{P}_{\mathcal{X}}$ denote the *projection operator* from $\Re^m$ onto a subspace $\mathcal{X} \subset \Re^m$. Then, $\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{A}})$ is the projection of the term-document matrix onto $\mathcal{X}$, and $\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_i)$ (which is a column of $\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{A}})$) is a new document vector. Hence, after projection onto $\mathcal{X}$, the similarity between the ith and jth documents is measured by

$$\cos(\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_i), \mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_j)) = \frac{\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_i)^T \mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_j)}{||\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_i)||_2 ||\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_j)||_2}.$$

The *document representation problem* is as follows: given $\widetilde{\mathbf{A}}$ — but *not* $\mathbf{S}$ (the true similarity matrix) or even any knowledge of what the underlying topics are — find a subspace $\mathcal{X}$ such that the entries of the *error matrix*

$$\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}) = \mathbf{S} - \mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{A}})^T \mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{A}})$$

are small. Note that the $[i,j]$th entry of $\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{A}})^T \mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{A}})$ is the inner product between the ith and jth new document vectors $\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_i)^T \mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_j)$. It suffices to consider the inner products rather than cosine because for any $\mathcal{X}$, if $\max_{i,j} |\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})_{[i,j]}| = \rho < 1$, then

$$\frac{\mathbf{S}_{[i,j]} - \rho}{1+\rho} \leq \cos(\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_i), \mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_j)) \leq \frac{\mathbf{S}_{[i,j]} + \rho}{1-\rho}.$$

(See footnote for the proof[1] .)

For subspace $\mathcal{X}$ we define the *average error*[2]

$$E_{\mathcal{X}}^{avg} \stackrel{\text{def}}{=} \frac{||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})||_{\text{F}}}{\sqrt{n}}$$

and the *maximum error*

$$E_{\mathcal{X}}^{max} \stackrel{\text{def}}{=} ||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})||_2 \,.$$

We call these quantities the 'maximum' and 'average' because

$$\begin{array}{rcccl} (E_{\mathcal{X}}^{max})^2 & = & ||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})||_2^2 & = & \max_i(\sigma_i[\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})]^2)\,, \\ (E_{\mathcal{X}}^{avg})^2 & = & ||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})||_{\text{F}}^2/n & = & \sum_{i=1}^{n}\sigma_i[\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})]^2/n\,. \end{array}$$

Intuitively, $E_{\mathcal{X}}^{max}$ measures the error regarding the most problematic documents, while $E_{\mathcal{X}}^{avg}$ measures the overall error. Note that these quantities are closely related mutually and to the maximum entry of $\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})$, by

$$\begin{array}{rcccl} E_{\mathcal{X}}^{max}/\sqrt{n} & \leq & E_{\mathcal{X}}^{avg} & \leq & E_{\mathcal{X}}^{max}\,, \\ E_{\mathcal{X}}^{max}/n & \leq & \max_{i,j}|\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})_{[i,j]}| & \leq & E_{\mathcal{X}}^{max}\,, \end{array}$$

derived from the properties of matrix norm; see Section 2.1.6.

We define the optimum subspace, which serves as the standard for comparison in our analysis, uniquely by

$$\mathcal{X}_{opt} \stackrel{\text{def}}{=} \arg\min_{\mathcal{X}\subseteq\text{range}(\widetilde{\mathbf{A}})} E_{\mathcal{X}}^{avg}\,,$$

resolving ties by the lowest dimensionality and then arbitrarily. The condition $\mathcal{X} \subseteq \text{range}(\widetilde{\mathbf{A}})$ excludes the dimensions that have no effect on $\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}$ [3]. Also, this condition ensures that $\dim(\mathcal{X}_{opt}) = \text{rank}(\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}}))$. Note that the average and maximum *optimum errors* $E_{\mathcal{X}_{opt}}^{avg}$ and $E_{\mathcal{X}_{opt}}^{max}$ need not be zero, as it may be impossible to project the given term-document matrix in such a way as to perfectly recover the true document similarities.

In the next sections, we study how good an LSI subspace is for document representation, in comparison with the optimum subspace.

3.3 Non-uniformity and LSI

We use the framework (and notation) of the previous sections to quantify the factors that affect the quality of the LSI subspaces. The main outline of our argument is to first show relations

---

[1] To simplify notation, let $\mathbf{d}_i$ be $\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{a}}_i)$. By definition, $\cos(\mathbf{d}_i,\mathbf{d}_j) = (\mathbf{d}_i^T\mathbf{d}_j)/\sqrt{(\mathbf{d}_i^T\mathbf{d}_i)(\mathbf{d}_j^T\mathbf{d}_j)}$. By assumption, $|\mathbf{S}_{[i,j]} - \mathbf{d}_i^T\mathbf{d}_j| \leq \rho$ and $|1 - \mathbf{d}_i^T\mathbf{d}_i| = |\mathbf{S}_{[i,i]} - \mathbf{d}_i^T\mathbf{d}_i| \leq \rho$. Therefore, we have $\mathbf{S}_{[i,j]} - \rho \leq \mathbf{d}_i^T\mathbf{d}_j \leq \mathbf{S}_{[i,j]} + \rho$ and $1-\rho \leq \sqrt{(\mathbf{d}_i^T\mathbf{d}_i)(\mathbf{d}_j^T\mathbf{d}_j)} \leq 1+\rho$. The conclusion immediately follows from $0 \leq \rho < 1$.

[2] Recall from Section 2.1 that the Frobenius norm $\|\cdot\|_{\text{F}}$ is the square root of the sum of squares of the entries.

[3] More formally, $\mathbf{P}_{(\text{range}(\widetilde{\mathbf{A}}))^\perp}(\widetilde{\mathbf{A}}) = \mathbf{0}$ and $\mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{A}}) = \mathbf{P}_{\mathcal{X}\cap\text{range}(\widetilde{\mathbf{A}})}(\widetilde{\mathbf{A}})$.

m		number of terms
n		number of documents
k		number of topics
$\widetilde{\mathbf{A}} \in \Re^{m \times n}$		term-document matrix
$\mathrm{rel}(t,d)$		relevance of the dth document to the tth topic
$\mathrm{sim}(d,d')$	$= \sum_{t=1}^{k} \mathrm{rel}(t,d)\mathrm{rel}(t,d')$	topic-based similarity between the dth and d'th documents
$\mathbf{S} \in \Re^{n \times n}$	$\mathbf{S}_{[d,d']} = \mathrm{sim}(d,d')$	True similarity matrix
$\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})$	$= \mathbf{S} - \mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{A}})^T \mathbf{P}_{\mathcal{X}}(\widetilde{\mathbf{A}})$	error matrix for $\mathcal{X}$
$E_{\mathcal{X}}^{avg}$	$= \|\|\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})\|\|_{\mathrm{F}}/\sqrt{n}$	average error of $\mathcal{X}$
$E_{\mathcal{X}}^{max}$	$= \|\|\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X})\|\|_2$	maximum error of $\mathcal{X}$
$\varepsilon_{\mathrm{VSM}}^{max}$	$= \left(E_{\mathcal{X}_{\mathrm{VSM}}}^{max}\right)^{1/2}$	root original error
$\mathcal{X}_{opt}$	$= \arg\min_{\mathcal{X} \subseteq \mathrm{range}(\widetilde{\mathbf{A}})} E_{\mathcal{X}}^{avg}$	optimum subspace
$\mathcal{X}_{\mathrm{VSM}}$	$= \Re^m$	VSM space
$\mathcal{X}_{LSI}$		LSI subspace
$\mathcal{T}_i$	$= \left(\sum_{d=1}^{n} \mathrm{rel}(i,d)^2\right)^{1/2}$	the ith largest topic dominance
μ	$= \left(\sum_{t \neq t'} \left(\sum_{d=1}^{n} \mathrm{rel}(t,d)\mathrm{rel}(t',d)\right)^2\right)^{1/2}$	topic mingling

Figure 3.2: Definitions and notational conventions.

between certain singular values and certain quantities, and then show how the distance between the LSI-subspace and the optimal subspace relates to these singular values. Proofs of our results, which make use of invariant subspace perturbation theorems (Davis and Kahan, 1970; Stewart and Sun, 1990; Golub and Van Loan, 1996), may be found in the appendix.

The bounds we derive are based on the following quantities.

Original error The bound we derive incorporates the *original error* contained in the given term-document matrix. Intuitively, if a "bad" term-document matrix is received as input, one cannot expect LSI to do well. More formally, let $\mathcal{X}_{\mathrm{VSM}}$ denote the VSM space, $\Re^m$ (spanned by terms). Then, as $\mathbf{P}_{\mathcal{X}_{\mathrm{VSM}}}(\widetilde{\mathbf{A}}) = \widetilde{\mathbf{A}}$,

$$E_{\mathcal{X}_{\mathrm{VSM}}}^{max} = ||\mathbf{S} - \widetilde{\mathbf{A}}^T \widetilde{\mathbf{A}}||_2$$

measures the maximum original error. To facilitate our presentation, we set $\varepsilon_{\mathrm{VSM}}^{max} = \left(E_{\mathcal{X}_{\mathrm{VSM}}}^{max}\right)^{1/2}$, and call it *root original error*.

Topic dominance Recall that we are dealing with a document collection with k underlying topics and $\mathrm{rel}(t,d)$ is the relevance of the dth document to the tth topic. An important quantity

in our analysis is the *dominance* $\mathcal{T}_i$ of a given topic i:

$$\mathcal{T}_i = \left(\sum_{d=1}^{n} \mathrm{rel}(i,d)^2 \right)^{1/2} .$$

Larger topic dominance indicates that more documents are more relevant to this topic. For convenience, we assume without loss of generality that $\mathcal{T}_1 \geq \mathcal{T}_2 \geq \cdots \geq \mathcal{T}_k$, and set $\mathcal{T}_i = 0$ if $i > k$. Note that in the special "single-topic documents" case, where each document is relevant to only one topic, $\mathcal{T}_i^2$ is exactly the number of documents relevant to the ith topic.

Topic mingling We also define

$$\mu = \left(\sum_{t \neq t'} \left(\sum_{d=1}^{n} \mathrm{rel}(t,d)\mathrm{rel}(t',d) \right)^2 \right)^{1/2}$$

to be the *topic mingling* to measure the degree to which documents are relevant to multiple topics. A large μ indicates that documents and topics have many-to-many relations, which may make it harder to discover the hidden topics. Note that in the "single-topic documents" case, $\mu = 0$.

Notation for related values To simplify our presentation, for $x_1 \geq \cdots \geq x_n \geq 0$ and $y_1 \geq \cdots \geq y_n \geq 0$, we write

$$\begin{array}{ll} x_i \overset{opt}{=} y_i & \text{if } \max_i |x_i^2 - y_i^2| \leq E^{max}_{\mathcal{X}_{opt}} \quad \text{and} \quad (\sum_{i=1}^n (x_i^2 - y_i^2)^2/n)^{1/2} \leq E^{avg}_{\mathcal{X}_{opt}}, \\ x_i \overset{\mu}{=} y_i & \text{if } (\sum_{i=1}^n |x_i^2 - y_i^2|^2)^{1/2} \leq \mu . \end{array}$$

The relation $x_i \overset{opt}{=} y_i$ (or $x_i \overset{\mu}{=} y_i$) indicates that x_i approximates y_i (and vice versa), and that the approximation becomes closer as the optimum error (or topic mingling) becomes smaller, respectively.

Our first result relates the singular values of the optimum projection of the initial term-document matrix $\widetilde{\mathbf{A}}$ to the dominances of the underlying topics.

Theorem 3.3.1 *There exist* $\dot{\mathcal{T}}_1 \geq \ldots \geq \dot{\mathcal{T}}_n \geq 0$ *such that* $\sigma_i[\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})] \overset{opt}{=} \dot{\mathcal{T}}_i$ *and* $\dot{\mathcal{T}}_i \overset{\mu}{=} \mathcal{T}_i$.
In particular, if each document is relevant to exactly one topic, then $\mu = 0$,
so $\sigma_i[\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})] \overset{opt}{=} \mathcal{T}_i$.

Thus, the singular values of the optimum projection $\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})$ in some sense reveal the topic dominances. The extent to which this holds depends on the optimum error and on the topic mingling. If the optimum error is high, then we cannot expect the optimum subspace to reveal the topic dominances; also, if there is high topic mingling in the collection, then topics will be fairly difficult to distinguish.

Our next result relates the largest singular value of $\mathbf{P}_{\mathcal{X}_{opt}^{\perp}}(\widetilde{\mathbf{A}})$ to the root original error $\varepsilon^{max}_{\mathrm{VSM}}$.

Theorem 3.3.2 *We have* $\sigma_i[\mathbf{P}_{\mathcal{X}_{opt}^{\perp}}(\widetilde{\mathbf{A}})] \stackrel{opt}{=} \sqrt{\sigma_i[\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\mathrm{VSM}})]}$.
In particular, $\sqrt{\sigma_1[\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\mathrm{VSM}})]} = \varepsilon_{\mathrm{VSM}}^{max}$.

As $\mathcal{X}_{opt}^{\perp}$ is the orthogonal complement of the optimum subspace, $\mathbf{P}_{\mathcal{X}_{opt}^{\perp}}(\widetilde{\mathbf{A}})$ is 'noise' in the sense that it should be factored out for the similarities to be close to the true topic-based similarities. Its largest singular value is related to the root original error $\varepsilon_{\mathrm{VSM}}^{max}$.

We are now in a position to show when the optimum subspace $\mathcal{X}_{opt}$ can be closely approximated by LSI.

Theorem 3.3.3 *Let* $h = \dim(\mathcal{X}_{opt})$. *To facilitate presentation let* $\hat{\mathcal{T}}_{max} = \sigma_1[\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})]$, $\hat{\mathcal{T}}_{min} = \sigma_h[\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})]$, *and* $\hat{\varepsilon}_{\mathrm{VSM}}^{max} = \sigma_1[\mathbf{P}_{\mathcal{X}_{opt}^{\perp}}(\widetilde{\mathbf{A}})]$. *(Recall that* $\mathcal{T}_i$ *and* $\varepsilon_{\mathrm{VSM}}^{max}$ *are approximated by* $\sigma_i[\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})]$ *and* $\sigma_1[\mathbf{P}_{\mathcal{X}_{opt}^{\perp}}(\widetilde{\mathbf{A}})]$, *respectively.) Let* $\mathcal{X}_{LSI}$ *be the* h*-dimensional LSI subspace associated with the first* h *singular values of* $\widetilde{\mathbf{A}}$.
If $\hat{\mathcal{T}}_{min} > \hat{\varepsilon}_{\mathrm{VSM}}^{max}$, *then we have*

$$\tan(\mathcal{X}_{\mathrm{LSI}}, \mathcal{X}_{\mathrm{opt}}) \;\leq\; \frac{\hat{\mathcal{T}}_{max}}{\hat{\mathcal{T}}_{min}} \cdot \frac{\hat{\varepsilon}_{\mathrm{VSM}}^{max}/\hat{\mathcal{T}}_{min}}{1 - (\hat{\varepsilon}_{\mathrm{VSM}}^{max}/\hat{\mathcal{T}}_{min})^2}$$

where the tangent function tan *measures the distance between subspaces (see Appendix A.2 for the definition).*

That is, the LSI subspace $\mathcal{X}_{LSI}$ *must be close to* $\mathcal{X}_{opt}$ *when the topic-document distribution is relatively uniform* and the original error is small in comparison to the hth largest topic dominance. Conversely, for a fixed original error (satisfying the conditions), we see that *our bound on LSI's performance weakens when the underlying topic-document distribution is highly non-uniform.*

We observe that Theorem 3.3.3 reveals the essential nature of LSI. The maximum original error $(\varepsilon_{\mathrm{VSM}}^{max})^2$ measures the most prominent error in the original term-document matrix. The error is prominent, for instance, the error is mostly due to one document, or when several documents show essentially similar error. In fact, the essence of LSI lies in that the SVD factors what is prominent[4]. As LSI leaves out all but the most prominent factors of the term-document matrix, the performance of LSI depends on *what is more prominent in the given term-document matrix.* Now note that among the topics, the smaller topics potentially have the largest risk to be 'swamped' by other more dominant topics and/or prominent error. Theorem 3.3.3 precisely shows that the quality of LSI subspace is bounded by such a 'risk' as measured by $\hat{\varepsilon}_{\mathrm{VSM}}^{max}/\hat{\mathcal{T}}_{min}$ and $\hat{\mathcal{T}}_{max}/\hat{\mathcal{T}}_{min}$.

The following theorem relates the lower bound of the tangent to the non-uniformity and the original error.

[4]SVD factors a matrix into projections onto one-dimensional subspaces, so that the ith subspace has the ith largest projection under the constraints of mutual orthogonality. The projection can be large when there are many and/or long column (or row) vectors pointing in similar directions, and that is what we mean by 'prominent'. See Section 2.2 for more precise statements.

Theorem 3.3.4 *In the notation of Theorem 3.3.3, suppose that* $\hat{\mathcal{T}}_{min} > \hat{\varepsilon}^{max}_{\mathrm{VSM}} > 0$ *and* $\tan(\mathcal{X}_{\mathrm{LSI}}, \mathcal{X}_{\mathrm{opt}}) \neq 0$.
Then, we can construct an ω *satisfying*

$$\begin{aligned} \tan(\mathcal{X}_{\mathrm{LSI}}, \mathcal{X}_{\mathrm{opt}}) &\geq \frac{2}{\omega + \sqrt{\omega^2 + 4}}, \\ \omega &\geq \frac{\hat{\mathcal{T}}_{min}}{\hat{\mathcal{T}}_{max}} \cdot \frac{1 - (\hat{\varepsilon}^{max}_{\mathrm{VSM}}/\hat{\mathcal{T}}_{min})^2}{\hat{\varepsilon}^{max}_{\mathrm{VSM}}/\hat{\mathcal{T}}_{min}}. \end{aligned}$$

Observe that ω *is bounded by the reciprocal of the upper bound of the tangent in Theorem 3.3.3.*

The lower bounds of the tangent are raised by smaller values of ω, and the lower bound of ω is smaller when the distribution is less uniform (i.e. $\hat{\mathcal{T}}_{min}/\hat{\mathcal{T}}_{max}$ is small) and when the original error is larger (i.e. $\hat{\varepsilon}^{max}_{\mathrm{VSM}}/\hat{\mathcal{T}}_{min}$ is large). Note that this does not imply that LSI *must* do poorly on high non-uniformity since ω may be large even if the distribution is highly non-uniform. However, ω can be small on high non-uniformity, and LSI can not perform well when ω is small.

Theorem 3.3.3 and 3.3.4 have shown bounds on how close the LSI subspace and the optimum subspace can be. Finally, the following theorem relates the tangent to the error of the LSI subspace.

Theorem 3.3.5 *In the setting of Theorem 3.3.3, if* $\hat{\mathcal{T}}_{min} > \hat{\varepsilon}^{max}_{\mathrm{VSM}}$, *then there exist non-negative* α_1 *and* α_2, *which are uniquely determined by* $\mathcal{X}_{opt}$ *and* $\widetilde{\mathbf{A}}$, *independently from* $\mathcal{X}_{LSI}$, *satisfying* $E^{avg}_{\mathcal{X}_{LSI}} \leq E^{avg}_{\mathcal{X}_{opt}} + \alpha_1 \tan(\mathcal{X}_{\mathrm{LSI}}, \mathcal{X}_{\mathrm{opt}}) + \alpha_2 \left(\tan(\mathcal{X}_{\mathrm{LSI}}, \mathcal{X}_{\mathrm{opt}})\right)^2$.

The tangent between $\mathcal{X}_{LSI}$ and $\mathcal{X}_{opt}$ is, indeed, coupled with the degree to which the error of LSI subspace becomes close to the optimum error (i.e. smaller): small tangent values reduce the upper bound of $E^{avg}_{\mathcal{X}_{LSI}}$.

3.4 The dimensionality of the LSI subspace

Based on our topic-based model, we analyze the relationship between the dimensionality and error of LSI subspaces.

Theorem 3.4.1 *Let* $\mathcal{X}_{LSI}$ *be an* x*-dimensional LSI subspace associated with the first* x *non-zero singular values of* $\widetilde{\mathbf{A}}$. *Then, for some* $\dot{\mathcal{T}}_1 \geq \cdots \geq \dot{\mathcal{T}}_n$ *satisfying*

$$\dot{\mathcal{T}}_i \stackrel{\mu}{=} \mathcal{T}_i,$$

and for some $E^{(1)}_{\mathcal{X}_{\mathrm{VSM}}}, \ldots, E^{(n)}_{\mathcal{X}_{\mathrm{VSM}}}$ *bounded by*

$$E^{(i)}_{\mathcal{X}_{\mathrm{VSM}}} \leq E^{max}_{\mathcal{X}_{\mathrm{VSM}}}, \qquad \sqrt{\sum_{i=1}^{n} (E^{(i)}_{\mathcal{X}_{\mathrm{VSM}}})^2/n} \leq E^{avg}_{\mathcal{X}_{\mathrm{VSM}}},$$

we have

$$E^{avg}_{\mathcal{X}_{LSI}} \geq \sqrt{\frac{\sum_{i=1}^{x}\left(E^{(i)}_{\mathcal{X}_{\mathrm{VSM}}}\right)^2 + \sum_{i=x+1}^{k}\dot{\mathcal{T}}_i^4}{n}} \quad \textit{for } x < k\,, \tag{3.1}$$

$$E^{avg}_{\mathcal{X}_{LSI}} \geq \sqrt{\frac{\sum_{i=1}^{k}\left(E^{(i)}_{\mathcal{X}_{\mathrm{VSM}}}\right)^2 + \sum_{i=k+1}^{x}\sigma_i[\widetilde{\mathbf{A}}]^4}{n}} \quad \textit{for } x \geq k\,, \tag{3.2}$$

$$|E^{avg}_{\mathcal{X}_{LSI}} - E^{avg}_{\mathcal{X}_{\mathrm{VSM}}}| \leq \sqrt{\frac{\sum_{i=x+1}^{n}\sigma_i[\widetilde{\mathbf{A}}]^4}{n}} \quad \textit{for any } x\,. \tag{3.3}$$

The quantity $\dot{\mathcal{T}}_i$ is associated with the ith largest topic dominance, and $E^{(i)}_{\mathcal{X}_{\mathrm{VSM}}}$ is bounded from above by the maximum and average original errors. Recall that k is the number of topics. The singular value $\sigma_i[\widetilde{\mathbf{A}}]$ measures the size of the portions of document vectors either brought in the LSI subspace (for $i \leq x$), or left out (for $i > x$) (see Section 2.2).

Suppose that the original error is not so large compared with the topic dominances. Then, the error bound in inequality (3.1) is dominated by $\sum_{i=x+1}^{k}\dot{\mathcal{T}}_i^4$, which measures the dominances of topics that might be left out from the x-dimensional LSI subspace. The bound becomes smaller (i.e. better) as x goes up to k, reducing the risk of leaving out topics; see the line labeled 'eq.(3.1)' in Figure 3.3.

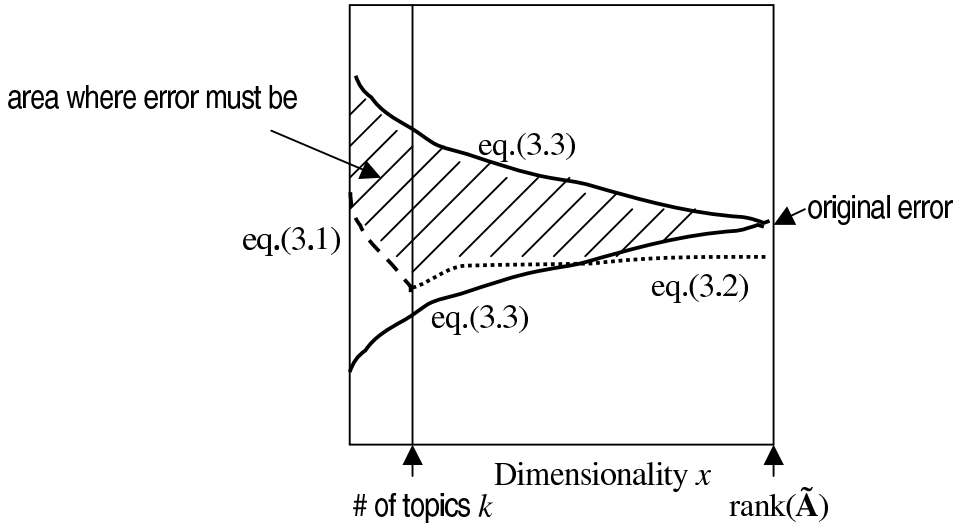

Figure 3.3: Schematic illustration of LSI error bounds and dimensionality.

In the case where the dimensionality is greater than the number of topics, the best-case error bound monotonically increases on the dimensionality; see (3.2). Note that the first term $\sum_{i=1}^{k}\left(E^{(i)}_{\mathcal{X}_{\mathrm{VSM}}}\right)^2$ is independent of x. The second term $\sum_{i=k+1}^{x}\sigma_i[\widetilde{\mathbf{A}}]^4$ corresponds to the magnitude of noise (i.e. the portions of vectors that raise the error) possibly brought into $\mathcal{X}_{LSI}$. This

best-case error bound is greater (i.e. worse) for larger x, increasing the risk of bringing noise in; see the line labeled 'eq.(3.2)' in Figure 3.3.

Thus, if the original error is relatively small compared with topic dominances, *the best-case error bound is the lowest when the dimensionality is the number of topics k.* Note that this does not imply that LSI *must* do well when the dimensionality is the number of topics since the actual error can be larger than the best-case error bound.

Inequality (3.3) shows that $E^{avg}_{\mathcal{X}_{LSI}}$ can be different from the average original error by at most the amount of the portions left out from $\mathcal{X}_{LSI}$ measured by $\sqrt{\sum_{i=x+1}^{n} \sigma_i[\widetilde{\mathbf{A}}]^4/n}$. When the dimensionality reaches rank($\widetilde{\mathbf{A}}$), the error is exactly the original error, as the original term-document matrix is reproduced. This indicates that *lower-dimensional LSI subspaces have more chance to differ (for better or for worse) from* VSM*, independently from the number of topics k* (which is an intuitive result); see the lines labeled 'eq.(3.3)' in Figure 3.3.

The actual error $E^{avg}_{\mathcal{X}_{LSI}}$ resides somewhere between the worst-case bound indicated by (3.3) and the worst over the best-case bounds indicated by (3.1)-(3.3), which is between bold lines in the example in Figure 3.3.

As shown above, the best dimensionality is weakly related to the number of topics k and the proportion of the term-document matrix left out of the LSI subspace. In particular, we see that *in case that LSI largely outperforms VSM, the dimensionality of the LSI subspace should be around the number of topics k.*

3.5 Related Work: Theoretical Analyses of LSI

As noted above, there have been several studies formally analyzing LSI. Overall, the previous studies start from the belief that LSI *is* effective and focus on explaining the effectiveness of LSI. In contrast, our analysis seeks to precisely quantify the factors that affect LSI's performance either positively or negatively.

We discuss the analyses that rely on one of the invariant subspace perturbation theorems (as our analysis does) in Section 3.5.1. Studies of finding the best dimensionality are reviewed in Section 3.5.2, and finally Section 3.5.3 discusses other types of analyses.

3.5.1 Generative models

Papadimitriou et al. (2000) and Azar et al. (2001) propose to explain LSI's effectiveness based on generative probabilistic models. Their main idea is that when the error portions of vectors are small and random, it is unlikely for them to get into the LSI subspace. Although their proofs rely on invariant subspace perturbation theorems, they derived different type of results. We summarize their results and discuss what caused the differences from our results.

Let $\mathbf{x}_i$ be a new document vector in the LSI subspace. Papadimitriou et al. (2000) have shown that, under certain conditions (shown below; which are on the corpus model, the term-document

matrix, and a sufficiently small constant ϵ),

$$\begin{array}{ll} \mathbf{x}_i^T \mathbf{x}_j \leq \delta ||\mathbf{x}_i||_2 ||\mathbf{x}_j||_2 & \text{for any cross-topic document pair,} \\ \mathbf{x}_i^T \mathbf{x}_j \geq 1 - \delta ||\mathbf{x}_i||_2 ||\mathbf{x}_j||_2 & \text{for any same-topic document pair,} \\ \text{where } \delta \in O(\epsilon)\,, & \end{array}$$

with probability $1 - O(m^{-1})$.

The conditions on their generative corpus model are as follows. Each document is on a single-topic. Each topic is associated with one of a set of mutually disjoint term sets. The probability that a topic assigns to the terms in the associated set is at most τ each, and at least $1 - \epsilon$ in total, where τ is a sufficiently small constant. This condition naturally bounds ϵ by $0 \leq \epsilon < 1$, as the probability can not be negative.

Let $\mathbf{A}$ be a *pure* term-document matrix, that is, generated by the corpus model, satisfying the conditions above. Let $\widetilde{\mathbf{A}}$ be the given term-document matrix. There are two conditions on $\mathbf{A}$ and $\widetilde{\mathbf{A}}$:

$$||\widetilde{\mathbf{A}} - \mathbf{A}||_2 \leq \epsilon < 1\,, \tag{3.4}$$

$$\sigma_k[\mathbf{A}]/\sigma_{k+1}[\mathbf{A}] > c\, \sigma_1[\mathbf{A}]/\sigma_k[\mathbf{A}]\,, \tag{3.5}$$

where c is a sufficiently large constant, and k is the dimensionality of the LSI subspace.

Compared with our analysis, the inequality (3.5), which is to assume a big gap between $\sigma_k[\mathbf{A}]$ and $\sigma_{k+1}[\mathbf{A}]$, may, from our perspective, be roughly related to the uniformity. However, in their conclusion, ϵ, which roughly corresponds to our root original error $\varepsilon_{\mathrm{VSM}}^{max}$, is the only factor to bound the quality of the LSI subspace. Such difference in results derives from the fact that Papadimitriou et al. (2000) regard most of the quantities as constant values[5] while they are variables for us. Thus, our notion of uniformity is disregarded as a constant part of $O(\epsilon)$ in their conclusion.

Azar et al. (2001) analyze LSI by comparing LSI applied to a noisy term-document matrix against LSI applied to a 'pure' term-document matrix.

Let $\widetilde{\mathbf{A}} \in \Re^{m \times n}$ be the term-document matrix, from which the LSI subspace $\mathcal{X}_{LSI}$ is derived. Let $\mathbf{A}$ be a pure term-document matrix, from which the *pure-LSI* [6] subspace $\mathcal{X}_{\mathrm{pure^L}}$ is derived. Azar et al. (2001) showed the following. Suppose that the noise matrix $\mathbf{E} = \widetilde{\mathbf{A}} - \mathbf{A}$ is a random matrix with mean zero and constant deviation. Let h be the dimensionality of $\mathcal{X}_{LSI}$ and $\mathcal{X}_{\mathrm{pure^L}}$. Assume that

$$\sigma_h[\mathbf{A}] - \sigma_{h+1}[\mathbf{A}] \in \omega(\sqrt{m+n})\,, \tag{3.6}$$

i.e., these singular values are sufficiently separated compared with the size of the document collection. Let $\widetilde{\mathbf{a}}_i^{\mathrm{L}}$ and $\mathbf{e}_i^{\mathrm{L}}$ be the projection of $\widetilde{\mathbf{a}}_i$ and $\mathbf{e}_i$ (noise portion) onto $\mathcal{X}_{LSI}$, respectively. Let $\mathbf{a}_i^{\mathrm{pure^L}}$ denote the projection of the pure document vector $\mathbf{a}_i$ onto $\mathcal{X}_{\mathrm{pure^L}}$. Define *good* documents

[5] For instance, in their Lemma 4 and its proof, singular values are bounded by specific values such as 1/20.

[6] 'Pure-LSI' is our naming for the sake of clarity.

to be those satisfying

$$||\mathbf{e}_i^{\mathrm{L}}||_2 \in o(||\widetilde{\mathbf{a}}_i^{\mathrm{L}}||_2)\,. \tag{3.7}$$

$$||\mathbf{a}_i^{\mathrm{pure^L}}||_2 \in \theta(||\mathbf{a}_i||_2)\,, \tag{3.8}$$

That is, the good documents are the ones whose error portions are mostly left out by LSI, and whose pure portions are mostly captured in the pure-LSI subspace $\mathcal{X}_{\mathrm{pure^L}}$. Then, it holds for *good* documents that

$$|\angle(\mathbf{a}_i^{\mathrm{pure^L}}, \mathbf{a}_j^{\mathrm{pure^L}}) - \angle(\widetilde{\mathbf{a}}_i^{\mathrm{L}}, \widetilde{\mathbf{a}}_j^{\mathrm{L}})| \in o(1)\,,$$

i.e., for good documents, the performance of LSI on the noisy term-document matrix is as good as LSI on the pure matrix.

Unlike Papadimitriou et al., Azar et al. do not assume single-topic documents. However, we note that Azar et al.'s assumptions are rather restrictive, which, from our perspective, results in *precluding the non-uniform case from their analysis* as follows. Let k be the number of topics, and suppose that the rank of the pure matrix is k, as Azar et al. suggest. Then, we observe that the ith singular value of the pure matrix $\sigma_i[\mathbf{A}]$ is roughly related to our notion of the ith largest topic dominance. Suppose that $h = k$. When the topic-document distribution is less uniform, $\sigma_h[\mathbf{A}]$ becomes relatively small, which makes it hard to satisfy the condition (3.6). Now suppose that $h < k$ so that a less uniform collection can satisfy (3.6). Then, it becomes harder for the documents on less dominant topics to be well-represented in the pure-LSI subspace because of the essential nature of LSI[7], i.e., the documents on less dominant topics tend to violate the condition (3.8) and fail to be good documents. Inequality (3.6) does not hold for $h > k$ since $\sigma_i[\mathbf{A}] = 0$ for $i > k$. All together, we see that, from our perspective, Azar et al's assumptions result in precluding the case of non-uniform document collections and/or the documents relevant to less dominant topics from their analysis.

We note that our approach is somewhat orthogonal to these two analyses: rather than starting from a *particular* set of conditions and showing that LSI provides good results under those conditions, we seek to specify exactly *what conditions* on *what quantities* are necessary for LSI to perform well. A strength of our approach is that, by carefully studying these quantities, we are able to not only explain LSI's behavior, but develop new methods for improving on it, as outlined in the next chapter.

3.5.2 Dimensionality selection

Zha et al. (1998) and Ding (1999) address the issue of dimensionality selection, in relation to their subspace-based model and probabilistic model, respectively.

Zha et al. (1998) propose a subspace-based model of LSI. Their work focuses on dimensionality selection based on *minimal description length* (MDL) and more accurate SVD-updating schemes.

[7] See the paragraph after our Theorem 3.3.3 for the argument of LSI's essential nature.

In their model, the ith term vector $\mathbf{t}_i \in \Re^n$ is

$$\mathbf{t}_i = \mathbf{C}\mathbf{w}_i + \epsilon_i \,,$$

where the columns of $\mathbf{C} \in \Re^{n \times h}$ represent the h underlying concepts, $\mathbf{w}_i \in \Re^h$ is the concept-weight vector assigned to the ith term, and ϵ_i is noise. The *latent-concept subspace* is range($\mathbf{C}$) of dimensionality h. Let $\mathcal{E}\{\cdot\}$ denote the expectation operator. Assuming no correlation among the noise components and between $\mathbf{w}_i$ and ϵ_i, they obtain

$$\mathbf{T} = \mathcal{E}\{\mathbf{t}_i \mathbf{t}_i^T\} = \mathbf{C}\mathcal{E}\{\mathbf{w}_i \mathbf{w}_i^T\}\mathbf{C}^T + \sigma^2 \mathbf{I}_n \,.$$

Then, the eigenvalues of $\mathbf{T}$ are in the form of

$$\mu_1 + \sigma^2, \ldots, \mu_h + \sigma^2, \sigma^2, \ldots, \sigma^2 \,,$$

i.e., the smallest $n - h$ eigenvalues are all equal to σ^2. Estimating $\mathbf{T}$ by $\mathbf{T} \approx \widetilde{\mathbf{A}}^T \widetilde{\mathbf{A}}/m$, ideally the dimensionality h should be determined by counting how many smallest eigenvalues are equal. However, in practice, none of the eigenvalues of $\widetilde{\mathbf{A}}^T \widetilde{\mathbf{A}}$ are equal to each other, and Zha et al. (1998) use MDL to determine h as is done in the context of array signal processing.

We note that the notion of noise (or error) is quite different between our approach and that of Zha et al. (1998). In our model, the error is defined with respect to a (hypothetical) human-assigned relevance between documents and topics, as our goal is the output sensible to humans. In Zha et al. (1998)'s approach, the noise is measured with respect to the encoding efficiency, where human judgment has no role to play.

Ding (1999) proposes a similarity-based probabilistic model for LSI. The assumption is that, given the basis vectors of a subspace $\mathbf{B}_k = \{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$, the documents are distributed with probability $p(\mathbf{d}|\mathbf{B}_k) = \exp(\sum_{j=1}^k (\mathbf{d}^T \mathbf{b}_j)^2)/Z(\mathbf{B}_k)$ where $\mathbf{d}$ denotes a document vector and $Z(\mathbf{B}_k) = \int \exp(\sum_{j=1}^k (\mathbf{x}^T \mathbf{b}_j)^2) d\mathbf{x}$ for normalization. This follows a Gaussian distribution when we regard $\mathbf{B}_k$ as the mean and an inner product as a measure of similarity. Assuming independence, the log-likelihood for the document vectors projected onto the subspace spanned by $\mathbf{B}_k$ is computed as

$$\begin{aligned} l_k &= \log(\prod_{i=1}^n p(\mathbf{d}_i|\mathbf{B}_k)) \\ &= \sum_{i=1}^n \sum_{j=1}^k (\mathbf{d}_i^T \mathbf{b}_j)^2 + (-n)\log(Z(\mathbf{B}_k)) \end{aligned}$$

The dimensionality of an LSI subspace is selected by seeking for $\arg\max_k l_k$.

Ding has argued that larger log-likelihood indicates a better statistical model, (i.e., a better semantic space). He argues for LSI that LSI maximizes the first term of l_k while the second term is negligible because it changes very slowly compared to the first term.

We note that l_k's second term ($(-n)\log(Z(\mathbf{B}_k))$), ignored by Ding, is somewhat related to the

uniformity of the topic-document distribution. Ding estimated the second term as

$$(-n)\log(\sum_{i=1}^{n}\exp(\sum_{j=1}^{k}(\mathbf{d}_i^T\mathbf{b}_j)^2))$$

by taking the documents as unbiased data drawn from the population, and by ignoring $d\mathbf{x}$, treating it as being independent of k. Now let $\hat{\mathbf{d}}_i$ be the ith document vector in the k-dimensional LSI subspace. Then, we observe that

$$||\hat{\mathbf{d}}_i||_2^2 = \sum_{j=1}^{k}(\mathbf{d}_i^T\mathbf{b}_j)^2\,.$$

Rewriting the estimated second term as $(-n)\log(\sum_{i=1}^{n}\exp(||\hat{\mathbf{d}}_i||_2^2))$, we note that, for $\sum_{i=1}^{n}||\hat{\mathbf{d}}_i||_2^2$ fixed, the estimation of the second term is maximized when

$$||\hat{\mathbf{d}}_1||_2^2 = \cdots = ||\hat{\mathbf{d}}_n||_2^2\,.$$

In other words, the second term is maximized when the lengths of the resultant document vectors are all the same. Now note that, according to the mathematical formulation of SVD, the documents on the less dominant topics tend to result in shorter document vectors in the LSI subspace. Hence, we see that *Ding's likelihood estimation for the LSI subspace is likely to be greater when the documents are distributed over topics more uniformly.*

In our previous work (Ando, 2000), we proposed to select the dimensionality h of the IRR subspace by

$$h = \arg\max_{x=1,\ldots,n}\hat{l_x} - (\sum_{i=x-c*n/2}^{x-1}\hat{l_i} + \sum_{i=x+1}^{i=x+c*n/2}\hat{l_i})/(c*n)$$

with constant $0 < c < 1$, so that h maximizes the log-likelihood with respect to the average over the nearby dimensionalities. We will discuss this method in Section 4.1.2.

Unlike us, neither Zha et al.'s model nor Ding's model incorporates human relevance judgment directly. Instead, Zha et al. (1998) essentially assume the optimal efficiency of natural language as an encoding (therefore, making MDL appropriate), and Ding assumes a Gaussian distribution on documents. We observe that such clear differences in approaches have resulted in our different type of results.

3.5.3 Other analyses of LSI

Bartell et al. (1992) have shown that LSI can be regarded as a solution of a special Multidimensional Scaling (MDS) problem. MDS finds the mapping from the given vectors to new vectors optimal with respect to the given constraints on the similarity (closeness) between vectors. Bartell et al. (1992) show that LSI provides the optimal solution to the special case of MDS such that the given similarity constraints are exactly the ones produced by the given term-document matrix. This study gives an insight into LSI, instead of investigating the conditions under which LSI is effective.

Interestingly to us, Bartell et al. (1992) show a locally-optimal solution to the problem of finding a matrix $\mathbf{W}$ that minimizes the F-norm of $\mathbf{C}^T\mathbf{C} - (\mathbf{WX})^T(\mathbf{WX})$ for a given general matrix $\mathbf{C}$ and $\mathbf{X}$. We note that this quantity corresponds to our average error $E_{\mathcal{X}}^{avg}$ for $\mathcal{X} = \text{range}(\mathbf{W}^T)$ when we set $\mathbf{C}_{[i,j]} = \text{rel}(i,j)$ and $\mathbf{X} = \widetilde{\mathbf{A}}$ and constrain the rows of $\mathbf{W}$ to be orthonormal. However, note that this solution does not solve our document representation problem in which true similarities are *hidden* from a method. Bartell et al. (1992) did not relate their solution to LSI but rather developed a new method *Metric Similarity Modeling* (MSM) (Bartell et al., 1995), which we will discuss in Section 8.1.

Story (1996) relates Bayesian regression models to LSI. While our model and other studies we have discussed focus on general inter-document similarity, Story's focus is specifically on the relevancy measure between a query and documents. From a Bayesian perspective, he expects that dropping smallest singular values removes "statistically dubious information" and reduces "specification errors"; however, formal definitions of these quantities are not provided. Although Story briefly discusses possible variations of LSI, the investigation of their effectiveness is beyond the scope of his study.

These two studies give insight into another aspect of LSI and vector-based methods in general.

Chapter 4

IRR: Overcoming non-uniform topic-document distributions

Our results from Chapter 3 indicate that we could improve the performance of LSI if we could somehow "smooth" the topic-document distribution (that is, effectively lower $\hat{\mathcal{T}}_{max}/\hat{\mathcal{T}}_{min}$). We propose an alternative method called *Iterative Residual Rescaling algorithm* (IRR), which accomplishes this task *without prior knowledge of the assignments of topics to documents.*

4.1 IRR algorithm

Recall from Section 2.2.4 that the left singular vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots$ form a recursive sequence as follows:

$$\begin{aligned} \mathbf{u}_i &= \arg\max_{\|\mathbf{x}\|_2=1} \sum_{j=1}^{n} \left(\|\mathbf{r}_j^{(i)}\|_2 \cos(\mathbf{x}, \mathbf{r}_j^{(i)}) \right)^2 , \\ \mathbf{R}^{(1)} &= \widetilde{\mathbf{A}} , \\ \mathbf{R}^{(i+1)} &= \mathbf{R}^{(i)} - \mathrm{proj}(\mathbf{R}^{(i)}, \{\mathbf{u}_i\}) . \end{aligned}$$

Note that the jth *residual* $\mathbf{r}_j^{(i)}$ is associated with the jth document as it derives from the jth column of a term-document matrix $\widetilde{\mathbf{A}}$. We observe that a residual vector's length is inversely related to the corresponding document's relevancy to dominant topics, as long as the initial term-document matrix somewhat reflects the hidden topic-document correlations. For example, suppose $\mathbf{u}_1$ points in the predominant direction of the initial document vectors, as shown in Figure 4.1(a). Any vector for a document relevant to the dominant topic is likely to lie in a direction similar to that of $\mathbf{u}_1$; hence, subtracting off its projection onto $\mathbf{u}_1$ shrinks it to a large degree (Figure 4.1(b)).

As the residuals for the dominant topic become smaller, one would expect that the next left singular vector should represent larger residuals on the less dominant topics. However, when the topic-document distribution is highly non-uniform, the cumulative influence of a large number of

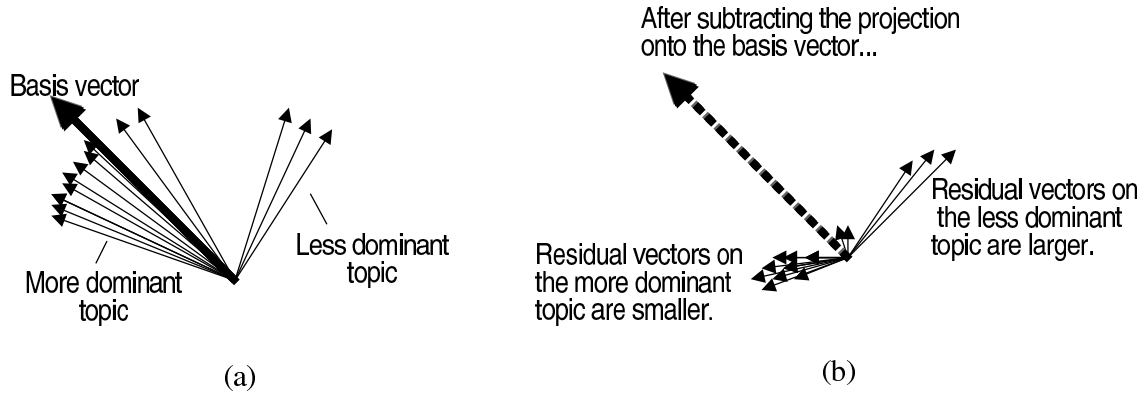

Figure 4.1: Lengths of residuals inversely reflect topic dominances.

small residuals for dominant topics can cause smaller topics to be ignored, as depicted in 4.2(a) and (b).

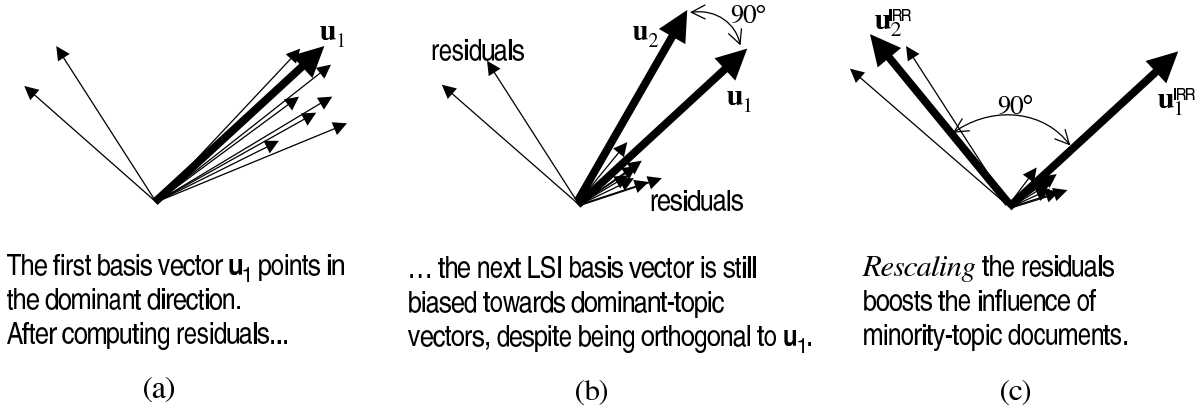

Figure 4.2: Effect of non-uniformity on LSI, and how IRR compensates.

IRR amplifies vectors for less dominant topics by sharpening the length differences among residual vectors, thus *compensating for non-uniform topic distributions* (Figure 4.2(c)). More precisely, IRR produces *IRR-basis vectors* $\mathbf{u}_1^{\text{IRR}}, \mathbf{u}_2^{\text{IRR}}, \ldots$, (instead of the left singular vectors), which are recursively defined as follows:

$$
\begin{aligned}
\text{pow}(\mathbf{r}, q) &= ||\mathbf{r}||_2^q \, \mathbf{r}\,, \\
\mathbf{u}_i^{\text{IRR}} &= \arg\max_{||\mathbf{x}||_2=1} \sum_{j=1}^{n} \left(||\text{pow}(\mathbf{r}_j^{(i)}, q)||_2 \, \cos(\mathbf{x}, \text{pow}(\mathbf{r}_j^{(i)}, q)) \right)^2 , \\
\mathbf{R}^{(1)} &= \widetilde{\mathbf{A}}\,, \\
\mathbf{R}^{(i+1)} &= \mathbf{R}^{(i)} - \text{proj}(\mathbf{R}^{(i)}, \{\mathbf{u}_i^{\text{IRR}}\})\,.
\end{aligned}
$$

Crucially, IRR's objective function incorporates a *scaling factor* q and scaling function $\text{pow}(\mathbf{r}, q)$. The maximization can be done by computing the first left singular vector of the rescaled residual

matrix:

$$\widehat{\mathbf{R}}^{(i)} = \left[\, \mathrm{pow}(\mathbf{r}_1^{(i)}, q) \quad \cdots \quad \mathrm{pow}(\mathbf{r}_n^{(i)}, q) \,\right] .$$

That is, IRR *rescales* each residual vector $\mathbf{r}_j^{(i)}$ at each IRR-basis vector computation, increasing the contrast between long and short residuals by setting $q > 0$. LSI is the special case in which $q = 0$.

The similarities between documents are measured in the *IRR subspace* spanned by $\mathbf{u}_1^{\mathrm{IRR}}, \mathbf{u}_2^{\mathrm{IRR}}, \ldots$.

The pseudocode for IRR in Figure 4.3 is high-level. An efficient implementation is discussed later in Section 5.1. An example data flow of the IRR computation may be found in Figure 7.6.

IRR(q,ℓ):
  $\mathbf{R} := \widetilde{\mathbf{A}}$ // initialize residuals by given term-doc matrix
  For $i := 1, 2, \ldots, \ell$ // create ℓ-dimensional IRR subspace
    For $j := 1, 2, \ldots, n$
      $\widehat{\mathbf{r}}_j := \mathrm{pow}(\mathbf{r}_j, q)$ // rescale residuals
    $\mathbf{u}_i^{\mathrm{IRR}} := \arg\max_{||\mathbf{x}||_2=1} \left(\sum_{j=1}^{n} (||\widehat{\mathbf{r}}_j||_2 \cos(\widehat{\mathbf{r}}_j, \mathbf{x}))^2 \right)$
    $\mathbf{R} := \mathbf{R} - \mathrm{proj}(\mathbf{R}, \{\mathbf{u}_i^{\mathrm{IRR}}\})$ // Recompute residuals
  // new document representation in IRR subspace
  $\mathbf{A}^{\mathrm{IRR}} := \mathrm{proj}(\widetilde{\mathbf{A}}, \{\mathbf{u}_1^{\mathrm{IRR}}, \mathbf{u}_2^{\mathrm{IRR}}, \ldots, \mathbf{u}_\ell^{\mathrm{IRR}}\})$

Figure 4.3: High-level pseudocode for IRR.

4.1.1 Scaling factor selection: The AUTO-SCALE method

Our discussion above argues that *the degree of rescaling should depend on the uniformity of the topic-document distribution*; more non-uniformity needs higher degree of rescaling. Our analysis in Chapter 3 allows us to exploit this connection to develop an effective estimation method — *automatic scaling factor determination* (AUTO-SCALE) — that approximates the topic-document non-uniformity *without prior knowledge of the underlying topics*.

AUTO-SCALE works as follows. Observing that a square sum of topic dominances is the number of documents n, i.e., $\sum_{i=1}^{k} \mathcal{T}_i^2 / n = 1$, we use the quantity

$$\sum_{i=1}^{k} \left(\frac{\mathcal{T}_i^2}{n} \right)^2 ,$$

as a measure of the non-uniformity, which increases when $\mathcal{T}_i^2$ distributes over n non-uniformly. AUTO-SCALE approximates this measure by

$$f(\widetilde{\mathbf{A}}) = \left(\frac{||\widetilde{\mathbf{A}}^T \widetilde{\mathbf{A}}||_F}{n} \right)^2 .$$

This approximation follows from

$$\begin{aligned}
||\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}||_{\mathrm{F}}^2 &\approx ||\mathbf{S}||_{\mathrm{F}}^2 \\
&= \sum_{d=1}^{n}\sum_{d'=1}^{n}\left(\sum_{t=1}^{k}\mathrm{rel}(t,d)\mathrm{rel}(t,d')\right)^2 \\
&= \sum_{t=1}^{k}\left(\sum_{d=1}^{n}\mathrm{rel}(t,d)^2\right)^2 + \sum_{s\neq u}\left(\sum_{d=1}^{n}(\mathrm{rel}(s,d)\mathrm{rel}(u,d))\right)^2 \\
&= \sum_{t=1}^{k}\left(\mathcal{T}_t^2\right)^2 + \left(\mu^2\right)^2 \\
&\approx \sum_{t=1}^{k}\mathcal{T}_t^4 .
\end{aligned}$$

The assumption is that the similarities produced by $\widetilde{\mathbf{A}}$ are close to the hidden topic-based similarities (i.e., $\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}} \approx \mathbf{S}$) and documents are approximately single-topic (i.e., $\mu \approx 0$). In practice, we set q to a linear function of $f(\widetilde{\mathbf{A}})$; this is discussed in Section 4.3.1.

Although the above approximations are rather coarse, AUTO-SCALE yields good empirical results: see Sections 4.3, 4.4, 5.4, and 5.5.

4.1.2 Dimensionality selection

IRR's second parameter is ℓ, the dimensionality of the IRR subspace. One systematic way to determine the dimensionality is to train some parameter on held-out data. As the best dimensionality is weakly related to the number of topics k (see Section 3.4), directly training the dimensionality parameter is not a good idea if one expects the number of topics varies across the collections. Recall inequality (3.3) in Theorem 3.4.1:

$$|E^{avg}_{\mathcal{X}_{LSI}} - E^{avg}_{\mathcal{X}_{\mathrm{VSM}}}| \leq \sqrt{\frac{\sum_{i=x+1}^{n}\sigma_i[\widetilde{\mathbf{A}}]^4}{n}},$$

i.e., the performance improvement (or degradation) of the x-dimensional LSI subspace over VSM is bounded by the singular values of the term-document matrix. Note that

$$\sum_{i=x+1}^{n}\sigma_i[\widetilde{\mathbf{A}}]^2 = ||\widetilde{\mathbf{A}} - \mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})||_{\mathrm{F}}^2,$$

and for an ℓ-dimensional IRR subspace, we have

$$||\widetilde{\mathbf{A}} - \mathbf{P}_{\mathcal{X}_{IRR}}(\widetilde{\mathbf{A}})||_{\mathrm{F}}^2 = ||\mathbf{R}^{(\ell+1)}||_{\mathrm{F}}^2 .$$

We found that learning thresholds on the *residual ratio*, which is defined for ℓ-dimensional LSI

and IRR subspaces by

$$\frac{||\mathbf{R}^{(\ell+1)}||_{\mathrm{F}}^2}{n},$$

as a stopping criterion is effective for both LSI and IRR.

While training on held-out data is reasonable, it is a relatively expensive process, so a faster alternative is desirable. In some settings, the number of topics k is specified (e.g., because the set of topics comes from a fixed class such as the TREC topic labels, or because the application allows the user to specify the appropriate level of granularity for his or her needs), in which case we could simply set the dimensionality equal to k, as Theorem 3.4.1 indicates k's weak connection to the best dimensionality. We describe experiments with both types of selection method in Sections 4.3 and 4.4.

In our previous work (Ando, 2000) we studied a method to select the dimensionality so that the log-likelihood (based on Ding (1999)'s estimation in Section 3.5.2) is maximized with respect to the average over the nearby dimensionalities. This *log-likelihood method* achieved *pair-wise average precision* (defined in Section 4.2) comparable to the residual ratio method (described above) in our previous experiments shown in Ando (2000). However, a disadvantage of the log-likelihood method is that there is no way to reflect the input from users (e.g., the level of topic granularity). Since we seek vector spaces corresponding to human interpretable concepts, we do not pursue the log-likelihood method in this thesis.

4.2 Evaluation metrics

Kappa average precision We use *pair-wise average precision*, adapted from the average precision measure commonly used in information retrieval, as an evaluation metric, under the reasonable assumption that the measured similarity for any two *same-topic* documents (i.e., that share at least one topic) should be higher than for any two *cross-topic* documents which have no topics in common. Let p_i denote the document pair with the ith largest measured similarity (cosine). Precision for an *same-topic* pair p_i is defined by

$$\mathrm{prec}(p_i) = \frac{\#\text{ of same-topic pairs } p_j \text{ such that } j \leq i}{i}.$$

The *pair-wise average precision* is the average of these precision values over all same-topic pairs.

To compensate for the effect of large topics (which increase the likelihood of chance same-topic pairs), we modify the pair-wise average precision to create a new metric, which we call the *kappa precision* in reference to the Kappa statistic (Siegel and Castellan, 1988; Carletta, 1996):

$$\mathrm{prec}_\kappa(p_i) = \frac{\mathrm{prec}(p_i) - \mathrm{chance}}{1 - \mathrm{chance}}$$
$$\text{where} \qquad \mathrm{chance} = \frac{\#\text{ of same-topic pairs}}{\#\text{ of document pairs}}.$$

The *kappa average precision* is defined to be the average of the kappa precision over all same-topic pairs, and is a linear function of the pair-wise average precision.

	topic 1	topic 2	topic 3	topic 4
cluster 1	5	10	**20**	0
cluster 2	5	10	5	0
cluster 3	0	0	0	**21**
cluster 4	**15**	5	0	0
cluster 5	0	0	0	4

Figure 4.4: Sample contingency table, with $g(C) = (15 + 20 + 21)/100 = 56\%$.

Clustering We also test how well the new subspaces represent document similarities by seeing whether document clustering improves when these new representations are used as input. For simplicity, we consider only single-topic documents.

Let C be a cluster-topic contingency table such that $C[i, j]$ is the number of documents in cluster i that are relevant to topic j, as in Slonim and Tishby (2000). We define $g(C) = \sum_{i,j} \mathcal{N}_{ij}/n$, where

$$\mathcal{N}_{ij} = \begin{cases} C[i,j] & \text{if } C[i,j] \text{ is the } \textit{unique} \text{ maximum in both its row and column} \\ 0 & \text{otherwise} \end{cases}$$

Note that this (rather strict) measure only considers the most tightly coupled topic-cluster assignment, and decreases when either cluster purity or topic integrity falls (see Figure 4.4).

To factor out the idiosyncrasies of particular clustering algorithms in our evaluation, we apply six standard clustering methods — single-link, complete-link, group average, and k-means with initial clusters generated by these three methods — to the document vectors in each proposed subspace, and record both the *ceiling* (highest) and *floor* (lowest) $g(C)$ scores. While the ceiling performance is perhaps more intuitive, we observe that floor performance also gives us important information about a subspace's representational power: if the floor is low, then there is at least one clustering algorithm for which the document subspace is not a good representation; otherwise, the representation is good for *all six* clustering algorithms.

4.3 Controlled-distribution experiments: validation of theorems

Our first suite of experiments study the dependence of LSI and IRR on increasingly less uniform topic-document distributions. The results strongly support our theoretical analysis of LSI's sensitivity to non-uniformity.

4.3.1 Experimental setting

To exclude factors irrelevant to our focus (in this section) on the uniformity of topic-document distributions, we first generated two-topic data from the TREC collection: 70 document sets of exactly the same size and over the same two TREC topics. These sets had the following seven (increasingly non-uniform) distribution types, with ten document sets for each: (25, 25), (30, 20), (35, 15), (40, 10), (43, 7), (45, 5), and (46, 4), where (i, j) denotes the number of

documents relevant to the first and second topic, respectively. To create the term-document matrices, we extracted single-word stemmed terms using TALENT (Boguraev and Neff, 2000), removed stop-words, and then length-normalized the document vectors (so that term weights were frequency-based).

We also created three-topic data sets in the same manner, where the distribution types were of the form $(i, j, j), i+2j = 50$ (using these restricted types makes uniformity comparison obvious), and similarly for four-topic and five-topic data sets.

To implement AUTO-SCALE, we set $q = \alpha \cdot f(\widetilde{\mathbf{A}}) + \beta$, where $\alpha = 3.5$ and $\beta = 0$ for *all* our experiments. These values (which are necessary to determine the "units" of the scale factor) were empirically determined once and for all from observations on data disjoint from our test sets. This contrasts with training q for every new test set encountered. Training is an expensive process, and we envision interactive applications such as organizing query results (a task we simulate in Section 4.4) in which what would serve as training data is not obvious. (Note that the degree of non-uniformity of training data should be similar to that of test data; for we predict higher non-uniformity requires larger scaling factor, and this prediction is empirically confirmed in the experiments in Section 4.3.3.) We thus view AUTO-SCALE as a practical alternative to the usual parameter training.

The dimensionality of LSI and IRR in our experiments was set to the number of topics, as the theoretical result in Section 3.4 indicates that if LSI outperforms VSM largely, the dimensionality should be around the number of topics. We show experimental results which validate this prediction in Section 4.3.3.

4.3.2 Controlled-distribution results

We first examine the kappa average precision results, shown in Figure 4.5. The x-axis represents the non-uniformity of the topic-document distribution as measured by $\mathcal{T}_{max}/\mathcal{T}_{min}$. First, we see that when the topic-document distribution is relatively uniform, LSI's performance is higher than 90%. However, as the non-uniformity increases, the performance of LSI declines sharply, as predicted by our theorems from Chapter 3.

Also, our interpretation of the scaling factor q as compensating for non-uniformity is borne out nicely. For very uniform distributions, the performance difference between $q = 0$ (at which IRR = LSI), $q = 2$, and $q = 4$ is not great. At medium non-uniformity, $q = 0$ degrades, but $q = 2$ still does about the same as $q = 4$. But as the non-uniformity increases even more, we see that $q = 2$ is not large enough to compensate, and so declines in comparison to $q = 4$.

Furthermore, we see that IRR with AUTO-SCALE (labeled 'IRR:q = auto') does extremely well across all levels of non-uniformity. Figure 4.9(a.1) shows that AUTO-SCALE indeed adjusts for more non-uniform distributions: the chosen scaling factor increases on average as the non-uniformity goes up. (We will examine Figure 4.9 in more detail later in Section 4.3.3.)

Now, one might conjecture that instead of using AUTO-SCALE, it would suffice simply to choose a single very large value of q. Intuitively, though, this is problematic, since too high a scaling factor would tend to completely eliminate residuals. Furthermore, the $q = 20$ curve in Figure

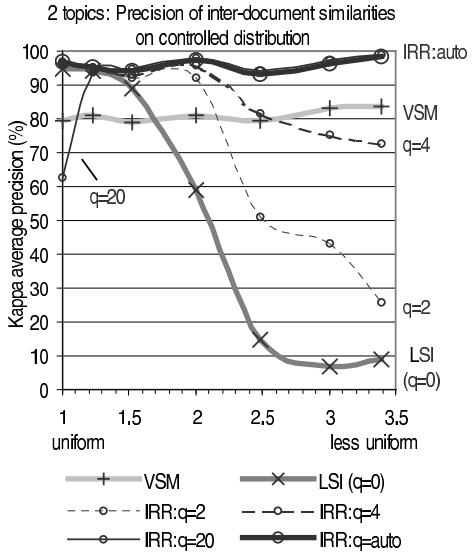

Figure 4.5: Kappa average precision results, two topics. Points are averages over ten document sets.

4.5 disproves the conjecture: we see that when the underlying topic-document distribution is relatively uniform, selecting an overly large scaling factor hurts performance, driving it below the baseline VSM curve.

We now examine our two-topic floor and ceiling clustering results, shown in Figure 4.6.

We see precisely the same types of behaviors as in the kappa average precision case. The floor performances are especially interesting, as they show that AUTO-SCALE-IRR exhibits very good performance for all six of our rather wide variety of clustering algorithms. They also indicate that VSM is 'fragile' for uniform distributions, in that sometimes it is a very poor representation for at least one of the (commonly-used) clustering algorithms we employed.

Finally, Figure 4.7 and 4.8 shows the results of the same evaluation experiments run on the three-topic, four-topic, and five-topic data. Again, the experimental results are completely in line with what we predicted, with AUTO-SCALE leading to the best performance overall. Note that the gap between LSI and VSM decreases in comparison to the $k = 2$ case; this is due to the fact that at higher dimensionalities, the subspace produced by LSI gets closer to that of the original term-document matrix.

These results all strongly support our theoretical claims.

4.3.3 Non-uniformity and scaling factor, the best dimensionality

Non-uniformity and scaling factor

To study the correlation between 'good' scaling factors and the degree of non-uniformity of the topic-document distribution, we measured kappa average precision on the two-topic and five-

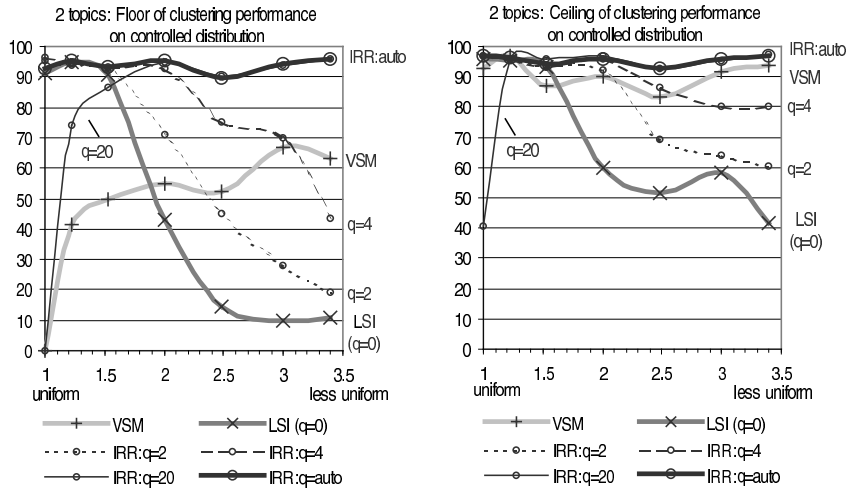

Figure 4.6: Floor and ceiling clustering results, two topics. Points are averages over ten document sets.

topic data sets in the previous sections by setting the scaling factor to values in the range of 0 to 19 in increments of 0.5. The dimensionality was set to the number of topics as in the previous section. The results (Figure 4.9) show that higher non-uniformity requires larger scaling factor as predicted, and that AUTO-SCALE produces larger scaling factor on higher non-uniformity as expected.

The dotted line (labeled 'best') in Figure 4.9 (a.1) shows the scaling factor values which produced the best kappa average precision results on average over the ten two-topic document sets. The x-axis represents the non-uniformity of the topic-document distribution as measured by $\mathcal{T}_{max}/\mathcal{T}_{min}$. We see that the trend of the best scaling factor increases with non-uniformity. The scaling factor selected by IRR AUTO-SCALE method (on average over ten sets; plotted by the solid line labeled 'auto') increases steadily with non-uniformity. Although the differences between 'best' and 'auto' scaling factors in (a.1) are apparently large, the differences in the produced kappa average precision are very small, as shown in (b.1). This is because, as observed in (c.1), the range of scaling factor to yield good performance (on average over ten sets) is relatively wide on this two-topic data.

As in the two-topic case, the trend of the best scaling factor on five-topic data increases with non-uniformity (Figure 4.9 (a.2)). The scaling factor selected by AUTO-SCALE method also increases with non-uniformity; however, it is smaller (larger) than the best value when the distribution is relatively uniform (non-uniform), respectively. As shown in (b.2), the performance produced by AUTO-SCALE rivals the best performance when the scaling factor is close to the best, and degrades otherwise. Figure 4.9 (c.2) indicates that five-topic data is more sensitive to the scaling factor setting than two-topic data; the curve for each distribution type has a clear peak of performance rather than a plateau as in two-topic case.

The results in Figure 4.9 show that the best scaling factor depends on the degree of non-uniformity of topic-document distribution; *larger non-uniformity requires larger scaling factor*, as predicted. It indicates that to obtain good performance by training the scaling factor (instead

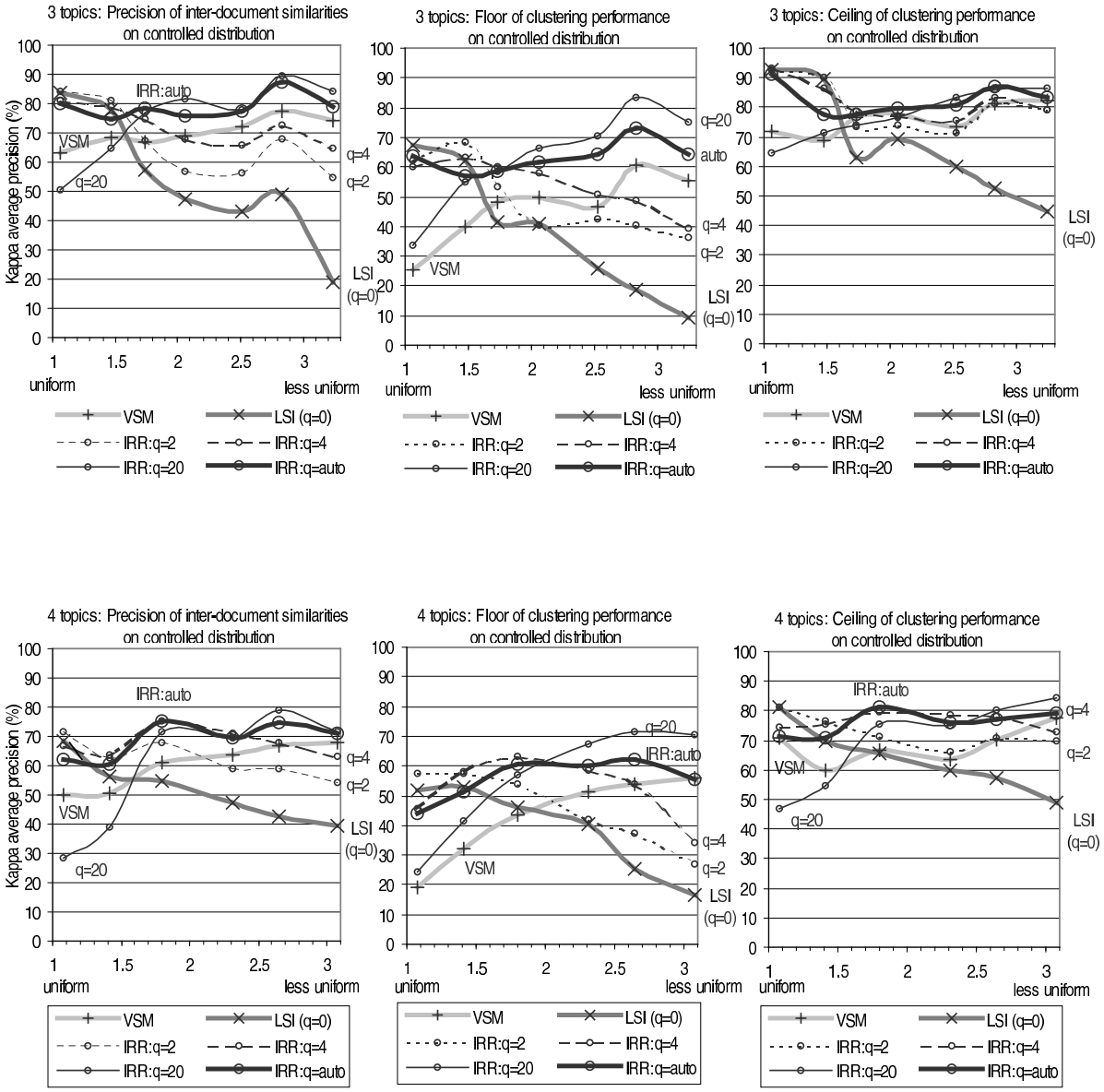

Figure 4.7: Kappa average precision and floor and ceiling clustering results, three topics, four topics. Points are averages over ten document sets.

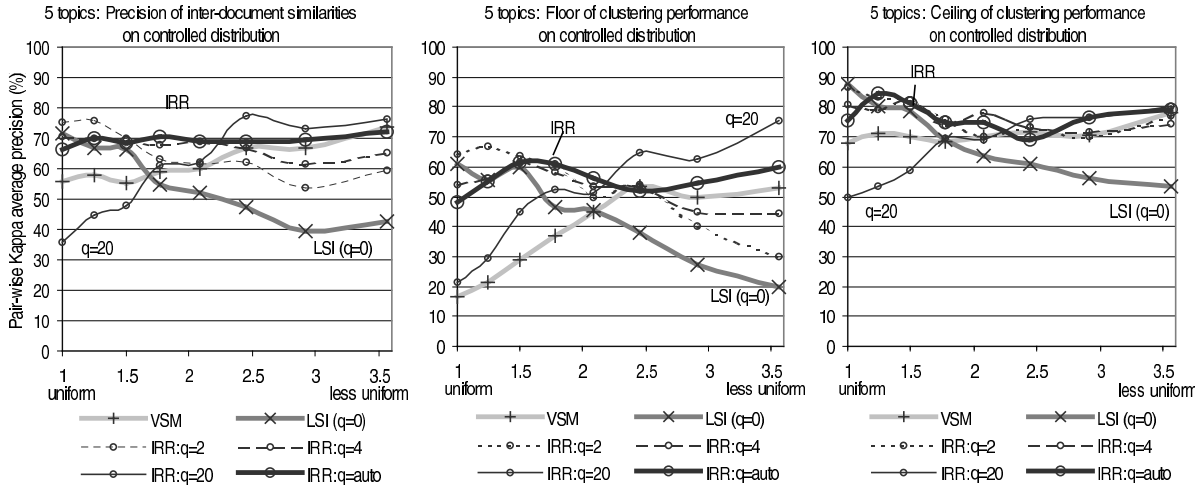

Figure 4.8: Kappa average precision and floor and ceiling clustering results, five topics. Points are averages over ten document sets.

of using the AUTO-SCALE method), the distribution type of training data must be similar to that of test data; Figure 4.9 (c.1) and (c.2) clearly show that mismatching of distribution type between training and test data would cause significant performance degradation. We confirm that the AUTO-SCALE method indeed produces larger scaling factors for larger non-uniformity as desired, and it serves as an alternative method of scaling factor selection in the case that the distribution type of the test data is unknown.

Best dimensionality

We studied the dimensionality with which LSI and IRR (with AUTO-SCALE) produced the best kappa average precision on each of the two-topic and five-topic data sets in the previous sections. The results show that *when the performance is largely improved over* VSM*, the dimensionality is close to the number of topics*, which validates our theoretical prediction in Section 3.4.

Each point in Figure 4.10 represents the best dimensionality (x-axis) and performance improvement over VSM (y-axis) on each of data sets over all distribution types as described in Section 4.3.1. The evaluation metric is kappa average precision. In all cases ((a) LSI on two topics, (b) IRR with AUTO-SCALE on two topics, (c) LSI on five topics, and (d) IRR with AUTO-SCALE on five topics), we see the points form two groups; in the first group the best dimensionality is close to the number of topics, and performance improvement is relatively large; in the second group the best dimensionality is close to the rank of the term-document matrix, and performance improvement is nearly zero. This observation confirms our prediction that the dimensionality should be around the number of topics if performance is to be largely improved over VSM. Also note that the term-document matrix is reproduced when the dimensionality reaches the rank of the term-document matrix; therefore, the performance of the second group can not differ much from VSM, as shown in Theorem 3.4.1 (and as is intuitively clear).

Furthermore, Figure 4.10 shows that LSI failed to gain large improvement over VSM on almost

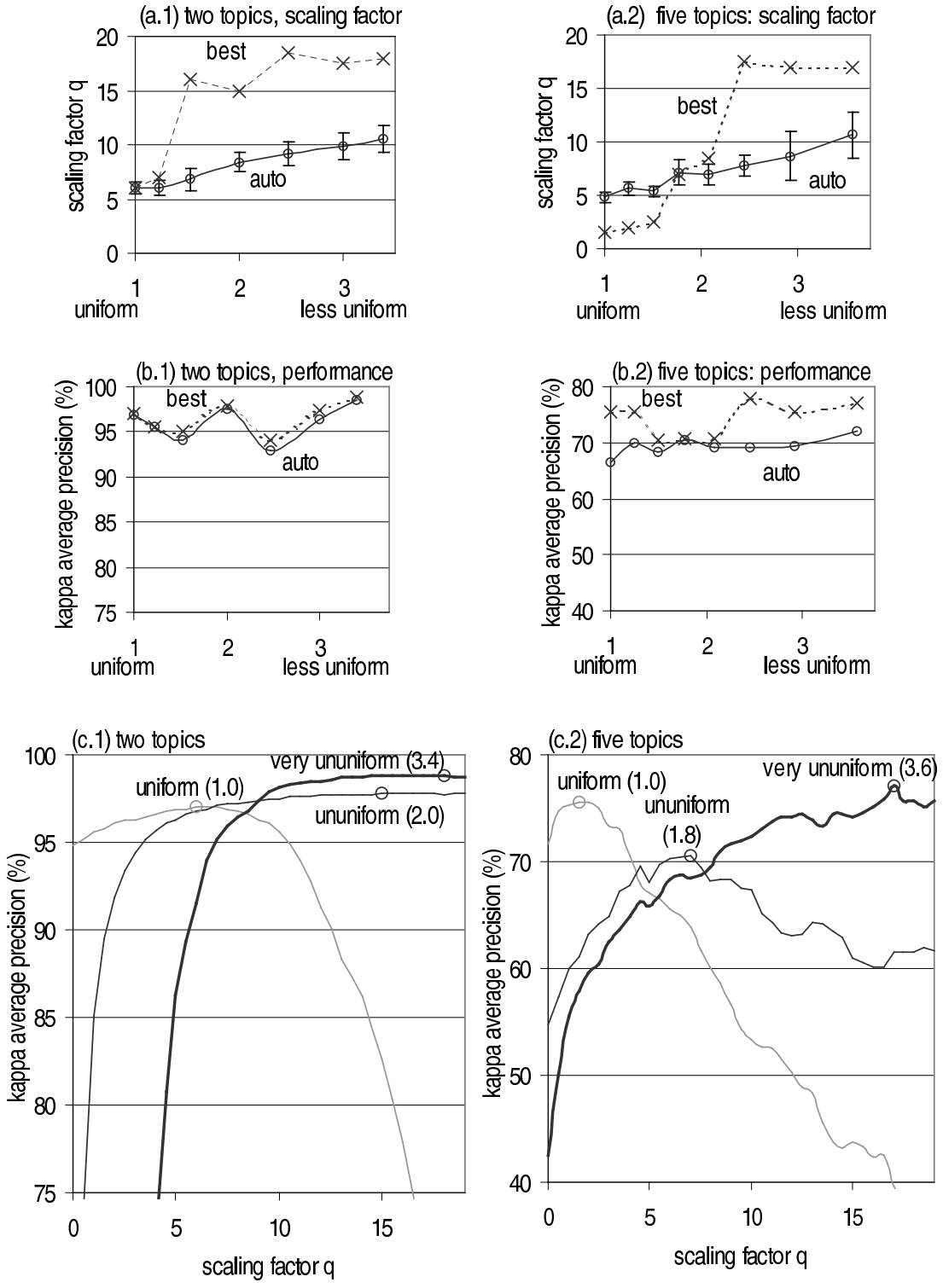

Figure 4.9: (a) scaling factor and non-uniformity. 'auto': selected by AUTO-SCALE; 'best': the values which produced the best performance. (b) kappa average precision results and non-uniformity; 'auto': performance produced by AUTO-SCALE; 'best': best over all the scaling factor settings. (c) kappa average precision results and scaling factor; the numbers in the labels are non-uniformity $\mathcal{T}_{max}/\mathcal{T}_{min}$. The circles are the peaks of the performance. All results are averages over 10 sets; error bars represent one standard deviation.

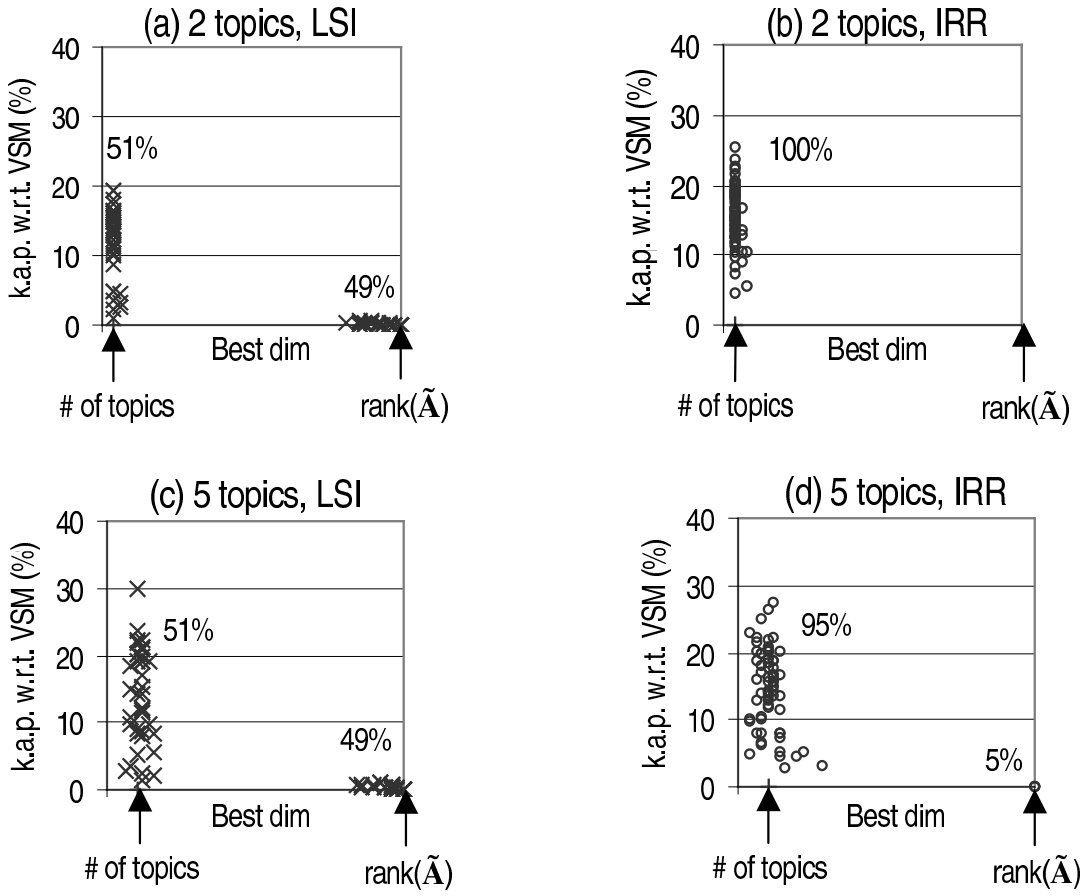

Figure 4.10: Best dimensionality and absolute improvement of kappa average precision over VSM. Each point represents the dimensionality to produce the best performance (x-axis) and the absolute improvement of best performance over VSM (y-axis), for each data set. (a) 2-topics, LSI; (b) 2-topics, IRR with AUTO-SCALE; (c) 5-topics, LSI; (d) 5-topics, IRR with AUTO-SCALE. Two groups are observed in each figure; the number is the percentage of population in each group.

half (49%) of the data sets both in 2-topic and 5-topic cases, while IRR outperformed VSM on most of the datasets (100% in 2-topic case and 95% in 5-topic case). Figure 4.11 and 4.12 show that as non-uniformity increases, LSI has less and less data sets belonging to the first group, i.e., LSI achieves performance improvement over VSM on less and less data sets.

	uniform				less uniform		
LSI (# of sets)	10	10	10	3	0	3	0
IRR	10	10	10	10	10	10	10

Figure 4.11: Two-topic data, size of the 'first group' (data sets whose best dimensionality was close to the number of topics k) on increasing non-uniformity, max size 10.

	uniform					less uniform		
LSI (# of sets)	10	10	9	6	5	0	0	1
IRR	10	10	10	10	10	9	8	9

Figure 4.12: Five-topic data, size of the 'first group' (data sets whose best dimensionality was close to the number of topics k) on increasing non-uniformity, max size 10.

4.4 Evaluation on unrestricted distributions

In this section, we experiment on the more realistic setting of document sets without distribution restrictions. We expect that in realistic data, topic-document distributions will be fairly non-uniform, so that IRR should perform well in comparison to LSI.

Figure 4.13 summarizes the evaluation settings.

Metric	Kappa average precision		Clustering	
Assume k (# of topics) is:	Given	Not given	Given	Not given
Dimensionality	k	Trained	k Trained	Trained
# of clusters	N/A		k	Dimensionality*
Section (Figure)	Section 4.4.2 (Table 4.14)		Section 4.4.3 (Figure 4.16)	

Figure 4.13: Evaluation settings for unrestricted distributions. (*) For VSM, we use the average number of topics.

4.4.1 Data

We used 648 TREC documents, each relevant to exactly one of twenty TREC topics. To perform parameter training, we randomly divided these documents into two disjoint document pools. We then simulated input from an information retrieval application by generating 15 document sets from each pool, where each set consisted of those documents containing one of 15 arbitrarily chosen keywords; this yielded a total of 30 document sets. Document sets from one pool were used as parameter training data for the other pool, and vice versa. Performance results are

dimensionality?	LSI	IRR
number of topics	-8.7	1.4
trained	0	3.95

Figure 4.14: Absolute improvement in pair-wise kappa average precision over VSM (51.4%), unrestricted distributions: averages over 30 runs.

Dimensionality	Trained	k
p (IRR vs VSM)	0.0383	*0.2938*
p (IRR vs LSI)	0.0383	< 0.0001

Figure 4.15: p-values from the paired t-test on the kappa average precision results. The performance differences are statistically significant ($p < 0.05$) in all but the one italicized.

averages over these 30 runs. The scaling factor for IRR was determined by AUTO-SCALE in all cases (with the same constants α and β as before). The term-document matrices were created in the same manner as in the controlled-distributed data experiments.

4.4.2 Kappa average precision

Recall that we consider two ways to choose the dimensionality of a document subspace. In the first case, the system knows the number k of topics underlying the collection (for example, this information could be given by a user as a way to control topic granularity, or by a set of predetermined classification labels), and sets the dimensionality to k. In the second case, we simply train the dimensionality parameter, using the residual ratio method from Section 4.1.2 on the held-out data from the other document pool, regardless of the availability of k.

From Figure 4.14, we see that IRR yields higher kappa average precision than LSI and VSM for both dimensionality selection methods, and therefore does a better job at representing inter-document similarities. LSI performs relatively poorly on this task; indeed, using k dimensions in the LSI case leads to worse results than VSM.

The paired t-test (Figure 4.15) shows that the mean differences of kappa average precision results are statistically significant ($p < 0.05$) between IRR and LSI with the both dimensionality selection methods, and between IRR and VSM when the dimensionality is trained. (The paired differences passed the Gaussian normality test in all cases.)

4.4.3 Clustering results

To derive floor and ceiling clustering performance results, there are two parameters we need to specify: the dimensionality of the representation subspace, and the desired number of clusters.

If k, the number of topics, is available, then it is the natural choice for the number of clusters. As for the dimensionality in this case, one option is to set it to k as well; Figure 4.16(a) shows the results. IRR's ceiling is 5% higher than VSM and 7.4% higher than LSI— note that LSI's

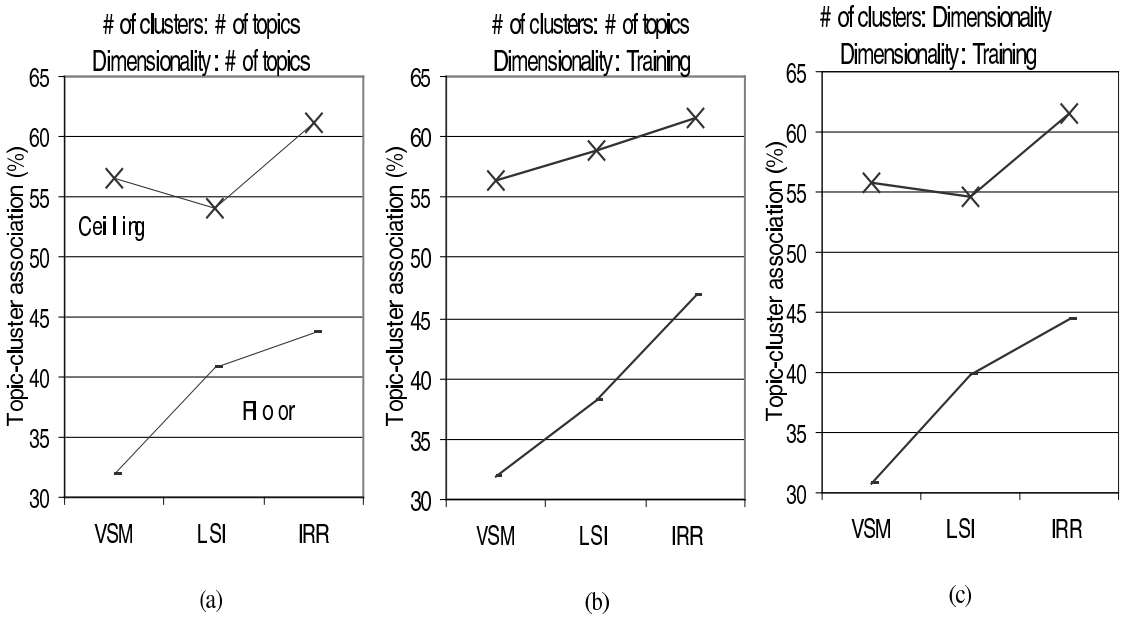

Figure 4.16: Document clustering performances, unrestricted distributions: averages over 30 runs.

ceiling is lower than VSM's. IRR's floor is also better: 12% higher than VSM and 3% higher than LSI.

When k is given but we train the dimensionality, using the residual ratio as specified in Section 4.1.2, then, according to Figure 4.16(b), the ceiling of IRR is 5.2% higher than VSM and 2.8% higher than LSI. The floor of IRR is 15% higher than that of VSM and 8.7% higher than that of LSI. Again, IRR provides a better subspace for the whole range of clustering algorithms we considered. We observe that for this type of data, training the dimensionality allows LSI to produce improved ceiling results.

We now consider the setting in which k is unknown. In this situation, we know of no alternative to training the dimensionality on held-out data. As for the number of clusters, a reasonable default is to simply set this value to the trained dimensionality. Of course, this doesn't apply to VSM since the dimensionality is not a free parameter for it; instead, we set the number of clusters to the average of the number of topics over the training document sets.

Figure 4.16(c) shows the clustering results for the unknown-k setting. LSI's ceiling degrades by 4.3% compared with when the number of topics is given, while those of VSM and IRR show almost no change. Furthermore, IRR clearly outperforms the other methods.

The paired t-test (Figure 4.17) shows that the mean differences of clustering performance results are statistically significant ($p < 0.05$) between IRR and VSM in all the settings, and between IRR and LSI in all but the ceiling performance when the dimensionality is trained and the number of clusters is k. (The paired differences passed the Gaussian normality test in all cases.)

Dimensionality	k		Trained		Trained	
# of clusters	k		k		Dimensionality	
Metric	ceiling	floor	ceiling	floor	ceiling	floor
p (IRR vs VSM)	0.0203	< 0.0001	0.0025	< 0.0001	0.0074	0.0017
p (IRR vs LSI)	0.0019	0.0212	*0.0775*	< 0.0001	0.0003	0.0196

Figure 4.17: p-values from the paired t-test on the clustering performance results. The performance differences are statistically significant ($p < 0.05$) in all the settings but the one italicized.

4.4.4 Discussion

In our experiments, LSI actually performed worse or essentially the same as VSM in 4 out of 8 combinations of practical settings and metrics. In particular, when the dimensionality is chosen to be the number of topics in realistic data, LSI performs relatively poorly. Dimensionality training improves LSI's kappa average precision scores, and also improves its clustering performance with respect to VSM as long as the correct number of clusters (i.e., the number of topics) is given. However, when the number of clusters is not given, LSI's ceiling clustering performance drops, again indicating that for LSI the dimensionality should not be tied to the number of clusters.

In contrast, IRR consistently performs better than LSI and VSM in all the settings and for all the metrics. In particular, IRR fares pretty well when the dimensionality is set to the number of topics as compared to when the dimensionality is actually trained. These results suggest that when the number of topics is known and dimensionality training is expensive, setting the dimensionality to the number of topics yields reasonable results. Furthermore, in clustering applications for which the number of topics is not known, we at least might be able to reduce the training effort by only searching for the dimensionality, setting the number of clusters to the same value.

Chapter 5

Scaling IRR up

In this chapter, we introduce an extension of IRR for faster computation on larger collections: *Sampled IRR* (SP-IRR). We show that SP-IRR speeds up IRR without significantly degrading the quality of the resultant subspace, rather, improving it in certain settings.

Notational conventions and assumptions As in the previous chapters, we let $\widetilde{\mathbf{A}} \in \Re^{m \times n}$ denote a term-document matrix and assume that $m > n$ (i.e., the number of terms is larger than the number of documents). The symbols ℓ and t refer to the dimensionality of an IRR subspace to be computed and the average number of term types in one document, respectively, (so that tn is precisely the number of non-zero entries in $\widetilde{\mathbf{A}}$). We assume $t \ll m$ and $\ell \ll n$.

5.1 Implementation and computation time for IRR

Term-document matrices, from which IRR computes subspaces, are typically sparse. Figure 5.1 shows an example of an efficient implementation of IRR and its asymptotic computation time assuming an efficient data structure for sparse matrices. Contrast this with the high-level pseudocode shown in Figure 4.3. It is useful to maintain an *inner product matrix* of residuals, $\mathbf{P} = (\mathbf{R}^{(i)})^T \mathbf{R}^{(i)}$, which can be used to determine the scaling factor q (Step 2) and to rescale the residuals (Step 4). The next right singular vector $\mathbf{v}_i^{\mathrm{IRR}}$ of the (implicit) rescaled residual matrix is obtained by computing the first eigenvector of the rescaled inner product matrix $\mathbf{P'}$ (Step 5), which runs in $O(n^2)$, e.g., by the Power method (Golub and Van Loan, 1996). It is turned into the next left singular vector $\mathbf{u}_i^{\mathrm{IRR}}$ in Step 6 by multiplying it with the rescaled term-document matrix $\widetilde{\mathbf{A}}$ diag$(s[1], \ldots, s[n])$ (where $s[i]$ is the degree to which the ith residual is rescaled) and subtracting the projections onto $\mathbf{u}_j^{\mathrm{IRR}}$ for $j < i$; observe that we have

$$\begin{aligned}
\mathbf{u}_i^{\mathrm{IRR}} &= \rho\, \mathbf{R}^{(i)}\, \mathrm{diag}(s[1], \ldots, s[n])\, \mathbf{v}_i^{\mathrm{IRR}} \\
&= \rho \left(\widetilde{\mathbf{A}} - \mathrm{proj}(\widetilde{\mathbf{A}}, \{\mathbf{u}_1^{\mathrm{IRR}}, \ldots, \mathbf{u}_{i-1}^{\mathrm{IRR}}\}) \right) \mathrm{diag}(s[1], \ldots, s[n])\, \mathbf{v}_i^{\mathrm{IRR}} \\
&= \rho\, \widetilde{\mathbf{A}}\, \mathrm{diag}(s[1], \ldots, s[n])\, \mathbf{v}_i^{\mathrm{IRR}} - \rho \sum_{j=1}^{i-1} \mathbf{u}_j^{\mathrm{IRR}} (\mathbf{u}_j^{\mathrm{IRR}})^T (\widetilde{\mathbf{A}}\, \mathrm{diag}(s[1], \ldots, s[n])\, \mathbf{v}_i^{\mathrm{IRR}}),
\end{aligned}$$

IRR(ℓ):

$\mathbf{P} := \widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}$ // Step 1: $O(tn^2)$; compute inner products

$q = \alpha * \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbf{P}_{[i,j]}\right)^2/n^2 + \beta$ // Step 2: $O(n^2)$; determine scaling factor q

For $i := 1, 2, \ldots, \ell$

  For $j := 1, 2, \ldots, n$ // Step 3: $O(\ell n)$; determine residual scale

    $s[j] := \left(\sqrt{\mathbf{P}_{[j,j]}}\right)^q$ // ($\sqrt{\mathbf{P}_{[j,j]}}$ is the jth residual length)

  For $j := 1, 2, \ldots, n$ // Step 4: $O(\ell n^2)$; rescale residual inner

    For $h := 1, 2, \ldots, n$ // products

      $\mathbf{P'}_{[j,h]} := \mathbf{P}_{[j,h]} * s[j] * s[h]$

  $\mathbf{v}_i^{\text{IRR}}$:= first eigenvector of $\mathbf{P'}$ // Step 5: $O(\ell n^2)$; right singualr vector

  $\mathbf{u'} := \widetilde{\mathbf{A}}\ \text{diag}(s[1], \ldots, s[n])\mathbf{v}_i^{\text{IRR}}$ // Step 6: $O(\ell tn + m\ell^2)$; left singular vector

  $\mathbf{u}_i^{\text{IRR}} := \mathbf{u'} - \sum_{j=1}^{i-1}\mathbf{u}_j^{\text{IRR}}(\mathbf{u}_j^{\text{IRR}})^T\mathbf{u'}$

  $\mathbf{u}_i^{\text{IRR}} := \mathbf{u}_i^{\text{IRR}}/||\mathbf{u}_i^{\text{IRR}}||_2$

  For $j := 1, 2, \ldots, n$ // Step 7: $O(\ell tn)$; projection size

    $\mathbf{A}_{[i,j]}^{\text{IRR}} := \widetilde{\mathbf{a}}_j^T\mathbf{u}_i^{\text{IRR}}$

  For $j := 1, 2, \ldots, n$ // Step 8: $O(\ell n^2)$; update residual inner

    For $h := 1, 2, \ldots, n$ // products

      $\mathbf{P}_{[j,h]} := \mathbf{P}_{[j,h]} - \mathbf{A}_{[i,j]}^{\text{IRR}} * \mathbf{A}_{[i,h]}^{\text{IRR}}$

// $\mathbf{A}^{\text{IRR}}$ is the new document representation in IRR subspace

Figure 5.1: Example implementation of IRR.

where ρ is a length-normalization constant. Thus, the next left singular vector is computed in $O(n^2 + tn + m\ell)$, faster than computing $\mathbf{u}_i^{\text{IRR}}$ directly in $O(m^2)$, as we assume $\ell \ll n < m$. The residual inner product matrix $\mathbf{P}$ is updated by $\mathbf{P}_{[j,h]} := \mathbf{P}_{[j,h]} - \mathbf{A}_{[i,j]}^{\text{IRR}} * \mathbf{A}_{[i,h]}^{\text{IRR}}$ (Step 8), which derives from the fact that

$$\begin{aligned}
(\mathbf{r}_j^{(i+1)})^T\mathbf{r}_h^{(i+1)} &= \left(\mathbf{r}_j^{(i)} - \text{proj}(\mathbf{r}_j^{(i)}, \{\mathbf{u}_i^{\text{IRR}}\})\right)^T\left(\mathbf{r}_h^{(i)} - \text{proj}(\mathbf{r}_h^{(i)}, \{\mathbf{u}_i^{\text{IRR}}\})\right) \\
&= (\mathbf{r}_j^{(i)})^T\mathbf{r}_h^{(i)} - \left((\mathbf{u}_i^{\text{IRR}})^T\mathbf{r}_j^{(i)}\right)\left((\mathbf{u}_i^{\text{IRR}})^T\mathbf{r}_h^{(i)}\right), \\
\mathbf{A}_{[i,j]}^{\text{IRR}} &= (\mathbf{u}_i^{\text{IRR}})^T\widetilde{\mathbf{a}}_j = (\mathbf{u}_i^{\text{IRR}})^T\mathbf{r}_j^{(i)}.
\end{aligned}$$

Note that when we let $\mathbf{U}^{\text{IRR}} = [\,\mathbf{u}_1^{\text{IRR}} \ \cdots \ \mathbf{u}_\ell^{\text{IRR}}\,]$, the implementation in Figure 5.1 results in the new document representation $\mathbf{A}^{\text{IRR}} = (\mathbf{U}^{\text{IRR}})^T\widetilde{\mathbf{A}}$, which is more compact representation than $\text{proj}(\widetilde{\mathbf{A}}, \{\mathbf{u}_1^{\text{IRR}}, \ldots, \mathbf{u}_\ell^{\text{IRR}}\}) = \mathbf{U}^{\text{IRR}}(\mathbf{U}^{\text{IRR}})^T\widetilde{\mathbf{A}}$ (given in Figure 4.3), but producing exactly the same inter-document similarities; observe that

$$\begin{aligned}
(\mathbf{A}^{\text{IRR}})^T\mathbf{A}^{\text{IRR}} &= \widetilde{\mathbf{A}}^T\mathbf{U}^{\text{IRR}}(\mathbf{U}^{\text{IRR}})^T\widetilde{\mathbf{A}} \\
&= (\mathbf{U}^{\text{IRR}}(\mathbf{U}^{\text{IRR}})^T\widetilde{\mathbf{A}})^T(\mathbf{U}^{\text{IRR}}(\mathbf{U}^{\text{IRR}})^T\widetilde{\mathbf{A}}) \\
&= \text{proj}(\widetilde{\mathbf{A}}, \{\mathbf{u}_1^{\text{IRR}}, \ldots, \mathbf{u}_\ell^{\text{IRR}}\})^T\text{proj}(\widetilde{\mathbf{A}}, \{\mathbf{u}_1^{\text{IRR}}, \ldots, \mathbf{u}_\ell^{\text{IRR}}\}).
\end{aligned}$$

As a result of taking advantage of the sparseness of $\widetilde{\mathbf{A}}$, an ℓ-dimensional IRR subspace is obtained

in $O(tn^2 + \ell n^2 + \ell tn + m\ell^2)$.

In practice, the eigenvector computation in Step 5 dominates the runtime when the number of documents n is large, as it is quadratic in n with a relatively large constant.

5.2 Random-IRR

To deal with a large TREC corpus of n documents, Dumais (1993, 1994, 1995) created an LSI subspace from *randomly sampled* $n' < n$ documents, which we call *random-LSI* in this thesis. (Note that it differs from *random projection* (Papadimitriou et al., 2000), which we will discuss in Section 8.2.2.) Given a term-document matrix $\widetilde{\mathbf{A}}$, random-LSI takes three steps:

- create $\widetilde{\mathbf{A}}'$ by randomly sampling n' columns of $\widetilde{\mathbf{A}}$,
- compute the first k left singular vectors $\mathbf{u}'_1, \ldots, \mathbf{u}'_k$ of $\widetilde{\mathbf{A}}'$, and
- project $\widetilde{\mathbf{A}}$ onto the subspace spanned by $\mathbf{u}'_1, \ldots, \mathbf{u}'_k$.

Since the SVD is applied to $\widetilde{\mathbf{A}}' \in \Re^{m \times n'}$, which is of smaller dimensions than $\widetilde{\mathbf{A}}$, computation time is reduced from being quadratic in n to quadratic in n'. This technique is sometimes referred as 'fold-in' since $n - n'$ documents are just folded in (projected onto)[1] the subspace. Jiang et al. (1999a) studied the performance of random-LSI; we discuss their results in Section 8.2.1.

Analogously, we can have a faster alternative to IRR, *random*-IRR, which also reduces (one factor of) the computation time n^2 to n'^2 by constructing an IRR subspace from n' randomly sampled documents and projecting all the document vectors onto that subspace.

In Chapter 4, we showed that the quality of an LSI subspace (as a document representation space) has a dependency on the uniformity of underlying topic-document distribution. We also confirmed that IRR can compensate for the non-uniformity when the scaling factor is adjusted to the degree of non-uniformity. Sampling of a sufficiently large number of documents would reflect the original distribution, so random-IRR and random-LSI should not degrade (or improve) the performance very much over IRR and LSI, respectively. However, when n' is relatively small, the distribution in the sampled documents may deviate highly from the original distribution. In particular, the documents on less dominant topics may not be sampled at all, which would result in degrading the quality of the subspace significantly. In the next section, we propose an alternative sampling method which attempts simultaneously to compensate for non-uniformity while reducing the computation time.

[1]Note that both in LSI and random-LSI all the document vectors in a collection are 'folded in' (projected onto) the subspace. The difference of random-LSI from LSI lies in that those $n-n'$ documents are not used to construct the subspace and merely folded in the subspace.

5.3 Sampled IRR (SP-IRR)

For faster computation on larger document collections, we extend IRR by *strategically reducing the number of documents* used in the process of eigenvector computation; we call this new algorithm SP-IRR (which stands for Sampled IRR).

Recall from Section 4.1 that the IRR-basis vector $\mathbf{u}_i^{\mathrm{IRR}}$ is defined by

$$\begin{aligned}
\mathrm{pow}(\mathbf{r}, q) &= ||\mathbf{r}||_2^q \, \mathbf{r}\,, \\
\mathbf{u}_i^{\mathrm{IRR}} &= \arg\max_{||\mathbf{x}||_2=1} \sum_{j=1}^{n} \left(||\mathrm{pow}(\mathbf{r}_j^{(i)}, q)||_2 \, \cos(\mathbf{x}, \mathrm{pow}(\mathbf{r}_j^{(i)}, q)) \right)^2 , \\
\mathbf{R}^{(1)} &= \widetilde{\mathbf{A}}\,, \\
\mathbf{R}^{(i+1)} &= \mathbf{R}^{(i)} - \mathrm{proj}(\mathbf{R}^{(i)}, \{\mathbf{u}_i^{\mathrm{IRR}}\})\,.
\end{aligned}$$

Letting $\mathbf{r}'_j = \mathrm{pow}(\mathbf{r}_j^{(i)}, q)$, $\mathbf{u}_i^{\mathrm{IRR}}$ satisfies

$$\mathbf{u}_i^{\mathrm{IRR}} = \lambda \mathbf{R}'\mathbf{R}'^T\mathbf{u}_i^{\mathrm{IRR}} = \lambda \sum_{j=1}^{n} \left(||\mathbf{r}'_j||_2 \, \cos(\mathbf{r}'_j, \mathbf{u}_i^{\mathrm{IRR}}) \right) \mathbf{r}'_j$$

for some λ. We see that in a sense $\mathbf{u}_i^{\mathrm{IRR}}$ is composed of a 'cluster' of rescaled residual vectors, and the vectors very far from this cluster make very small contributions. Based on this observation, we seek to leave out the vectors that eventually make little contribution to $\mathbf{u}_i^{\mathrm{IRR}}$ *before* computing $\mathbf{u}_i^{\mathrm{IRR}}$. This is done by selecting *a subset of large relatively close residual vectors*. Given $n' < n$, in the ith iteration, we select the largest residual vector $\mathbf{r}_{largest}^{(i)}$ first, and $(n'-1)$ vectors closest to $\mathbf{r}_{largest}^{(i)}$ next, as illustrated in Figure 5.2. This computation is in $O(n + n\log(n))$, assuming that $(\mathbf{R}^{(i)})^T\mathbf{R}^{(i)}$ (i.e., $\mathbf{P}$ in Figure 5.1) is maintained throughout the process. The motivation for choosing $\mathbf{r}_{largest}$ is *compensation for non-uniform topic-document distributions by amplifying less dominant topics*, since the larger residual vectors are likely to represent the documents on less dominant topics; see Section 4.1.

Let S be the set of the indices of thus chosen documents. We compute the ith basis vector $\mathbf{u}_i^{\mathrm{SP}}$ for $i > 1$ by

$$\mathbf{u}_i^{\mathrm{SP}} = \arg\max_{||\mathbf{x}||_2=1} \sum_{j\in S} \left(||\mathrm{pow}(\mathbf{r}_j, q)||_2 \, \cos(\mathrm{pow}(\mathbf{r}_j, q), \mathbf{x}) \right)^2 ,$$

using the scaling factor q and scaling function $\mathrm{pow}(\mathbf{r}, q)$ as in IRR.

In the first iteration where there is no clue for selection, we can approximate $\mathbf{u}_1^{\mathrm{IRR}}$ by following, more efficient calculation:

$$\mathbf{u}_1^{\mathrm{IRR}} \approx \mathbf{u}_1^{\mathrm{SP}} = \frac{\sum_{j=1}^{n} \widetilde{\mathbf{a}}_j}{||\sum_{j=1}^{n} \widetilde{\mathbf{a}}_j||_2}\,,$$

which satisfies

$$\mathbf{u}_1^{\mathrm{SP}} = \arg\max_{||x||_2=1} \sum_{j=1}^{n} ||\widetilde{\mathbf{a}}_j||_2 \, \cos(\widetilde{\mathbf{a}}_j, \mathbf{x})\,,$$

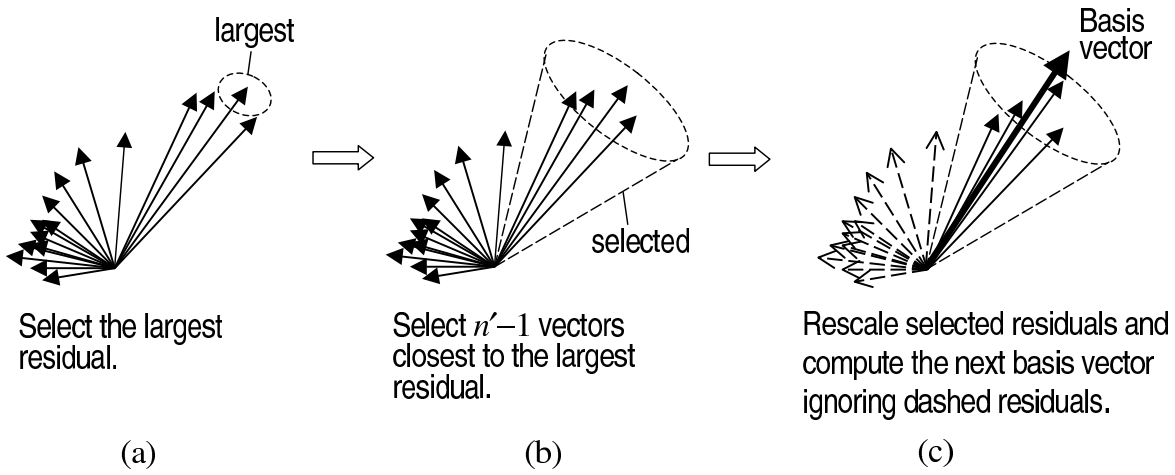

Figure 5.2: Schematic illustration of SP-IRR's residual vector selection.

instead of

$$\mathbf{u}_1^{\mathrm{IRR}} = \arg \max_{||x||_2=1} \sum_{j=1}^{n} (||\widetilde{\mathbf{a}}_j||_2 \ \cos(\widetilde{\mathbf{a}}_j, \mathbf{x}))^2 \ .$$

Although the total asymptotic computation time remains the same, SP-IRR reduces the computation time of the bottleneck process (the eigenvector computation in Step 5 of Figure 5.1) from $O(\ell n^2)$ to $O(\ell n'^2)$. Figure 5.3 shows the high-level pseudocode of SP-IRR.

SP-IRR(q,ℓ, n'):
- $\mathbf{R} := \widetilde{\mathbf{A}}$ // initialize residuals
- // by given term-document matrix
- $\mathbf{u}_1^{\mathrm{SP}} := \sum_{i=1}^{n} \mathbf{r}_i / || \sum_{i=1}^{n} \mathbf{r}_i ||_2$ // approximate the eigenvector
- For $j := 2, 3, \ldots, \ell$ // create ℓ basis vectors in total
  - $p := \arg\max_i ||\mathbf{r}_i||_2$ // select the largest residual vector
  - $S := \{p\} \cup \{h \mid \mathbf{r}_h$ is the ith closest to $\mathbf{r}_p$ where $i < n'\}$
  - $\mathbf{u}_j^{\mathrm{SP}} := \arg\max_{\mathbf{x}} \left(\sum_{i \in S} (||\mathrm{pow}(\mathbf{r}_i, q)||_2 \ \cos(\mathrm{pow}(\mathbf{r}_i, q), \mathbf{x}))^2 \right)$
  - // Compute eigenvector
  - $\mathbf{R} := \mathbf{R} - \mathrm{proj}(\mathbf{R}, \{\mathbf{u}_j^{\mathrm{SP}}\})$ // subtract projections from residuals
- // new document representation
- $\widetilde{\mathbf{A}}_{\mathrm{SP-IRR}} := \mathrm{proj}(\widetilde{\mathbf{A}}, \{\mathbf{u}_1^{\mathrm{SP}}, \ldots, \mathbf{u}_\ell^{\mathrm{SP}}\})$

Figure 5.3: High-level pseudocode for SP-IRR.

We expect that SP-IRR is more robust than random-IRR in that when we gradually reduce n', random-IRR should 'collapse' earlier than SP-IRR. That is, since random-IRR samples documents just once at the beginning, if some (minority) topics are 'wiped out' from the sample, there is no chance for random-IRR to recover them. SP-IRR resamples documents with higher priority on the documents likely to be on less dominant topics, so we expect more robust compensation effects.

Algorithm	random-LSI	random-IRR	SP-LSI	SP-IRR
How to select documents	randomly (once)	randomly (once)	SP (per iteration)	SP (per iteration)
Scaling factor	$q = 0$	AUTO-SCALE	$q = 0$	AUTO-SCALE

Figure 5.4: Comparison of the algorithms.

5.4 Experiments of SP-IRR on controlled distributions

We study the performance dependence of SP-IRR and random-IRR on the uniformity of topic-document distributions and the number of (re)selected documents n'. The comparison with random-IRR confirms our expectation that SP-IRR compensates better for non-uniformity.

5.4.1 Settings

To study the performance dependence on the uniformity of topic-document distributions, we generated 10-topic data from the TREC collection resulting in ten sets of 500 documents each over the same ten TREC topics. Five sets were completely uniform ($\mathcal{T}_{max}/\mathcal{T}_{min} = 1$; fifty documents for each topic), and the other five sets were highly non-uniform ($\mathcal{T}_{max}/\mathcal{T}_{min} = 3.17$; 275 documents for one topic and 25 each for other nine topics). Documents were arbitrarily chosen from each topic. The term-document matrices were created in the same manner as in Section 4.3.1.

We applied four algorithms to these document sets: SP-IRR, random-IRR, and their LSI versions SP-LSI and random-LSI, summarized in Figure 5.4. The scaling factor was determined by AUTO-SCALE from all the documents for SP-IRR, and from the selected n' documents for random-IRR (with the same constants α and β as before).

5.4.2 SP-IRR: Controlled-distribution results

The figures in this section plot the average performance over five document sets for each distribution type when the dimensionality is the number of topics. For random-LSI and random-IRR we did five runs with different 'seeds' of random numbers. The x-axis is the number of (re)selected documents n'. Note that when $n' = 500 (= n)$, SP-IRR and random-IRR become IRR, and SP-LSI and random-LSI become LSI.

We start with a general overview of the kappa average precision (defined in Section 4.2) results on the non-uniform topic-document distribution (Figure 5.5 (a)). We see clear performance trends. When n' is relatively large (250 to 500), the IRR-based methods (plotted by circles) perform better than the LSI-based methods (plotted by '$\times$'s). The difference of performance between these two groups is as large as 20%. For smaller n', the SP-based methods (solid lines) largely outperform the randomized methods (dotted lines). For instance, when only 25 documents are selected, the SP-IRR outperforms random-IRR by 25.9%, and the SP-LSI outperforms random-LSI by 33.6%.

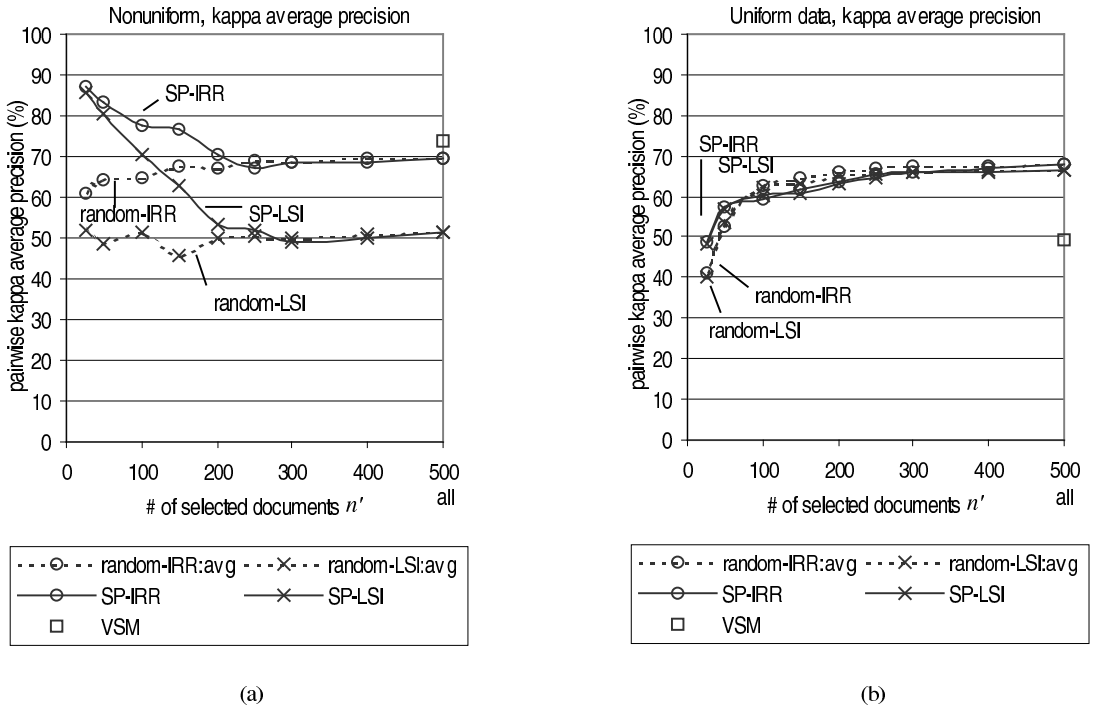

Figure 5.5: Kappa average precision on (a) non-uniform distribution and (b) uniform distribution of 500 documents over ten topics.

Now we turn to the kappa average precision results on the completely uniform data (Figure 5.5 (b)). The performance differences between methods are larger when fewer documents are selected, however, they are much smaller than on non-uniform data. When only 25 documents are selected, the SP-based methods outperform the randomized methods by 7.3 - 7.8%. The IRR-based methods are better than the LSI-based methods for larger n', although the differences are no greater than 2%.

Overall, SP-IRR outperforms the other three methods.

We note that, on the highly non-uniform data we generated, the SP-based methods perform better with smaller n', which indicates faster computation and better quality simultaneously. (For instance, SP-LSI with $n' = 25$ outperforms LSI by 34%.) Although this may look surprising, it is explained by the 'compensation effect' of SP, as discussed in the previous section. Recall that SP strategically *reselects* n' documents each time it adds one more dimension to the subspace. Even if only one-tenth of the documents are used at any one time (which speeds the computation up by nearly 100 times), reselection allows more documents to be utilized in total, while counteracting the non-uniformity of the distribution by favoring those possibly on less dominant topics. The results also indicate that IRR with AUTO-SCALE alone did not smooth out the non-uniformity largely enough for this data. In fact, the kappa average precision results produced by IRR (i.e., SP-IRR with $n' = n = 500$) were slightly lower than VSM on average, even though they were much higher than LSI. (This is due to the fact that constants α and β, which are used in AUTO-SCALE and determined once for all, were not optimal for this data. By exploring the range 0

to 30 in increments of 1, we found that the best scaling factor for this data is 29 on average, which produces kappa average precision 9% higher than VSM, while the scaling factor selected by AUTO-SCALE was 4.5 on average.) Conversely, though, we see that in a sense, SP complements IRR and AUTO-SCALE to compensate for the large non-uniformity.

For the completely uniform data, as long as a sufficiently large number of documents are selected, the methods do not differ from each other so much, as there is no non-uniformity to be counteracted. (The original document set is uniform, and the randomly selected documents should distribute somewhat uniformly, reflecting the original distribution.) However, a very small number of random samples may enlarge the deviation from the original distribution, increasing the non-uniformity in this case and possibly 'wiping out' some topics. This explains the relatively poor performance of the randomized methods for small n'.

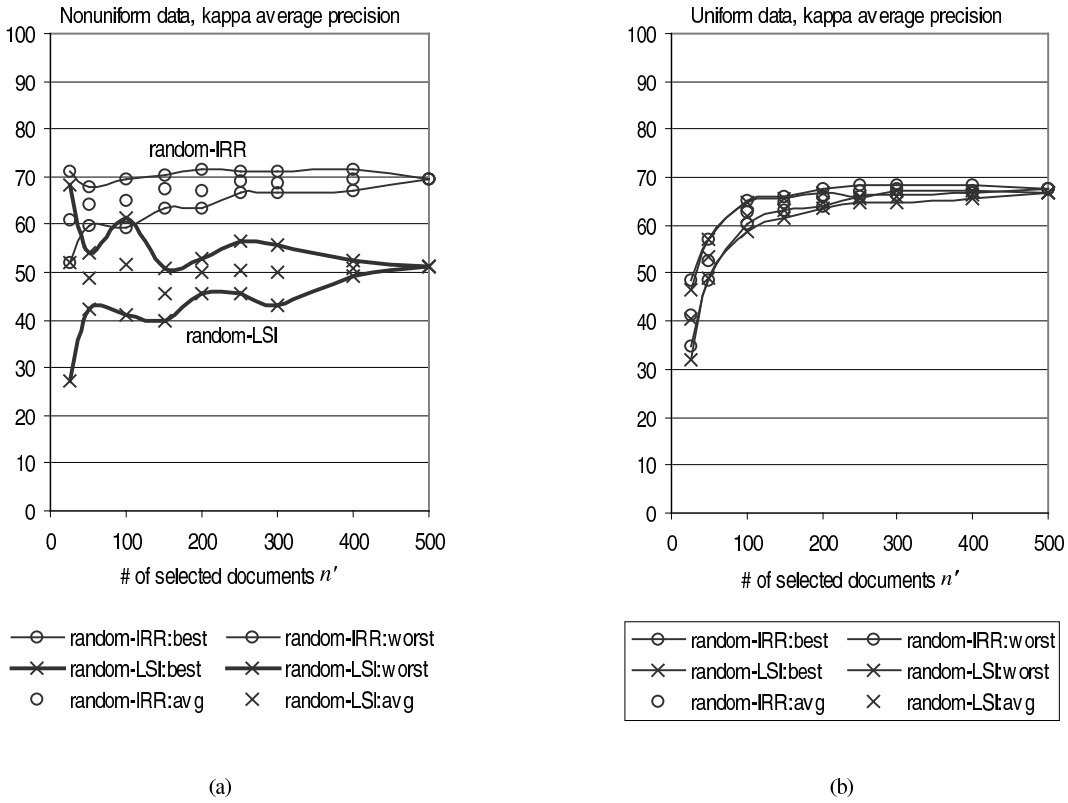

Figure 5.6: Kappa average precision of the randomized methods; the best, worst, and average over five runs with different 'seeds'.

Figure 5.6 plots the best and worst kappa average precision over five runs with different random seeds, for the randomized methods. The points are averages over five test sets. The variation in performance (over five seeds) is aggravated as n' becomes smaller, reaching differences of 41.1% (random-LSI) and 19% (random-IRR) on the non-uniform distribution (Figure 5.6 (a)), and 14.3% (random-LSI) and 13.8% (random-IRR) on the uniform distribution (Figure 5.6 (b)).

The trends of the ceiling and floor clustering performance results are similar to the kappa average precision results; see Figure 5.7 (and 5.8) for the ceiling (and floor) results. The floor results

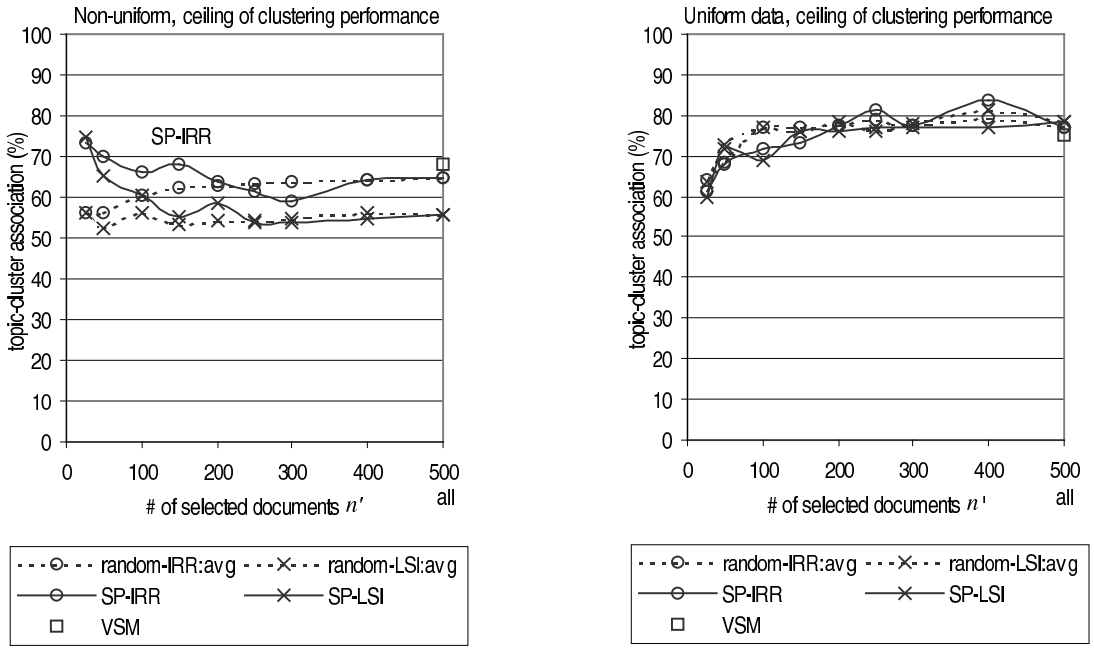

Figure 5.7: Ceiling clustering performance results on (a) non-uniform distribution and (b) uniform distribution of 500 documents over ten topics.

on uniform data (Figure 5.8(b)) exclude single-link, as single-link performed very poorly in all cases.

As a whole, the results meet our prediction that the randomized methods should be fragile for small n' especially on non-uniform distributions, and confirm that SP compensates for the non-uniformity. This also suggests that there should be a way to optimize the number of sampled documents n' (together with the scaling factor) in relation to non-uniformity.

5.5 Evaluation of SP-IRR on unrestricted distributions

In this section we experiment on the more realistic setting of document sets without distribution restrictions.

5.5.1 Data

We used 2091 TREC documents, each relevant to exactly one of fifteen TREC topics. We divided the documents into five disjoint document sets, each of which spans a time period of one year, based on the date tag contained in the document text. This setting simulates time-sliced focused collections, which would be encountered, for instance, when patent documents in a certain field are analyzed to find trends. For each set, we used the other four sets as training data when the dimensionality parameter was trained. Performance results are averages over five test sets. For

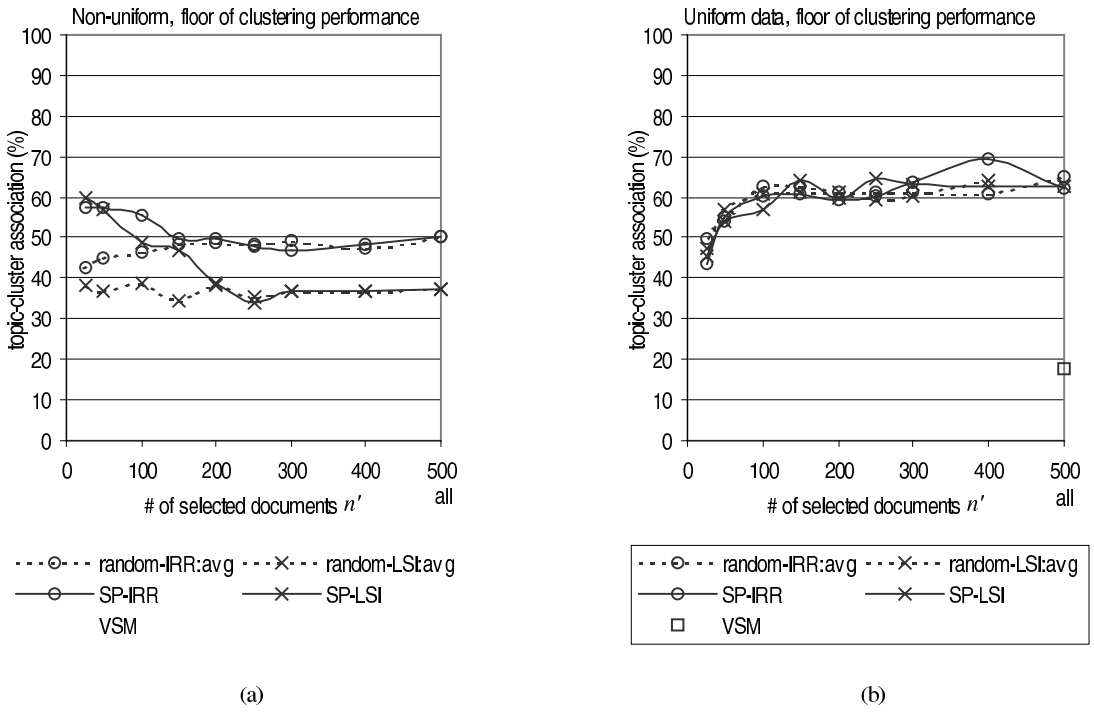

Figure 5.8: Floor clustering performance results on (a) non-uniform distribution and (b) uniform distribution of 500 documents over ten topics. The results in (b) excludes single-link.

the randomized methods, we did five runs for each document set (i.e., 25 runs in total). The scaling factor for IRR was determined by AUTO-SCALE as in Section 5.4.1. The term-document matrices were created in the same manner as in Section 4.3.1. We report the performance results for $r = n'/n = 0.05, 0.5, 1$.

In addition to the ceiling and floor clustering performance results over all the clustering methods, we plot the floor performance results excluding single-link by the dotted line because it performed significantly worse than the other clustering methods in all settings.

5.5.2 Kappa average precision

As in the evaluation of IRR (Section 4.4), we consider two ways to choose the dimensionality of a document subspace. In one case we assume that the system knows the number k of topics and set the dimensionality to k. In the other case, we train the dimensionality parameter, using the residual ratio method from Section 4.1.2.

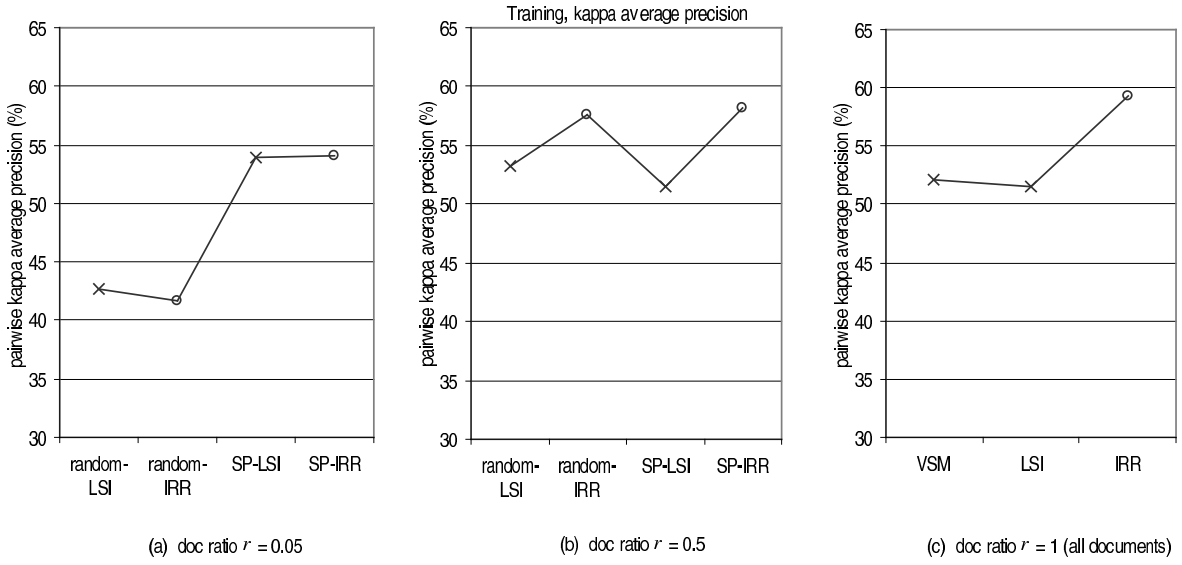

Figure 5.9: Kappa average precision when the dimensionalities are trained.

When the dimensionality parameter is trained, we observe the same trend as in the controlled distribution experiments. When r is very small ($r = 0.05$; Figure 5.9 (a)), SP-based methods perform better than the randomized methods largely by up to 12.3%. For large values of r (Figure 5.9 (b) and (c)), IRR-based methods outperform LSI-based methods by up to 7.8%. As a result, SP-IRR shows the overall advantage over the other three methods (and VSM).

Setting the dimensionality to the number of topics k, the trend of the differences across methods is the same as in the setting where the dimensionality parameter is trained. However, we see that for small r ($r = 0.05$; Figure 5.10 (a)), the kappa average precision results on the SP-based methods degrade quite a bit, although they remain better than the randomized methods. For larger r (Figure 5.10 (b) and (c)) the performance of each method is equivalent to or better than when the dimensionality parameter is trained.

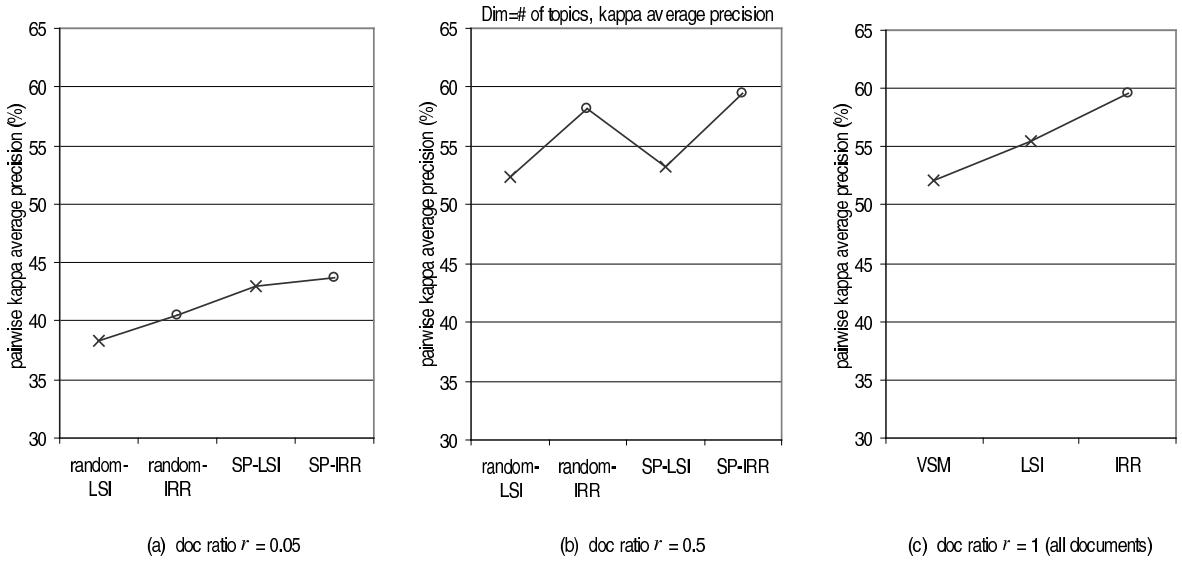

Figure 5.10: Kappa average precision results where the dimensionalities are set to the number of topics.

5.5.3 Clustering results

To derive clustering performance results, we need to specify the number of clusters, in addition to the dimensionality of the subspace. We set the number of clusters to the number of topics k, since, in this setting which simulates focused collections, it is natural to assume that the system approximately knows k.

As shown in Figure 5.11 and 5.12, the ceiling clustering results show trends quite similar to the kappa average precision results. Whether the dimensionality is set to k or trained, IRR-based methods outperform the randomized methods for small r, and SP-based methods do better than LSI-based methods for larger r. Setting the dimensionality to k degrades performance when r is small, while it is essentially as good as training if r is large.

The floor clustering performance results show similar trends, although it is less clear, presumably because of the idiosyncrasies of the clustering methods. But at any rate, the SP-IRR's floors are not substantially worse than others.

5.5.4 Discussion

Figure 5.14, 5.15, and 5.16 list the performance improvement with respect to VSM in all the settings of experiments on unrestricted distribution data. We see that SP-IRR and random-IRR generally outperform their LSI versions. For large r ($r = 0.5$), SP-IRR and random-IRR rival IRR. SP-IRR is more robust than random-IRR for small r ($r = 0.05$), if the dimensionality is appropriately chosen by training.

Using SP-IRR or random-IRR, the computation time is reduced largely (as shown in Figure 5.13) with very small degradation of the quality.

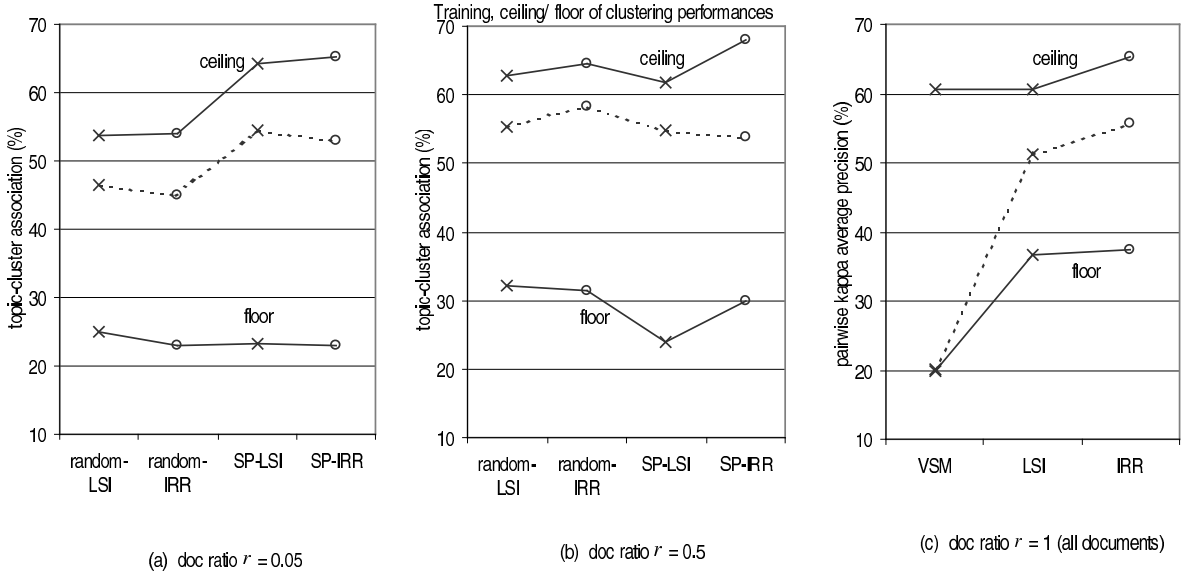

Figure 5.11: Ceiling and floor clustering performance results when the dimensionality is trained. The dotted lines are the floor excluding single-link.

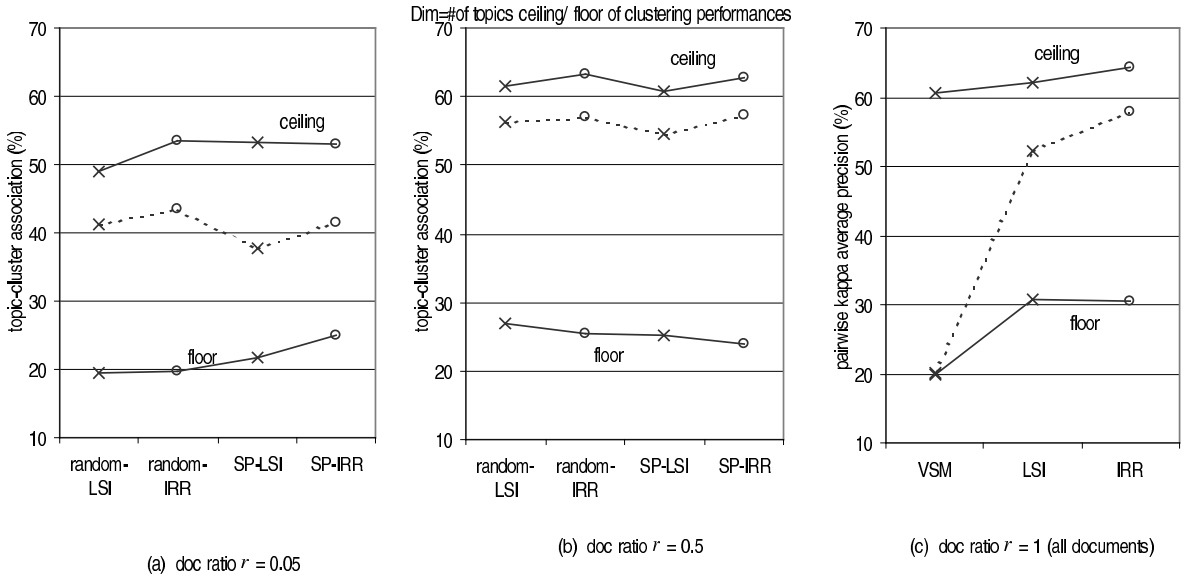

Figure 5.12: Ceiling and floor clustering performance results where the dimensionalities are set to the number of topics. The dotted lines are the floor excluding single-link.

algorithm	CPU sec	k.a.p. (%)
IRR	110.54	0
SP-IRR ($r = 0.50$)	38.36	-0.8%
random-IRR ($r = 0.50$)	20.57	-1.7%

Figure 5.13: CPU time in seconds and kappa average precision with respect to IRR when the dimensionalities are trained. Averages over five runs are shown.

Interestingly, the number of topics k has turned out to be a bad choice for the dimensionality when r is very small, while for large r it is as good as training. When r is very small, naturally the amount of information encoded into one dimension becomes small; therefore, SP-IRR subspaces need more dimensions to be 'good'. In contrast, random-IRR with very small r is not improved much even if we add more dimensions (by training), because the source of information is fixed to nr documents at the beginning. If something is missing (e.g., less dominant topics wiped out), it is never recovered, as random-IRR does not reselect documents.

We note that the performance trend on unrestricted distribution data is somewhere between the trend on the uniform data and the highly non-uniform data in the previous section. As the unrestricted distribution data we used simulates a practical setting, the distributions of documents over topics are fairly (but not extremely) non-uniform. This indicates that the results in this section are consistent with those on the controlled distribution data, and all together validate our prediction that the performance can be improved by counteracting the non-uniformity of document distributions over the underlying topics.

rank	SP or random?	IRR or LSI?	r	dim	k.a.p (%)
1	–	IRR	1.0	k	+7.4
1	SP	IRR	0.5	k	+7.4
3	–	IRR	1.0	trained	+7.2
4	SP	IRR	0.5	trained	+6.0
4	random	IRR	0.5	k	+6.0
6	random	IRR	0.5	trained	+5.5
7	–	LSI	1.0	k	+3.4
8	SP	IRR	0.05	trained	+1.9
9	SP	LSI	0.05	trained	+1.8
10	random	LSI	0.5	trained	+1.1
10	SP	LSI	0.5	k	+1.1
12	random	LSI	0.5	k	+0.2
13	–	VSM	–	–	0
14	SP	LSI	0.5	trained	-0.6
14	–	LSI	1.0	trained	-0.6
16	SP	IRR	0.05	k	-8.6
17	SP	LSI	0.05	k	-9.1
18	random	LSI	0.05	trained	-9.4
19	random	IRR	0.05	trained	-10.4
20	random	IRR	0.05	k	-11.6
21	random	LSI	0.05	k	-13.8

Figure 5.14: Kappa average precision results with respect to VSM in the descending order.

rank	SP or random?	IRR or LSI?	r	dim	ceiling (%)
1	SP	IRR	0.5	trained	+7.4
2	–	IRR	1.0	trained	+4.6
2	SP	IRR	0.05	trained	+4.6
4	random	IRR	0.5	trained	+4.0
5	–	IRR	1.0	k	+3.8
6	SP	LSI	0.05	trained	+3.7
7	random	IRR	0.5	k	+2.7
8	SP	IRR	0.5	k	+2.1
9	–	LSI	1.0	k	+1.4
10	random	LSI	0.5	trained	+2.1
11	SP	LSI	0.5	trained	+1.2
12	random	LSI	0.5	k	+0.8
13	–	LSI	1.0	trained	+0.1
13	SP	LSI	0.5	k	+0.1
15	–	VSM	–	–	0
16	random	IRR	0.05	trained	-6.5
17	random	LSI	0.05	trained	-6.9
18	random	IRR	0.05	k	-7.1
19	SP	LSI	0.05	k	-7.2
20	SP	IRR	0.05	k	-7.6
21	random	LSI	0.05	k	-11.7

Figure 5.15: Ceiling clustering performance results with respect to VSM in the descending order.

rank	SP or random?	IRR or LSI?	r	dim	floor (%)
1	random	IRR	0.5	trained	+38.6
2	–	IRR	1.0	k	+38.2
3	SP	IRR	0.5	k	+37.1
4	random	IRR	0.5	k	+36.8
5	random	LSI	0.5	k	+36.1
6	–	IRR	1.0	trained	+35.4
7	random	LSI	0.5	trained	+35.0
8	SP	LSI	0.5	trained	+34.7
9	SP	LSI	0.05	trained	+34.3
9	SP	LSI	0.5	k	+34.3
11	SP	IRR	0.5	trained	+33.5
12	SP	IRR	0.05	trained	+32.8
13	–	LSI	1.0	k	+31.9
14	–	LSI	1.0	trained	+31.0
15	random	LSI	0.05	trained	+26.3
16	random	IRR	0.05	trained	+24.8
17	random	IRR	0.05	k	+23.4
18	SP	IRR	0.05	k	+21.4
19	random	LSI	0.05	k	+21.1
20	SP	LSI	0.05	k	+17.5
21	–	VSM	–	–	0

Figure 5.16: Floor clustering performance results excluding single-link with respect to VSM.

Chapter 6

Other ways to compensate for non-uniformity

In the previous chapters, we proposed IRR and SP-IRR, which seek to outperform LSI by compensating for the non-uniformity of topic-document distributions. This chapter introduces other ways to compensate for non-uniformity. The methods discussed here may not be as useful as IRR or SP-IRR, but answer some interesting questions.

6.1 Pseudo-document SVD

We introduce another way to 'smooth out' non-uniform distributions of documents over topics, which we call *pseudo-document SVD (*PDSVD). The definitions of the basis vectors of the LSI and IRR subspaces in Section 4.1 can be rewritten by:

$$\begin{aligned}
\text{LSI:}\ \mathbf{u}_i &= \arg\max_{||\mathbf{x}||_2=1} \sum_{j=1}^{n} ||\mathbf{r}_j^{(i)}||_2^2\ \cos^2(\mathbf{x}, \mathbf{r}_j^{(i)})\,, \\
\text{IRR:}\ \mathbf{u}_i &= \arg\max_{||\mathbf{x}||_2=1} \sum_{j=1}^{n} ||\mathbf{r}_j^{(i)}||_2^s\ \cos^2(\mathbf{x}, \mathbf{r}_j^{(i)}) \qquad \text{where } s = 2q+2\,,
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{R}^{(1)} &= \widetilde{\mathbf{A}}\,, \\
\mathbf{R}^{(i+1)} &= \mathbf{R}^{(i)} - \text{proj}(\mathbf{R}^{(i)}, \{\mathbf{u}_i\})\,.
\end{aligned}$$

One natural question is, what if we parameterize the order of the power of the cosine, which is always 2 for both LSI and IRR:

$$\hat{\mathbf{u}}_i = \arg\max_{||\mathbf{x}||_2=1} \sum_{j=1}^{n} ||\mathbf{r}_j^{(i)}||_2^s\ \cos^p(\mathbf{x}, \mathbf{r}_j^{(i)}) \quad ? \tag{6.1}$$

Through our preliminary experiments, it has turned out that in practice the effect of changing the factor s is negligible when $p \geq 3$ because the shrinking/enlarging effect of $\cos^p(\mathbf{x}, \mathbf{r}_j^{(i)})$ is typically far larger than that of $||\mathbf{r}_j^{(i)}||_2^s$. We simply set $s = p$, thereby replacing equation (6.1) by

$$\hat{\mathbf{u}}_i \;=\; \arg\max_{||\mathbf{x}||_2=1} \sum_{j=1}^{n} \left(\mathbf{x}^T \mathbf{r}_j^{(i)}\right)^p . \tag{6.2}$$

As it is easy to show that when $p = 1$, $\hat{\mathbf{u}}_i$ for $i > 1$ becomes non-unique[1], which is not good for our purposes, we focus on settings where $p \geq 3$. A larger p sharpens the distance difference from $\mathbf{x}$ (as measured by the cosine) among the residual vectors, and forces the creation of a local 'representative' of a smaller neighborhood by shrinking farther vectors more, as illustrated in Figure 6.1.

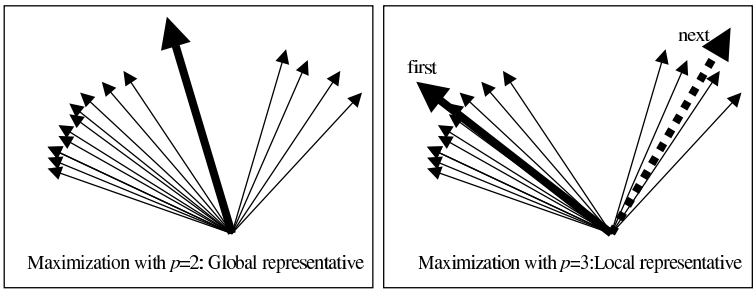

Figure 6.1: The effect of changing p. With $p = 3$, *local* representative is found while $p = 2$ finds more *global* representative.

For $p \geq 3$, there is no easy way to find a global maximum, but we can derive a property which when satisfied yield the critical point.

Theorem 6.1.1 *Let p be a positive integer, and let*

$$f(\mathbf{x}) = \sum_{j=1}^{n} \left(\frac{\mathbf{x}^T}{||\mathbf{x}||_2} \mathbf{r}_j\right)^p .$$

[1]We have $\hat{\mathbf{u}}_1 = \sum_{i=1}^n \tilde{\mathbf{a}}_i / ||\sum_{i=1}^n \tilde{\mathbf{a}}_i||_2$. For any $\mathbf{x}$, we have

$$\begin{aligned}
\sum_{j=1}^{n} \mathbf{x}^T \mathbf{r}_j^{(2)} &= \sum_{j=1}^{n} \mathbf{x}^T (\tilde{\mathbf{a}}_j - \hat{\mathbf{u}}_1 \hat{\mathbf{u}}_1^T \tilde{\mathbf{a}}_j) \\
&= \mathbf{x}^T \sum_{j=1}^{n} \tilde{\mathbf{a}}_j - (\mathbf{x}^T \sum_{i=1}^{n} \tilde{\mathbf{a}}_i / ||\sum_{i=1}^{n} \tilde{\mathbf{a}}_i||_2)(\sum_{i=1}^{n} \tilde{\mathbf{a}}_i / ||\sum_{i=1}^{n} \tilde{\mathbf{a}}_i||_2)^T \sum_{j=1}^{n} \tilde{\mathbf{a}}_j \\
&= \mathbf{x}^T \sum_{j=1}^{n} \tilde{\mathbf{a}}_j - (\sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{\mathbf{a}}_i^T \tilde{\mathbf{a}}_j / ||\sum_{i=1}^{n} \tilde{\mathbf{a}}_i||_2^2) \mathbf{x}^T \sum_{i=1}^{n} \tilde{\mathbf{a}}_i = \mathbf{0}
\end{aligned}$$

Therefore, $\arg\max_{||\mathbf{x}||_2=1} \sum_{j=1}^n \left(\mathbf{x}^T \mathbf{r}_j^{(2)}\right)$ is not unique.

Suppose that for some λ,

$$\lambda \mathbf{y} = \sum_{j=1}^{n} (\mathbf{y}^T \mathbf{r}_j)^{p-1} \mathbf{r}_j \,.$$

Then, $f(\mathbf{y})$ is a critical point, i.e., either a local maximum, a local minimum, or a saddle point.

This theorem (see Appendix B.8 for the proof) leads us to the iteration in Figure 6.2, which becomes the Power iteration for eigenvector computation if $p = 2$. Clearly, it follows from Theorem 6.1.1 that $\hat{\mathbf{u}}_i$ thus computed produces an extreme value of our objective function in (6.2), if it converges, and empirically it almost always converges to a local maximum in our setting. We initially set $\hat{\mathbf{u}}_i$ to the most promising residual $\arg\max_{\exists h.\mathbf{x}=\mathbf{r}_h^{(i)}} \sum_{j=1}^{n} \left(\mathbf{x}^T \mathbf{r}_j^{(i)}\right)^p$ expecting that it converges to a larger local maximum.

Repeat
  $\hat{\mathbf{u}}_i := \sum_{j=1}^{n} (\hat{\mathbf{u}}_i^T \mathbf{r}_j^{(i)})^{p-1} \mathbf{r}_j^{(i)}$
  $\hat{\mathbf{u}}_i := \hat{\mathbf{u}}_i / ||\hat{\mathbf{u}}_i||_2$
until convergence

Figure 6.2: Iteration to find the vector which produces an extreme value of $\sum_{j=1}^{n} (\mathbf{x}^T \mathbf{r}_j^{(i)})^p$ subject to $||\mathbf{x}||_2 = 1$.

The $\hat{\mathbf{u}}_i$ thus computed satisfies

$$\lambda \hat{\mathbf{u}}_i = \sum_{j=1}^{n} \left(\hat{\mathbf{u}}_i^T \mathbf{r}_j^{(i)}\right)^{p-1} \mathbf{r}_j^{(i)} \,. \tag{6.3}$$

for some λ. Instead of letting the $\hat{\mathbf{u}}_i$s span the subspace (for reasons explained later in Section 6.3), we define a *pseudo-document vector* $\mathbf{p}_i$ for each i by

$$\mathbf{p}_i = \sum_{j=1}^{n} \left(\hat{\mathbf{u}}_i^T \mathbf{r}_j^{(i)}\right)^{p-1} \widetilde{\mathbf{a}}_j \,. \tag{6.4}$$

Observe that (6.4) is a weighted average of the initial document vectors $\widetilde{\mathbf{a}}_j$s rather than of the residual vectors $\mathbf{r}_j^{(i)}$s as in (6.3). Consequently, each of the pseudo document vectors $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{n''}$ represents each of n'' groups of somewhat clustered documents.

We let the left singular vectors of the *pseudo document matrix*

$$\left[\frac{\mathbf{p}_1}{||\mathbf{p}_1||_2} \quad \frac{\mathbf{p}_2}{||\mathbf{p}_2||_2} \quad \cdots \quad \frac{\mathbf{p}_{n''}}{||\mathbf{p}_{n''}||_2} \right] \tag{6.5}$$

span the subspace, which we call the PDSVD subspace. (We call this computation pseudo-document SVD.) Crucially, length normalization of $\mathbf{p}_i$ in (6.5) gives the same weight to each group of clustered documents, no matter how dominant the group is. Thus, we expect that

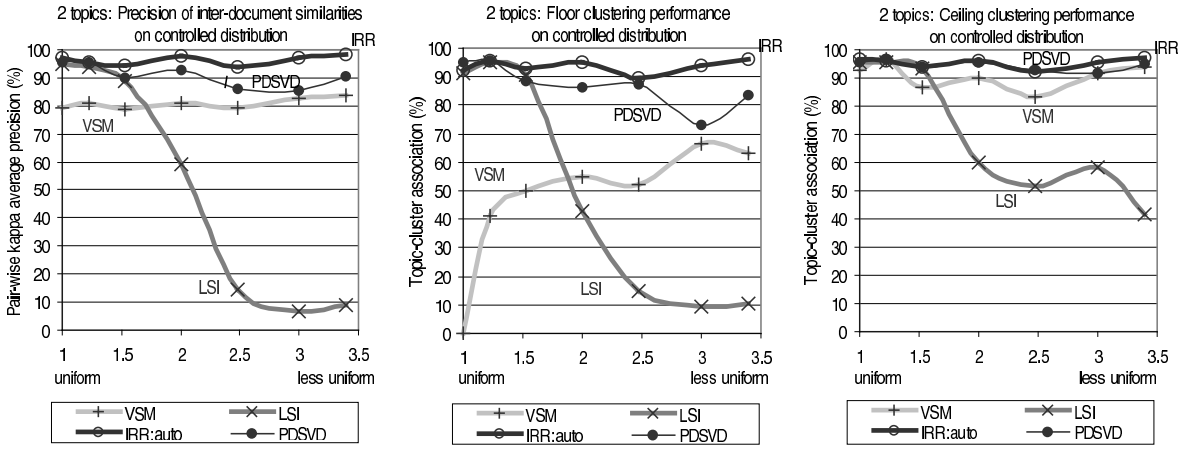

Figure 6.3: PDSVD kappa average precision and clustering performance, controlled distributions over two topics. Points are averages over ten document sets.

PDSVD should *smooth the non-uniformity of topic-document distributions.*

6.2 PDSVD experiments

We evaluated PDSVD in exactly the same settings as the evaluation of IRR in Section 4.3 (controlled-distribution data) and Section 4.4 (unrestricted-distribution data).

The number n'' of pseudo document vectors was determined by

$$n'' = n\left(\gamma \cdot f(\widetilde{\mathbf{A}}) + \delta\right),$$

where

$$f(\widetilde{\mathbf{A}}) = \left(\frac{||\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}||_F}{n}\right)^2,$$

from the AUTO-SCALE method in Section 4.1.1, and γ and δ are experimentally determined once and for all from observations on data disjoint from our test sets. We set $p = 3$ experimentally.

The performance results of PDSVD on controlled distributions (Figure 6.3) are slightly worse than IRR with AUTO-SCALE, but still apparently compensate for non-uniformity. The results on the unrestricted data are also close to those of IRR; see Figure 6.4 and 6.5 and the mean differences between PDSVD and IRR in Figure 6.6. PDSVD rivals IRR, but it does not quite outperform IRR. The performance difference between IRR and PDSVD is not statistically significant in any setting of the unrestricted-data experiments; see Figure 6.6.

6.3 Discussion

The performance of PDSVD on controlled distribution data indicates that PDSVD compensates for the non-uniformity of topic-document distributions, as we expected. The results on unre-

dimensionality?	LSI	PDSVD
number of topics	-8.7	2.75
trained	0	4.00

Figure 6.4: PDSVD absolute improvement in pair-wise kappa average precision over VSM (51.4%), unrestricted distributions: averages over 30 runs.

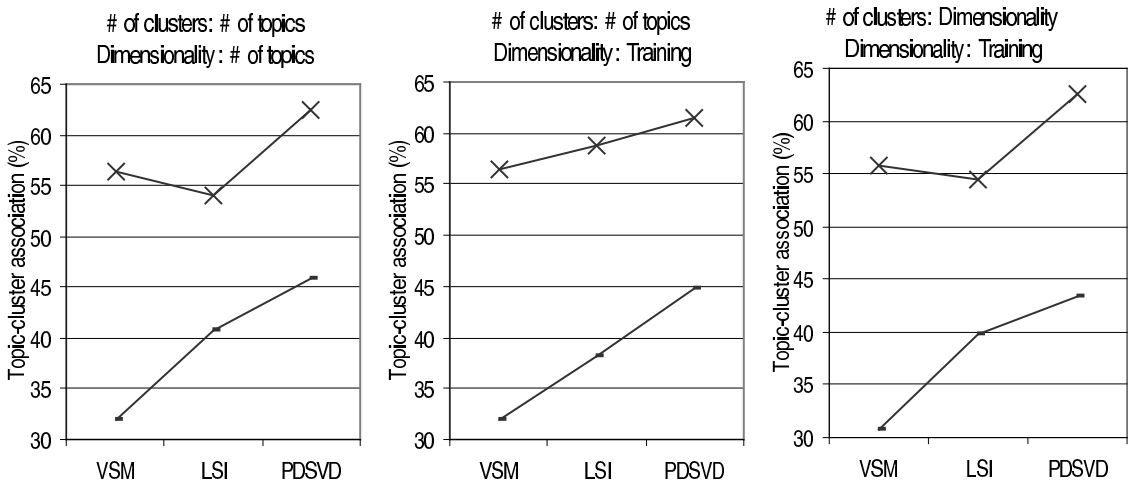

Figure 6.5: PDSVD document clustering performances, unrestricted distributions: averages over 30 runs.

stricted distribution data show that PDSVD generally outperforms LSI and rivals IRR. However, the disadvantage of PDSVD is that its computation is almost twice more expensive than IRR and LSI, as it computes $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n''}$ (and $\mathbf{p}_1, \ldots, \mathbf{p}_{n''}$) first, and it computes the SVD next.

If we let $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n''}$ span the subspace, it is as fast as IRR. Since a larger p amplifies the distance difference among residuals, non-uniformity may be compensated for if we can choose the initial vector (for the local maximization in Figure 6.2) from less dominant topics, and if the residuals that are far from it and on dominant topics, are ignored. We conducted experiments to evaluate the subspace spanned by $\hat{\mathbf{u}}_i$'s for various settings of the initial vector, and observed that when $p = 3$ (which turned out better than $p = 4, 5$), the performance, at best, tended to be either largely improved or largely degraded with respect to VSM and LSI, and on average worse than IRR.

We examined each of $\hat{\mathbf{u}}_i$'s and found that in some sense $p \geq 3$ is 'too large'. After computing the first several $\hat{\mathbf{u}}_i$'s, the cosine values among residuals are very small (i.e., very far from each other), and therefore, when we set $p \geq 3$, the distance differences are over-amplified (compared with length differences), and the iteration for (sub-)optimization (Figure 6.2) often converges to the vector very close to the initial vector. That is, roughly speaking, by setting $p = 3$, after several large topics are captured into the subspace, documents somewhat different from others (not necessarily on less dominant topics but sometimes outliers on dominant topics) are captured one by one. It is not clear that the non-uniformity is compensated for in this process. Setting $p = 3$ causes a drastic change from $p = 2$, and note that we cannot set, for instance, $p = 2.5$

dim	trained	k	trained		k		trained	
# of clusters	N/A		k		trained		dimensionality	
metric	Kappa	avg prec	ceiling	floor	ceiling	floor	ceiling	floor
mean diff (%)	+0.33	+0.70	+0.70	-1.37	+1.91	-1.12	+0.54	+3.68
p-value	0.809	0.524	0.633	0.237	0.198	0.555	0.701	0.093

Figure 6.6: Comparison of IRR and PDSVD performance results on unrestricted data by the paired t-test. All the groups (each consisting of performance results from 30 runs) passed the Gaussian normality test. The difference of mean ("mean diff") is PDSVD's mean with respect to that of IRR. No statistically significant difference ($p < 0.05$) was found.

because the cosine value raised by p may be negative. As p has to be an integer, fine adjustment is not possible.

In a sense, the SVD (i.e., $p = 2$) in the second step of PDSVD 'mitigates' the drastic effect of $p = 3$ in the first step, resulting in performance rivaling IRR.

We observe that IRR has advantages over the two methods we have discussed in this chapter. First, as mentioned above, IRR is computationally less expensive than PDSVD. Second, unlike p, the scaling factor of IRR has a clear relation to the degree of the compensation for the non-uniformity. Finally, the scaling factor does not have to be an integer; therefore, fine adjustment of the degree of compensation is possible.

Chapter 7

An Application of IRR: Multi-document Summarization

This chapter introduces an example application of IRR, multi-document summarization. Our starting point is the assumption that a tightly connected (and therefore intuitively interpretable) set of coherent texts would act as a 'prompting' device when appropriately presented to the user. We show that the IRR semantic space works as a framework to find such coherent texts and effectively present them to the user.

7.1 Multi-document summarization as an enabling technology for IR

The rapid growth of electronic documents has created a great demand for a navigation tool for traversing a large corpus. Information retrieval technologies allow us to access documents presumably matching our interests. However, a traditional hit list-based architecture, which returns linearly organized single document summaries, no longer suffices, given the size of a typical hit list (e.g., queries like "natural language summarization" or "automatic summarization" against the AltaVista search engine (`http://www.altavista.com`) return over 100,000 hits each.)

To allow a more comprehensive and screen space-efficient presentation of query results, we propose a technology for summarizing collections of multiple documents. In our work, we focus on identifying themes representative of a document and possibly running across documents. Even if we are unable to 'embody' a theme in coherently generated prose, we start with the assumption that a mapping exists between a theme and a tightly connected (and therefore intuitively interpretable) set of coherent linguistic objects, which would act as a 'prompting' device when presented to the user in an appropriate context. We refer to such themes as *topics*.

Our view of multi-document summarization combines three premises: coherent topics can be identified reliably; highly representative topics, running across subsets of the document collection, can function as multi-document summary surrogates; and effective end-use of such topics can be facilitated by a visualization environment which clarifies the relationship between topics and documents. This work specifically addresses the following considerations.

• **Multiple general topics** We regard the ability to respond to multiple topics in a document collection — in contrast to a prevailing trend in multi-document summarization that seeks to present a single, possibly pre-determined, topic (see below) — to be crucial to applications such as summarization of query results. In this work we choose not to narrow the topic detection process by the given query, since in IR it is a well-known concern that user-specified queries do not necessarily convey the user's real interests thoroughly. Thus, we need to deal with multiple general topics.

• **Textual and graphical presentation** Since our multi-document summaries will, by definition, incorporate multiple topics, the question arises of optimal representation of the relationships among the topics, the linguistic objects comprising each topic, and the documents associated with (possibly more than one) topic. In particular, for IR, we want to show the relationships between topics and documents so that a user can access documents in the context of the topics. A topic by itself can clearly be represented largely by a set of text objects. However, we need also to present arbitrary number of such topics as part of the same summary. We believe that, for adequate representation of the resulting many-to-many relationships (which is crucial for the end-user fully understanding the summary), additional graphical components are needed in the interface.

To our knowledge, existing studies of multi-document summarization do not place emphasis on these considerations. Radev and McKeown (1998) have shown a methodology for 'briefing' news articles reporting the same event. Barzilay et al. (1999) have proposed a method for summarizing "news articles presenting different descriptions of the same event". Both these studies focus on a single topic in a document collection. Mani and Bloedorn (1999) have addressed summarizing of similarities and differences among related documents with respect to a specified query or profile. In their study, several presentation strategies are suggested. Although they mention a graphical strategy, such as plotting documents sharing more terms closer together, no implementation is reported.

There are a number of different studies that address graphical presentation of multi-document (or document corpus) visualization: these include the VIBE system (Olsen et al., 1993; Korfhage and Olsen, 1995), Galaxy (Rennison, 1994), SPIRE Themescapes (Wise et al., 1995), LyberWorld (Hemmje et al., 1994), and applications of self-organizing map utilizing neural network technique (Kohonen, 1997; Lin, 1993; Lagus et al., 1996). In general, these studies consider documents as objects in a typically high-dimensional model space and provide 2-D or 3-D representation of this space. Their focus is on detecting and presenting structural relationships among documents in a corpus.

From our viewpoint, these two fields of research address two different perspectives on the multi-document analysis problem: multi-document summarization efforts largely deliver their results in textual form, while document corpus visualization research, which focuses on means for graphical representation of a document space, does not (explicitly) perform any summarization work. While we believe that both textual and graphical representations are essential in the context of IR, the technologies from the two fields, in general, cannot be easily combined because of methodological differences (such as differences in modeling the document set, calculating similarity measures, and choosing linguistic objects in terms of which a summary would be con-

structed).

Motivated by these observations, we propose one uniform framework that provides both textual and graphical representations of a document collection. In this framework, topics underlying a document collection are identified and described by means of linguistic objects in the collection. Relationships, typically many-to-many, among documents and topics are graphically presented, together with the topic descriptions, by means of a graphical user interface specifically designed for this purpose. We focus on relatively small document collections (e.g., 100 or so top-ranked documents), observing that in a realistic environment users will not look much beyond such a cut-off point. Our approach maps linguistic objects onto a multi-dimensional space (called *semantic space*). As we will see below, the mapping is defined in a way that allows for topics with certain properties to be derived and for linguistic objects at any granularity to be compared as semantic concepts.

7.2 Semantic space

A semantic space is derived on the basis of analyzing relationships among linguistic objects — such as terms, sentences, and documents — in an entire multi-document collection. A *term* can be simply a 'content word', in the traditional IR sense, or it can also be construed as a phrasal unit, further representative of a concept in the document domain. In our implementation, we do, in fact, take that broader definition of terms, to incorporate all types of non-stop lexical items as well as phrasal units such as named entities, technical terminology, and other multi-word constructions (see Section 7.4 below).

We map linguistic objects (such as terms, sentences, and documents) to vectors in a multi-dimensional space. We construct this space so that vectors for objects behaving statistically similarly (and therefore presumed to be semantically similar) point in similar directions. The vectors are called *document vectors*, *sentence vectors*, and *term vectors*, according to the original linguistic objects they are derived from; however, all vectors hold the same status in the sense that they represent some concepts. In this work, we call this multi-dimensional space *semantic space*, and implement it by the IRR algorithm in Chapter 4. In essence, in our semantic space terms related to each other are mapped to vectors having similar directions, while a traditional vector space model treats all terms as independent from each other.

To detect topics underlying the document collection, we create a set of vectors in the semantic space so that every document vector is represented by (or close to) at least one vector (called *topic vector*). In other words, we provide viewpoints in the semantic space so that every document can be viewed somewhat closely from some viewpoint. Given such vector representations for topics, we can quantitatively measure the degree of associations between topics and linguistic objects by using a standard cosine similarity measure between topic vectors and linguistic object vectors. The linguistic objects with the strongest association would represent the topic most appropriately.

We adapt the IRR algorithm because it achieves high precision in similarity measurement among *all* the documents by capturing information more *evenly from every document*. This algorithm fits well in our framework since we want to find topics by referring the similarities of *all* pairs of

documents (shown later), and also we want to assume that *all documents are equal*. Since our focus is on relatively small document collections, we do not consider SP-IRR.

7.3 Visual presentation of a semantic space: combining text and graphics

In this section, we illustrate how textual and graphical presentation are combined, by showing a summary created from 50 documents (the documents were extracted from the TREC collection as being relevant to the TREC topic "non-proliferation treaty").

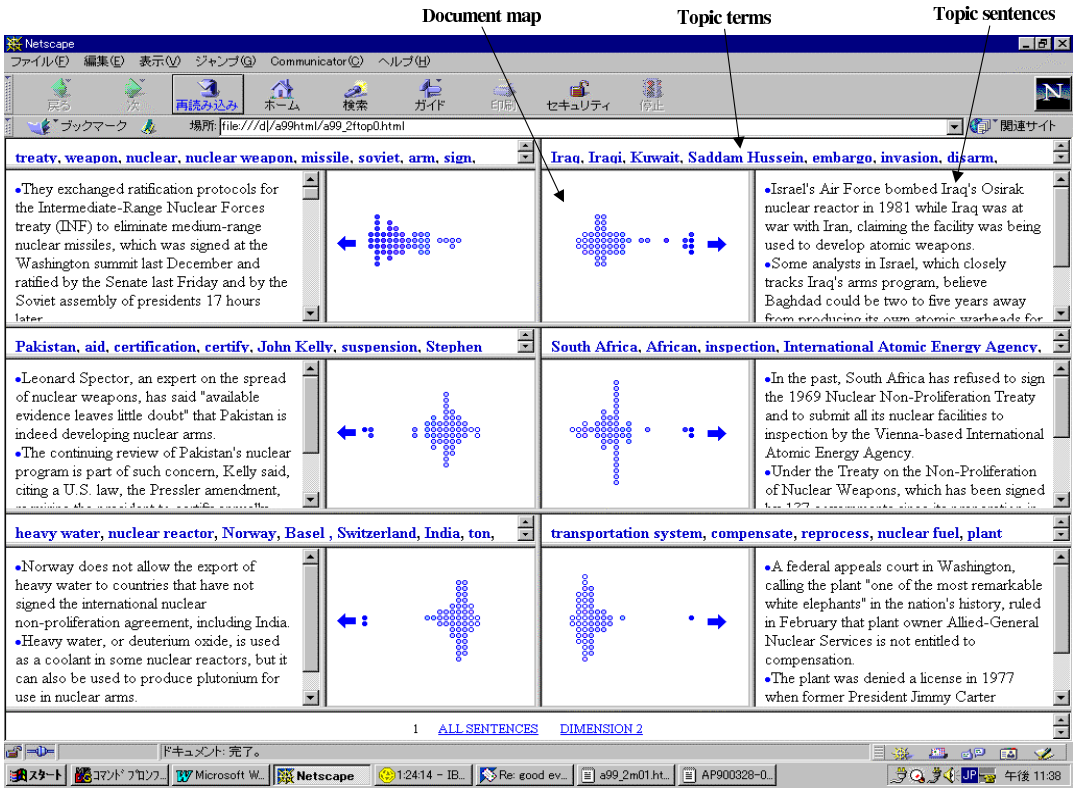

Figure 7.1: Topical summary of a multi-document set: initial screen display.

The summary is presented in one full screen, in relation to the underlying topics. Six topics were detected in this document set; in general, the number of topics detected depends on the document set and on the settings of parameter adjusting the granularity of analysis (Figure 7.1). For each topic, three types of information are presented: a list of terms (*topic terms*), a list of sentences (*topic sentences*), and a visual representation of relevance of each document to the topic (*document map*). The processes of identifying topic terms and topic sentences are described in some detail in Sections 7.4.2 and 7.4.3.

Below we highlight some features of the interface.

• **Topic terms and topic sentences:** The topic presented at the upper right corner of Figure 7.1 has the topic terms *"Iraq"*, *Iraqi"*, *"Kuwait"*, *"Saddam Hussein"*, *"embargo"*, *"invasion"*,

"disarm", and so on. (The frame is scrollable, thus accommodating all topic terms.) A topic typically will be addressed by more than one sentence, presented in a closely associated scrollable frame. The first topic sentence for this topic is *"Israel's Air Force bombed Iraq's Osirak ..."*. Together, the sets of topic terms and sentences describe the topic, i.e., one 'thread' discussed in possibly several documents.

• **Document proxies** — a *dot* represents a document: In a document map, a dot image represents each document (i.e., *document proxy*). A dot before a topic sentence is also a document proxy representing the document containing that sentence.

• **Document maps** — topic-document relevance shown by document proxy placement and color gradation: In a document map, the horizontal placement (closeness to the direction of the arrow) represents the degree of relevance of the corresponding document to the topic. The color intensity of the dot also represents the degree of relevance. For instance, in the document map at the upper right corner of Figure 7.1, we see that there are six documents closely related to this 'IRAQ-TOPIC'. These six dots are placed on the right (the direction of the arrow), and their colors are more intense than the other document proxies. We see one more document to the left to the six documents, also with a relatively strong connection to this topic. Two documents, represented by dots almost at the center of the map, are only somewhat related to this topic. The rest of the documents, having dots that are almost transparent and placed on the left, are not very related to this topic. Thus, users can tell, at a glance, how many documents are related to each topic and how strongly they are related. Note that each document map contains proxies for all the documents. Unlike a typical clustering approach, we do not divide documents into groups. Clusters of documents, if any, are naturally observed in the document map.

• **Highlighting of document proxies** — the relationships between a document and multiple topics: When the mouse rolls over a dot, the title of the document appears, and the color of the dots representing the same document in all the document maps changes (from blue to red) (Figure 7.2). This color change facilitates understanding the relationships between a document and multiple topics.

• **A hot-link from a document proxy to full text:** When a dot is clicked, the full text of the corresponding document is displayed in a separate window. This allows us to browse documents in the context of document-topic relationships.

• **Highlighting a topic sentence in the full text:** When the clicked dot is associated with a topic sentence, the full text is displayed in a separate window, with the topic sentence highlighted. This highlighting helps the user to understand the context of the sentence quickly, and thus further facilitates focusing on the information of particular interest.

• **Topic sentences:** Finally, we illustrate some of the extracted topic sentences below. For each topic, the two sentences most closely related to the topic are shown.

'Iraq-topic':

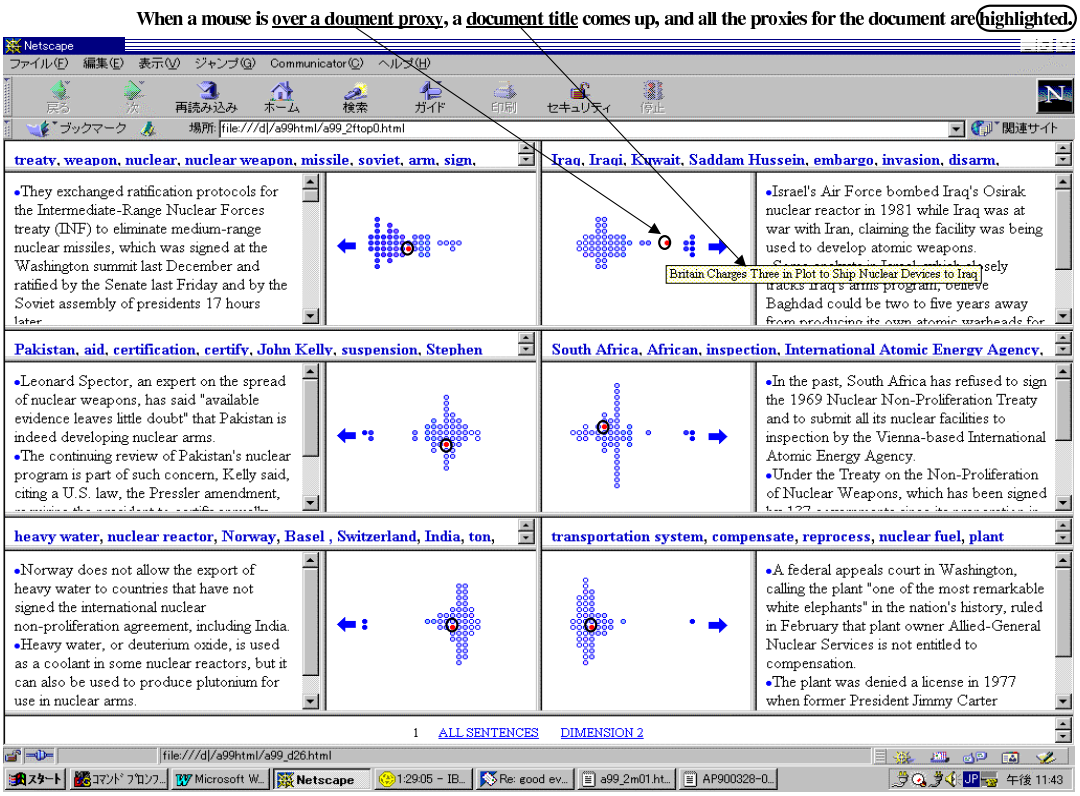

Figure 7.2: Topical summary of a multi-document set: dynamics of the display when the mouse rolls over a document proxy.

- *Israel's Air Force bombed Iraq's Osirak nuclear reactor in 1981 while Iraq was at war with Iran, claiming the facility was being used to develop atomic weapons.*
- *Some analysts in Israel, which closely tracks Iraq's arms program, believe Baghdad could be two to five years away from producing its own atomic warheads for missiles or nuclear bombs to be dropped from jets.*

'Pakistan-topic':

- *Leonard Spector, an expert on the spread of nuclear weapons, has said "available evidence leaves little doubt" that Pakistan is indeed developing nuclear arms.*
- *The continuing review of Pakistan's nuclear program is part of such concern, Kelly said, citing a U.S. law, the Pressler amendment, requiring the president to certify annually that Pakistan does not possess a nuclear weapon.*

'South Africa-topic':

- *In the past, South Africa has refused to sign the 1969 Nuclear Non-Proliferation Treaty and to submit all its nuclear facilities to inspection by the Vienna-based International Atomic Energy Agency.*

- *Under the Treaty on the Non-Proliferation of Nuclear Weapons, which has been signed by 137 governments since its preparation in 1969, countries without such weapons open their nuclear facilities to inspection by experts from the International Atomic Energy Agency, a U.N. agency based in Vienna.*

Both for 'Iraq-' and 'Pakistan-topic', the two topic sentences address two different aspects of the similar 'doubt' or 'concern'. For 'South Africa-topic', the second topic sentence gives background knowledge for the specific fact described in the first topic sentence. We find it interesting that, despite the fact that the two topic sentences are extracted from different documents, they appear to be consecutive sequences from a uniform source.

In essence, the design seeks to facilitate quick appreciation of the contents of a document space by supporting browsing through a document collection with easy switching between different views: topic highlights (terms), topical sentences, full document text, and inter-document relationships. At present, there is no attempt to handle redundancy between topic sentences.

7.4 Mapping a document collection into semantic space

In this section, we outline the process flow of mapping a multi-document collection into semantic space. The process is illustrated schematically in Figure 7.3.

Throughout this section, we use the three small 'documents' shown below as an illustrative example of a data set.

Document #1:	Mary Jones has a little lamb. The lamb is her good buddy.
Document #2:	Mary Jones is a veterinarian for ABC University.
	ABC University has many lambs.
Document #3:	Mike Smith is a programmer for XYZ Corporation.

The data flow from these three documents to the final output is shown in Figure 7.6.

7.4.1 Term extraction and vector creation

As mentioned above, we use the IRR algorithm to create a semantic space. The linguistic objects in a document collection are mapped into vectors in an IRR subspace as follows. First, we extract single- and multi-word stemmed terms, using TALENT (Boguraev and Neff, 2000), and represent documents, sentences, and terms in a collection by the vectors in a VSM space (spanned by terms). We then compute an IRR subspace (i.e., our semantic space) from the term-document matrix (whose columns are length-normalized VSM document vectors), and we project the VSM vectors we created above onto the IRR subspace to obtain the IRR vectors.

For the example mini-documents above, the terms remaining after removal of common stop words are listed at top of Figure 7.6. In our implementation, we use term frequency to fill the entries of the VSM vectors, as shown in the same figure. In this example, we create a 2-dimensional

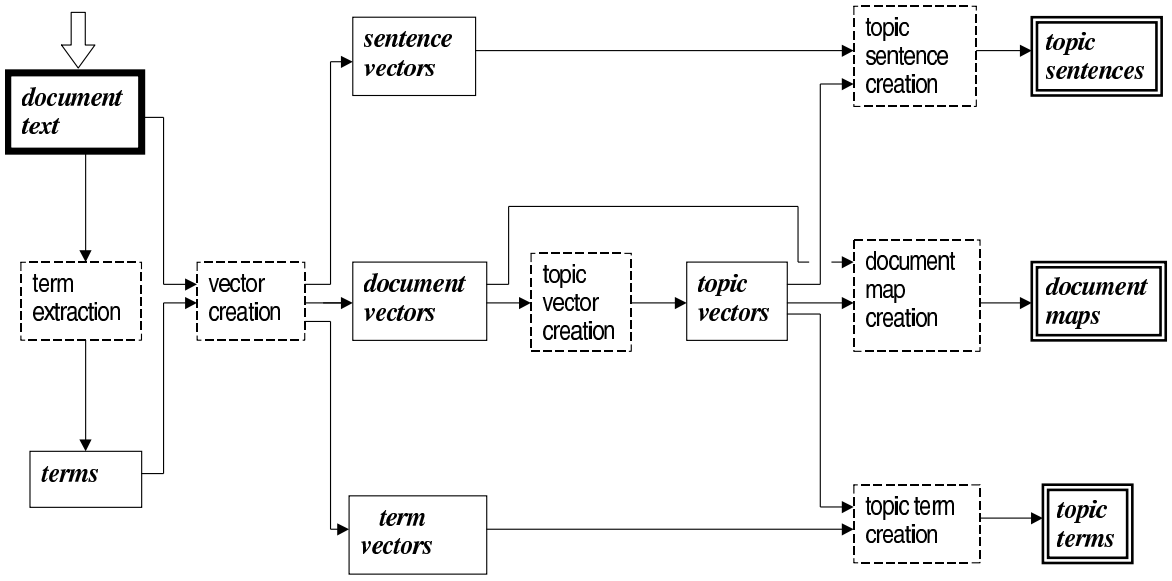

Figure 7.3: Overview of the process. Dashed rectangles are sub-processes. Other rectangles represent data.

IRR subspace; see the two IRR basis vectors in the figure. Note that an IRR basis vector (and any VSM vector) can be regarded as a linear combination of terms. For instance, the first IRR basis vector for our example represents

$$\begin{array}{llllll} 0.45 * & \text{“Mary Jones”} & + & 0.22 * & \text{“little”} & + \\ 0.67 * & \text{“lamb”} & + & 0.22 * & \text{“good buddy”} & + \\ 0.22 * & \text{“veterinarian”} & + & 0.45 * & \text{“ABC University”} & . \end{array}$$

Let the columns of $\mathbf{U}$ be the IRR basis vectors. Then, any VSM vector $\mathbf{x}$ can be turned into an IRR vector (i.e., projected onto the IRR subspace) by computing $\mathbf{U}\mathbf{U}^T\mathbf{x}$ (see Section 2.1.5 for details of the projection operator). Also note that $\mathbf{U}^T\mathbf{x}$ yields a more compact representation of IRR vectors (see Section 2.1.4), which is used in the examples of IRR vectors in Figure 7.6.

7.4.2 Identifying topics

Ultimately, our multi-document summaries rely crucially on identifying topics representing all the documents in the set. This is done by creating topic vectors so that *each document vector is close to (i.e., represented by) at least one topic vector*. We implement this topic vector creation process as follows. First, we create a document graph from the document vectors. In the document graph, each node represents a document vector, and two nodes have an edge between them if and only if the similarity between the two document vectors is above a threshold. Next, we detect the connected components in the document graph, and we create the topic vectors from each connected component by applying the procedure `DetectTopic` (Figure 7.4) recursively.

`DetectTopic` works as follows. The left singular vector $\mathbf{u}$ of a matrix whose columns are the document vectors in a set S is computed. It is a representative direction of the document vectors

```
Procedure DetectTopic(S)
Input: a set of document vectors S
Output: topic vectors

u = the left singular vector of a matrix of
    document vectors in S;
Loop for each document vector d in S
    If similarity between d and u is below a threshold
    Then Begin
        divide S into S1 and S2;
        Call DetectTopic(S1);
        Call DetectTopic(S2);
        Exit the procedure;
    End If
End Loop
Return u as a topic vector;
```

Figure 7.4: Topic vector creation.

```
Step1:  Create a sub-graph Gs for S from the document graph.
Step2:  Compute the minimum cut of all the node pairs in Gs.
Step3:  Evaluate each minimum cut,
          and choose the cut (A, B) maximizing h(A, B) as (S1, S2).
        Function h(A, B) = sqrt(|A| * |B|) * f and f is a cut value of (A, B).
        A cut that divides S more evenly with a smaller cut value
          (i.e., with fewer crossing edges) is chosen.
```

Figure 7.5: Document vector division procedure.

in S. If the similarity between $\mathbf{u}$ and any document vector in S is below a threshold, then S is divided into two sets S_1 and S_2 (see Figure 7.5), and the procedure is called for S_1 and S_2 recursively. Otherwise, $\mathbf{u}$ is returned as a topic vector. The granularity of topic detection can be adjusted by the setting of threshold parameters.

Note that such a topic vector creation procedure essentially detects 'cluster centroids' of document vectors (not sentence vectors), although grouping documents into clusters is not our purpose. This indicates that general vector-based clustering technologies could be integrated into our framework if it brings further improvement.

7.4.3 Associations between topics and linguistic objects

Associations between topics and linguistic objects (documents, sentences, and terms) can be measured by computing the cosine or inner products between the topic vectors and linguistic object vectors. In our implementation, we use the cosine for topic-document associations and the inner product for others. The degree of association between topics and documents is used to create document maps. The terms and sentences with the strongest associations are chosen to be the topic terms and the topic sentences, respectively.

As a result, for our example we get the output shown at the bottom of Figure 7.6.

7.5 Further work

We have proposed a framework for multi-document summarization that leverages graphical elements to present a summary as a 'constellation' of topical highlights. In this framework, we detect topics underlying a given document collection, and we describe the topics by extracting related terms and sentences from the document text. Relationships among topics and documents are graphically presented using gradation of color and placement of image objects. We have illustrated interactions with our prototype system and present algorithms that formalize our framework. We re-emphasize that the framework presented here derives its strength in equal part from two components: the results of topical analysis of the document collection are displayed by means of a multi-perspective graphical interface specifically designed to highlight this analysis.

We have described one possible interface, focusing on certain visual metaphors for highlighting collection topics. The relationships between the topics in a document collection and the documents themselves can be visualized through a number of different perspectives. It is possible to imagine alternative renderings of topic groups within the semantic space; typically these would be mediated by different granularities of document fragments (sentences, phrases, or terms). We have done some work on determining the effects of analyzing linguistic objects (e.g., sentence- and clause-level phrasal units, and different semantic categories of phrasal types), specifically for the purpose of representing closely related topical documents to the user.

Outside of the scope of this work remain at least two open questions. What kinds of phrases are adequate 'carriers' of topical content? How much would operations over sentences, such as sentence merging or reduction, offer alternative ways of visualizing topical content? Until we have clear answers to these, we cannot argue strongly in favor of any particular interface metaphor.

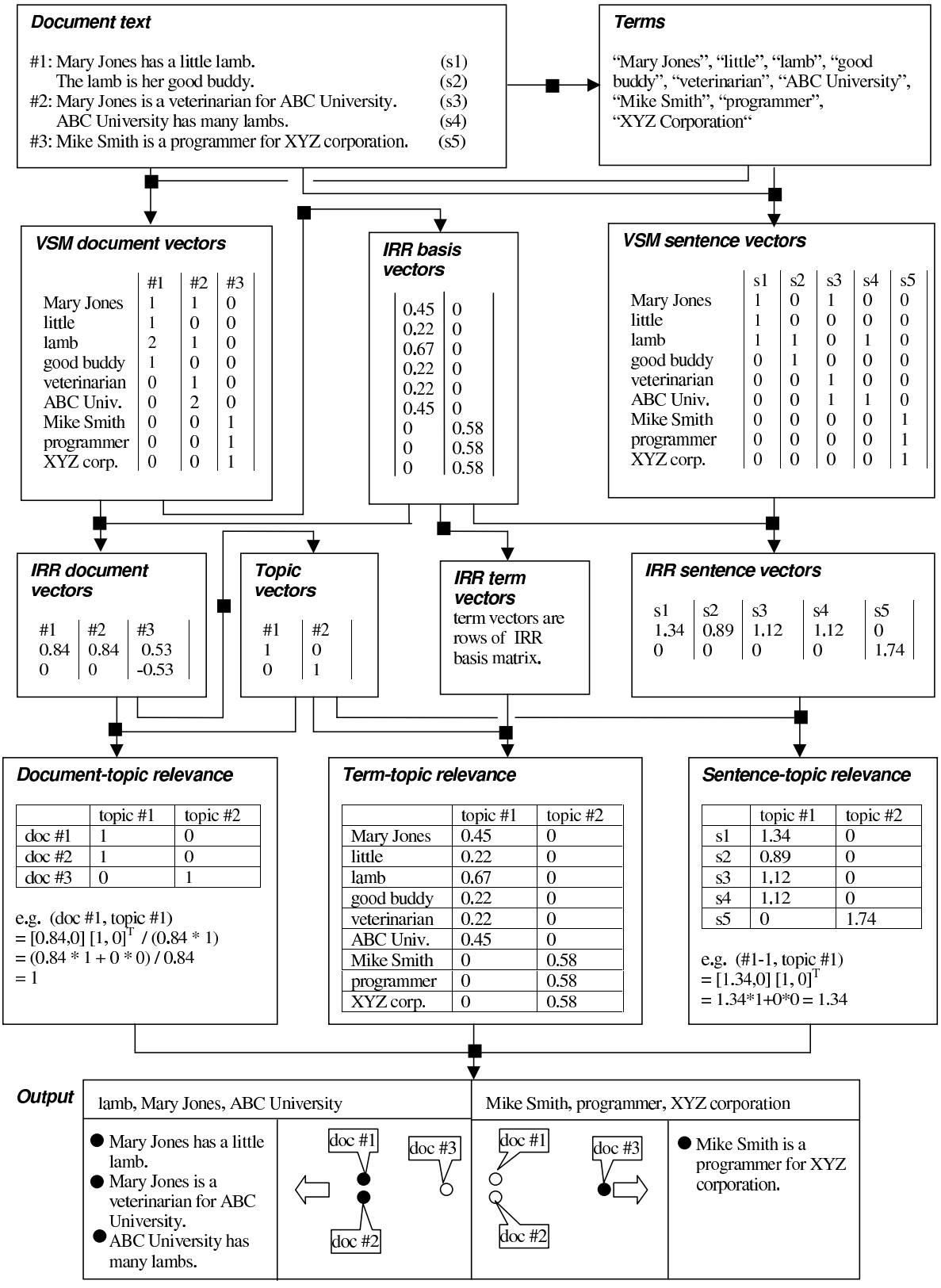

Figure 7.6: Schematic illustration of an example of data flow.

Chapter 8

Related work: Document Representation

This chapter reviews previous studies in relation to our work: general methods related to document representation (Section 8.1) and extensions of LSI (Section 8.2).

8.1 Methods regarding document representation

We summarize previous studies on improving document similarity measurements in relation to LSI. We see that our starting point — the new finding of LSI's performance dependence on the uniformity of underlying topic-document distributions — distinguishes our approach from the previous studies.

8.1.1 Training by prior knowledge of inter-document similarity

We summarize two methods which, unlike IRR, make direct use of prior knowledge of inter-document similarities or concept-document relevance as input: Metric Similarity Modeling (MSM) (Bartell et al., 1995) and an application of Linear Least Squares Fit (LLSF) (Yang and Chute, 1993) to an information retrieval task.

MSM (Bartell et al., 1995) seeks a linear mapping from given initial document vectors to the new document vectors that optimally approximates the given inter-document similarities, called *similarity constraints*. Two kinds of information are required as input: a term-document matrix $\widetilde{\mathbf{A}} \in \Re^{m \times n}$ and $\mathbf{S} \in \Re^{n \times n}$ containing similarity constraints satisfying $\mathbf{S} = \mathbf{X}^T\mathbf{X}$ for some $\mathbf{X}$. MSM has its basis in the analytic (locally-optimal[1]) solution to the problem of finding a linear mapping $\mathbf{W} \in \Re^{h \times m}$ that minimizes $||\mathbf{S} - (\mathbf{W}\widetilde{\mathbf{A}})^T(\mathbf{W}\widetilde{\mathbf{A}})||_{\mathrm{F}}$. After computing such a $\mathbf{W}$ from $\mathbf{S}$ and $\widetilde{\mathbf{A}}$, the term-document matrix $\widetilde{\mathbf{A}}$ and a query vector $\mathbf{q}$ are mapped into new h-dimensional representations, $\mathbf{W}\widetilde{\mathbf{A}}$ and $\mathbf{Wq}$, respectively. The expectation is that the similarity between queries (which were not used for optimization) and documents should be measured more accurately, i.e.,

[1] To be precise, Bartell et al. (1995)'s analytic solution finds a critical point, and they state that empirically it is always a local maximum.

better performance should be in an information retrieval task. We note that the rows of $\mathbf{W}$ are not necessarily orthonormal, i.e., unlike LSI and IRR, MSM does not constrain the linear mapping to orthogonal projection. However, MSM crucially requires pre-defined inter-document similarity measures. This is in contrast to our model in Chapter 3 where true inter-document similarities are hidden from the method.

Yang and Chute (1993) have applied *Linear Least Squares Fit* (LLSF) to information retrieval. LLSF, which finds $\mathbf{W}$ such that

$$||\mathbf{W}\widetilde{\mathbf{A}} - \mathbf{X}||_{\mathrm{F}}$$

is minimized, has been applied to the text categorization task as well (Yang and Chute, 1994; Yang, 1995). For information retrieval, they set $\widetilde{\mathbf{A}}$ to be a term-document matrix, and $\mathbf{X}$ to be a concept-document matrix instantiated by using pre-defined concepts assigned to documents; both are from a training corpus. Fitting by LLSF, $\mathbf{W}$ is expected to capture a mapping from a term-based space to a concept-based space.

These two approaches share the spirit with IRR in seeking vector representations for documents corresponding to human relevance judgments. However, as these methods, unlike IRR and LSI, essentially rely on training (or 'fitting'), we conjecture that the availability of appropriate training data may be a critical issue in practical situations.

8.1.2 Factor analysis

Factor analysis seeks to summarize data in a concise but accurate manner, so that each factor is a generalization of the relationships among the features of the subjects and is quantitatively distinct from any other factor. There are numerous factor analysis methods (see e.g., Kruskal (1978) and Gorsuch (1983)) widely used to process scientific data.

In general, factor analysis seeks to discover 'factors' so that the data can be comprehensively explained by interpreting the factors. From the perspective of our subspace-based approach, a set of 'factors' in factor analysis roughly correspond to an orthonormal basis of a subspace. However, differently from factor analysis, our goal is not on interpreting the basis of the subspace but on using the geometrical closeness (or angle) between the vectors in the subspace as similarity measurement.

Principal Component Analysis

According to Gerbrands (1981), SVD and *principal component analysis (PCA)* were often confused in the literatures on digital image processing. PCA is a factor analysis method whose mathematical formulation is related to SVD. Similarly, the mathematical formulation of LSI is related to SVD; however, LSI (or IRR) is not PCA. The definition of PCA crucially incorporates the deviations from the expectation or mean

$$\mathbf{X} - \mathcal{E}(\mathbf{X})$$

where $\mathbf{X}$ is a matrix of data described by features and $\mathcal{E}(\cdot)$ is the expectation operator. PCA seeks the factors that maximally describe the variance of the features, which results in computing

the SVD of a matrix deriving from $\mathbf{X} - \mathcal{E}(\mathbf{X})$. In contrast, neither LSI nor IRR incorporates the deviation from the mean.

Non-negative matrix factorization

Non-negative matrix factorization (NMF) approximates a matrix $\mathbf{X}$ by

$$\mathbf{X} \approx \mathbf{W}\mathbf{Y}\,,$$

where the elements of $\mathbf{W}$ and $\mathbf{Y}$ are constrained to be non-negative (Lee and Seung, 1999b; Lee and Seung, 1999a). The columns of $\mathbf{W}$ are the factors discovered from $\mathbf{X}$, and $\mathbf{Y}$ represents how the factors should be combined to (approximately) reproduce $\mathbf{X}$. As $\mathbf{Y}$ (as well as $\mathbf{W}$) has non-negativity constraints, $\mathbf{X}$ is reproduced by using only addition (i.e., no subtraction).

Lee and Seung (1999b) show that the factors discovered by NMF from face image data are more interpretable than those discovered by PCA. NMF factors the 'parts' of a face such as an eye, a nose, and so on, while each of the factors discovered by PCA is more mixed and not recognizable as a part of a face.

They also show example results of applying NMF to a term-document matrix. We see that, because of non-negativity, the factors that NMF derives from a term-document matrix should be more interpretable than the basis vectors of an LSI or IRR subspace. However, our focus is not on obtaining interpretable factors but on obtaining the vector representations for each document whose geometrical closeness serves as similarity measurement. Lee and Seung do not report evaluation regarding similarity measurement.

8.1.3 Probabilistic Latent Semantic Indexing

Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999b; Hofmann, 1999a) has its basis in a probabilistic latent class model. Hofmann (1999b) argues for PLSI that, unlike LSI, PLSI has a solid statistical foundation, and that experiments show that it can potentially outperform LSI. However, several issues need to be worked out in practice (see below).

Let D, W, and Z be a set of documents, words, and latent classes, respectively, and let $n(d, w)$ be the count of word w in document d. For a given document collection, PLSI (locally) maximizes the log-likelihood function

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in W} n(d, w) \log P(d, w) = \sum_{d \in D} \sum_{w \in W} n(d, w) \sum_{z \in Z} P(z) P(w|z) P(d|z)\,,$$

by the tempered Expectation Maximization (TEM) algorithm. Let

$$\begin{aligned} \mathbf{U}_{[i,h]} &= P(d_i|z_h)\,, \\ \mathbf{V}_{[j,h]} &= P(w_j|z_h)\,, \\ \mathbf{\Sigma} &= \mathrm{diag}(P(z_1), \ldots, P(z_{|Z|}))\,. \end{aligned}$$

Then, the resultant model can be written as a matrix product

$$\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T ,$$

analogous to SVD.

Hofmann (1999b) reports the results of various trials applying PLSI to an IR task. The relevance of a document to a query was measured by the cosine between corresponding vectors. In one scheme, each document vector was instantiated by $P(w|d)$ with/without being weighted by the inverse document frequency (idf). A query vector was formed as in VSM. In the other scheme, a query was 'folded-in' the model, i.e., $P(z|d)$ and $P(z|q)$ were used to fill document vectors and a query vector, respectively, while each latent class was weighted by $\sum_w P(w|z) \cdot \text{idf}(w)$. The obtained cosine was weighted-averaged with the cosine obtained in VSM. PLSI models were trained at five different numbers of latent classes. Although the best performance over all these various trials outperforms LSI substantially, in practice, the issue of selecting an appropriate variation and parameters need to be worked out.

8.1.4 Vector Space-based Methods

Jiang and Littman (2000) consider that the difference between VSM, LSI, and the *generalized vector space model* (GVSM) lies in the weights of the dimensions. Let $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD of a term-document matrix $\widetilde{\mathbf{A}}$. Let $\mathbf{I}'_k$ and $\mathbf{\Sigma}'_k$ be matrices of the same shape as $\mathbf{\Sigma}$ so that:

$$\begin{aligned} \mathbf{I}'_k &= \text{diag}(\overbrace{1,\ldots,1}^{k \text{ entries}},0,\ldots,0) , \\ \mathbf{\Sigma}'_k &= \text{diag}(\sigma_1[\widetilde{\mathbf{A}}],\ldots,\sigma_k[\widetilde{\mathbf{A}}],0,\ldots,0) . \end{aligned}$$

Jiang and Littman rewrite the inner product similarity measured by these three methods by:

$$\text{Sim}_{\mathbf{\Psi}}(\mathbf{d},\mathbf{q}) = \mathbf{d}^T\mathbf{U}\mathbf{\Psi}^2\mathbf{U}^T\mathbf{q} ,$$

and

$$\mathbf{\Psi} = \begin{cases} \mathbf{\Sigma} & \text{GVSM} , \\ \mathbf{I}'_k & \text{LSI} , \\ \mathbf{I} & \text{VSM} . \end{cases}$$

Based on the assumption that the dimensions associated with smaller singular values are noise, Jiang and Littman (2000) propose *approximate dimension equalization* (ADE), which tapers off the smaller singular values instead of zeroing them out as in LSI. This is done by setting

$$\mathbf{\Psi} = \mathbf{I}'_k + \frac{1}{\sigma_k[\widetilde{\mathbf{A}}]}\mathbf{\Sigma} - \frac{1}{\sigma_k[\widetilde{\mathbf{A}}]}\mathbf{\Sigma}'_k .$$

The reported average precision results of ADE on an IR task are better than LSI on the TREC AP 1990 collection and similar to LSI on the Cranfield collection.

We note that IRR does not fit in Jiang and Littman's generalization above. IRR can be regarded

as (implicitly) altering a term-document matrix so that the topic-document distribution becomes more uniform, rather than changing the weights of dimensions.

8.2 Extensions of LSI

There have been several studies to extend LSI. Their main focuses are on reducing the computation time of SVD (the 'heart' of LSI), rather than improving the accuracy of inter-document similarity measurement.

We discuss Jiang et al. (1999a)'s study in document sampling, which is related to random-LSI (see Chapter 5) in Section 8.2.1, and introduce LSI by *random projection* (Papadimitriou et al., 2000) in Section 8.2.2. We briefly summarize other studies in Section 8.2.3.

8.2.1 Document sampling

Jiang et al. (1999a) propose a document sampling method for faster computation of LSI seeking a smaller degradation of the quality of LSI subspaces compared with uniform sampling ('folding in' (Dumais, 1993) or random-LSI in Chapter 5). They sample the documents with probability distributed proportional to the initial document vector lengths, which they call *weighted sampling*. The LSI subspace is computed only from those sampled documents. Their motivation is that longer document vectors are more influential in the SVD computation, so weighted sampling approximates the LSI subspace better than uniform sampling.

They report experiments comparing their weighted sampling against uniform sampling on both the accuracy of the approximation of the term-document matrix and on performance in an information retrieval task. Weighted sampling clearly achieved higher approximation accuracy than uniform sampling. In contrast, its average precision in document retrieval showed no clear improvement over the uniform sampling.

From our perspective, interestingly, Jiang et al.'s results provide experimental evidence that *a subspace that approximates the term-document matrix more accurately does not necessarily provide better document representations*; in other words, the fact that the LSI subspace approximates the term-document matrix optimally does not mean that the LSI subspace is the best subspace for document representation.

8.2.2 Random projection

Papadimitriou et al. (2000) studied *LSI by random projection* for faster computation. Let k be the number of topics underlying the corpus. The method projects the initial term-document matrix onto a random subspace of dimensionality l ($l > k$), and applies rank $O(k)$ LSI. If $l \ll m$, the computation time for LSI is greatly reduced. Papadimitriou et al. (2000) mathematically show how well the thus created $2k$-dimensional subspace preserves the term-document matrix in comparison to the proper k-dimensional LSI subspace.

However, their study is of theoretical interest rather than of practical importance. Let m and n

be the number of terms and documents, respectively. They show that when the dimensionality l of a random subspace satisfies

$$24 \log m < l < \sqrt{m}, \tag{8.1}$$

$$l > c\frac{\log n}{\epsilon^2} \quad \text{for a sufficiently large constant } c, \tag{8.2}$$

LSI by random projection approximates the term-document matrix $\widetilde{\mathbf{A}}$ with accuracy (as measured by the F-norm of residuals) lower than that of the $l/2$-dimensional LSI subspace by at most $2\epsilon||\widetilde{\mathbf{A}}||_{\mathrm{F}}$. For (8.1) to be satisfied, the number of terms m must be very large; for instance, $m = 180000$ allows some l to satisfy (8.1), but $m = 160000$ does not, as $24 \log 160000 > 400 = \sqrt{160000}$. Thus, their results are applicable only to document collections having a large number of unique terms, and therefore, presumably a large number of documents. Now suppose that $m = 180000$ and $n = 100000$. If we set $\epsilon = 0.01$, which ensures that degradation of accuracy is at most 2%, then according to (8.2) l has to satisfy

$$l > c(\log 100000)/0.01^2 > 166096c.$$

Although the authors did not specify how large c should be, we see that to obtain good approximation ratio in practice, the dimensionality l to be chosen may not be far smaller than the original dimensionality m. No experimental results are reported.

8.2.3 Other extensions of LSI

The studies we discuss in this section either have no evaluation on the quality of document representation or report performance no better than LSI.

SVD-updating In a practical information retrieval system, the document corpus could change frequently. SVD-updating schemes for when the corpus changes have been proposed as an alternative to recomputing SVD for the entire corpus (Berry et al., 1995b; Witter and Berry, 1998; Zha et al., 1998; Zha and Simon, 1999; Zha and Zhang, 1999). This issue is essentially out of scope of our work.

ULV decomposition Berry and Fierro (1996) propose to approximate the LSI subspace by a ULV subspace obtained by the *ULV decomposition* of a term-document matrix. The *ULV decomposition* is denoted by

$$\mathbf{X} = \mathbf{U} \begin{bmatrix} \mathbf{L} \\ \mathbf{0} \end{bmatrix} \mathbf{V}^T,$$

so that the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal,

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_h & \mathbf{0} \\ \mathbf{H} & \mathbf{E} \end{bmatrix},$$

$\mathbf{L}_h \in \Re^{h \times h}$ is a lower triangle matrix whose singular values approximate the first h singular values of $\mathbf{X}$, and $||\begin{bmatrix} \mathbf{H} & \mathbf{E} \end{bmatrix}||_2$ depends on the $(h+1)$th singular value of $\mathbf{X}$. Berry and Fierro (1996) demonstrate that the ULV decomposition is computationally less expensive than SVD, but do

not evaluate the quality of the document representation.

Semidiscrete decomposition The *semidiscrete decomposition* (SDD) was originally developed for image compression (O'Leary and Peleg, 1983). It decomposes a matrix into the sum of rank-one outer products $\sum_{i=1}^{h} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ (as in SVD) with the constraints that the elements of $\mathbf{u}_i$ and $\mathbf{v}_i$ are either -1, 0, or 1. According to Kolda and O'Leary (1998) (see also Kolda and O'Leary (1996a); Kolda and O'Leary (1996b)), the average precision performance on an information retrieval task is not significantly degraded (or improved) by replacing SVD of LSI by SDD when the parameters are appropriately chosen. Although the computation of SDD is more expensive than SVD, SDD saves the storage and query time because of its discreteness.

Survey of extensions of LSI Berry et al. (1999) gives a general overview of SVD-updating, downdating (removal of documents and terms), and SDD. They also suggest the possibility of using QR factorization to create a document subspace; however no evaluation is reported.

Chapter 9

Conclusion

In this work we have studied methods to construct a vector space in which the comparison of vector directions serves as a measurement of semantic similarity between documents. Our starting point is a model in which a document collection encompasses 'topics', which are (human-interpretable) hidden constituent concepts. Throughout this thesis we have considered the uniformity of the document distributions over topics. Starting from our finding that high non-uniformity has a negative effect on LSI's performance, we have investigated how to counteract the hidden non-uniformity underlying document collections.

9.1 Main results

We have provided a theoretical analysis of LSI based on a formal model for framing the question of how well a subspace projection-based method uncovers hidden true similarities between documents. We show a precise relation between LSI's performance and both the uniformity of the underlying topic-document distribution and the quality of the term-document matrix given as input. Furthermore, we analyze a relationship between LSI's performance and the dimensionality of an LSI subspace, showing that the best dimensionality is loosely related to the number of topics underlying the document collection and the proportion of the term-document matrix left out from the LSI subspace. In particular, if LSI greatly outperforms VSM, the dimensionality should be around the number of topics.

Based on this analysis, we have proposed an alternative method, IRR, which attempts to outperform LSI by effectively lowering the non-uniformity without prior knowledge of topic-document correspondences. IRR compensates for the non-uniformity of topic-document distributions by rescaling residuals (derives from a term-document matrix) upon the computation of every new dimension of the subspace, which amplifies the effect of documents predicted to be on less dominant topics. We also have proposed a method, called AUTO-SCALE, to automatically determine the scaling factor (which adjusts the degree of rescaling for compensation) based on a prediction that higher non-uniformity requires larger compensation, therefore, a larger scaling factor.

We have presented performance measurements both on document sets in which the topic-document distributions were carefully controlled, and on unrestricted datasets as would be found

in application settings. The results confirm our theoretical predictions regarding the performance dependence of LSI on the uniformity of the topic-document distributions, the dependence of the optimal scaling factor on the uniformity, and the effectiveness of IRR.

We have also proposed SP-IRR, which augments IRR by document sampling, for faster computation on larger document sets. Each time a new dimension of the subspace is computed (by IRR), we resample documents in favor of documents so far poorly represented in the subspace. By doing so, we seek to reduce computation time and to compensate for the non-uniformity simultaneously. Our expectation was that compared with random-IRR (which samples documents randomly once at the beginning as commonly done with LSI), SP-IRR should be more robust when a relatively small number of documents are sampled from highly non-uniform distributions. Our performance measurements, both on document sets with controlled distributions and with unrestricted distributions in application settings, have confirmed our expectations.

Finally, we have presented an application of IRR, a multi-document summarization system based on three premises: coherent themes can be identified reliably; highly representative themes can function as multi-document summary surrogates; and effective end-use of such themes can be facilitated by a visualization environment which clarifies the relationship between themes and documents. IRR works as a uniform framework from which we can find such themes and effectively present them to users.

9.2 Future directions

As commonly done, we used single- or multi-word terms to create term-document matrices in our performance evaluation. This means that we took 'a bag of words' approach, which does not take account of word order or syntactic information. However, we note that 'a bag of words' is not a crucial constituent of our model or methods, since 'terms' do not have to be words. The essence of our approach is to seek to uncover true relations between objects (e.g., documents) corresponding to their correlation with hidden objects (e.g., topics) by analyzing given relationships between two types of objects (e.g., documents and terms). We are interested in making use of syntactic information in our methods, for instance, by using subject-verb pairs or verb-object pairs as terms.

On the other hand, the fact that our approach does not rely on linguistic information suggests that the usage of our methods is not restricted to texts. We are interested in applying our methods to non-text cases such as link structure analysis.

Appendix A

Theorems and Definitions from Previous Studies

The proofs of our theorems are based on the following theorems.

A.1 Perturbation of singular values for symmetric matrices

Theorem A.1.1 *Let* $\mathbf{X}$ *and* $\mathbf{F}$ *be symmetric matrices in* $\Re^{r\times r}$*, and let* $\mathbf{X}' - \mathbf{X} = \mathbf{F}$.
Then $|\sigma_i[\mathbf{X}'] - \sigma_i[\mathbf{X}]| \leq ||\mathbf{F}||_2$ *for any* i.
(Corollary 8.6.2, pp.449 in Golub and Van Loan (1996))
Furthermore, $\sqrt{\sum_{i=1}^{n}(\sigma_i[\mathbf{X}'] - \sigma_i[\mathbf{X}])^2} \leq ||\mathbf{F}||_F$.
(Theorem 8.6.4, pp.450 in Golub and Van Loan (1996))

A.2 Canonical angle between subspaces

We use the *canonical angle* (Davis and Kahan, 1970; Stewart, 1973) to measure the difference between two subspaces of the same dimensionality.

Let $\mathcal{X}$ and $\mathcal{X}'$ be h-dimensional subspaces of $\Re^r$, and let the columns of $\mathbf{X}$, $\mathbf{X}_\perp$, $\mathbf{X}'$, and $\mathbf{X}'_\perp$ form the orthonormal bases of $\mathcal{X}$, $\mathcal{X}^\perp$, $\mathcal{X}'$, and $\mathcal{X}'^\perp$, respectively. Let $q = \min(h, r-h)$.
Then, we have $0 \leq \sigma_i[\mathbf{X}_\perp^T\mathbf{X}'] \leq 1$, so we set $\theta_i \in [0, \pi/2]$ for $1 \leq i \leq q$ so that

$$\sin\theta_i = \sigma_i[\mathbf{X}_\perp^T\mathbf{X}'] \,.$$

We define[1] the canonical angle (matrix) between $\mathcal{X}$ and $\mathcal{X}'$ to be

$$\mathbf{\Theta}(\mathcal{X}, \mathcal{X}') \stackrel{\text{def}}{=} \text{diag}(\theta_1, \theta_2, \ldots, \theta_q) \,.$$

Note that $\cos(\pi/2) = 0$, so the tangent can not be defined for $\pi/2$, as $\tan\theta = \sin\theta/\cos\theta$.

[1] Our definition of canonical angle is a simplification of the definition found in Stewart (1973) and Stewart and Sun (1990) in which each θ_i is called a canonical angle.

If $\theta_i \neq \pi/2$ for any i, the tangent between $\mathcal{X}$ and $\mathcal{X}'$ is measured by

$$||\tan(\mathbf{\Theta}(\mathcal{X}, \mathcal{X}'))|| = ||\mathrm{diag}(\tan\theta_1, \tan\theta_2, \ldots, \tan\theta_q)||\,,$$

using an appropriate matrix norm $||\cdot||$.

In particular, we let $\tan(\mathcal{X}, \mathcal{X}')$ denote the 2-norm of the tangent between $\mathcal{X}$ and $\mathcal{X}'$, that is,

$$\tan(\mathcal{X}, \mathcal{X}') = ||\tan(\mathbf{\Theta}(\mathcal{X}, \mathcal{X}'))||_2\,.$$

A.3 Tangent theorem

To analyze LSI, we use the following theorem to quantify the tangent of subspaces (Davis and Kahan, 1970); see also Theorem 3.10, pp.253 in Stewart and Sun (1990)), which we simplify and change the notation for clarity.

Theorem A.3.1 *Let* $\mathbf{W} \in \Re^{r\times r}$ *be a symmetric matrix, and let* $[\,\widetilde{\mathbf{X}} \quad \widetilde{\mathbf{X}}_\perp\,]$ *be an orthogonal matrix, with* $\widetilde{\mathbf{X}} \in \Re^{r\times p}$*, so that* range$(\widetilde{\mathbf{X}})$ *forms an invariant subspace of* $\mathbf{W}$ *(i.e. if* $\mathbf{z} \in$ range$(\widetilde{\mathbf{X}})$*, then* $\mathbf{Wz} \in$ range$(\widetilde{\mathbf{X}})$*). For any matrix* $\mathbf{X} \in \Re^{r\times p}$ *with orthonormal columns, we define the* residual *matrix* $\mathbf{R}$ *of* $\mathbf{X}$ *as* $\mathbf{R} = \mathbf{WX} - (\mathbf{XX}^T)\mathbf{WX}$*. Suppose that the eigenvalues of* $\mathbf{X}^T\mathbf{WX}$ *lie in the range* $[\alpha, \beta]$*, and that there exists* $\delta > 0$ *such that the eigenvalues of* $\widetilde{\mathbf{X}}_\perp^T\mathbf{W}\widetilde{\mathbf{X}}_\perp$ *either all lie in the interval* $(-\infty, \alpha - \delta]$ *or are all in* $[\beta + \delta, \infty)$*. Then*

$$\tan(\mathrm{range}(\widetilde{\mathbf{X}}), \mathrm{range}(\mathbf{X})) \leq \frac{||\mathbf{R}||_2}{\delta}\,.$$

A.4 Relationship between tangent and orthonormal bases

Theorem A.4.1 *Suppose that the columns of* $\mathbf{Z}$ *and* $\mathbf{Z}'$ *form orthonormal bases of the same subspace (i.e.,* range$(\mathbf{Z}) =$ range$(\mathbf{Z}')$*). Set* $\mathbf{X}$*,* $\mathbf{Y}$*,* $\mathbf{X}'$*, and* $\mathbf{Y}'$ *so that*

$$\begin{aligned} \mathbf{Z} &= [\,\mathbf{X} \quad \mathbf{Y}\,]\,, \\ \mathbf{Z}' &= [\,\mathbf{X}' \quad \mathbf{Y}'\,]\,, \end{aligned}$$

and that $\mathbf{X}$ *and* $\mathbf{X}'$ *have the same shape.*
If the tangent between range$(\mathbf{X})$ *and* range$(\mathbf{X}')$ *can be defined, then there exists unique* $\mathbf{P}$ *such that*[2]

$$\begin{aligned} \mathbf{X}' &= (\mathbf{X} + \mathbf{YP})(\mathbf{I} + \mathbf{P}^T\mathbf{P})^{-1/2}\,, \\ \mathbf{Y}' &= (\mathbf{Y} - \mathbf{XP}^T)(\mathbf{I} + \mathbf{PP}^T)^{-1/2}\,, \\ ||\mathbf{P}||_2 &= ||\tan(\mathbf{\Theta}(\mathcal{X}, \mathcal{X}'))||_2 = \tan(\mathcal{X}, \mathcal{X}')\,, \\ ||\mathbf{P}||_\mathrm{F} &= ||\tan(\mathbf{\Theta}(\mathcal{X}, \mathcal{X}'))||_\mathrm{F}\,. \end{aligned}$$

[2] $\mathbf{X}^{-1/2}$ *denotes a matrix such that its square is the inverse of* $\mathbf{X}$.

(This is a modification of Lemma 3.11, pp.255 in Stewart and Sun (1990).)

Appendix B

Proofs

B.1 Definitions and notational conventions

We introduce the definitions and notational conventions that we use throughout this appendix.

We define $\mathbf{T} \in \Re^{k \times n}$, so that

$$\mathbf{T}_{[i,j]} = \mathrm{rel}(i,j)\,,$$

i.e., topic-document matrix. Note that $\mathbf{T}^T\mathbf{T} = \mathbf{S}$.

For clarity we let

$$\begin{aligned} \mathbf{A} &= \mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})\,, \\ \bar{\mathbf{A}} &= \mathbf{P}_{\mathcal{X}_{opt}^{\perp}}(\widetilde{\mathbf{A}})\,. \end{aligned}$$

Note that $\widetilde{\mathbf{A}} = \mathbf{A} + \bar{\mathbf{A}}$ and $\mathbf{A}^T\bar{\mathbf{A}} = \mathbf{0}$.
Let

$$\begin{aligned} \mathbf{A} &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\,, \\ \bar{\mathbf{A}} &= \bar{\mathbf{U}}\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^T\,, \end{aligned}$$

be the SVD; note that $\mathrm{range}(\mathbf{U}) = \mathcal{X}_{opt}$ and the columns of $[\,\mathbf{U} \quad \bar{\mathbf{U}}\,]$ form an orthonormal basis of $\mathrm{range}(\widetilde{\mathbf{A}})$.

Let $\widetilde{\mathbf{u}}_i$ be the ith left singular vector of $\widetilde{\mathbf{A}}$ including those associated with zero singular values. Given an LSI subspace $\mathcal{X}_{LSI}$, let $h = \dim(\mathcal{X}_{LSI})$. We set

$$\begin{aligned} \widetilde{\mathbf{U}}_{\mathrm{LSI}} &= [\,\widetilde{\mathbf{u}}_1 \quad \cdots \quad \widetilde{\mathbf{u}}_h\,]\,, \\ \widetilde{\mathbf{U}}_{\overline{\mathrm{LSI}}} &= [\,\widetilde{\mathbf{u}}_{h+1} \quad \cdots \quad \widetilde{\mathbf{u}}_{\mathrm{rank}(\widetilde{\mathbf{A}})}\,]\,, \\ \widetilde{\mathbf{U}}_{\mathrm{LSI}^{\perp}} &= [\,\widetilde{\mathbf{u}}_{h+1} \quad \cdots \quad \widetilde{\mathbf{u}}_m\,]\,, \end{aligned}$$

so that $\mathrm{range}(\widetilde{\mathbf{U}}_{\mathrm{LSI}}) = \mathcal{X}_{LSI}$, the columns of $[\,\widetilde{\mathbf{U}}_{\mathrm{LSI}} \quad \widetilde{\mathbf{U}}_{\overline{\mathrm{LSI}}}\,]$ form an orthonormal basis of $\mathrm{range}(\widetilde{\mathbf{A}})$, and $[\,\widetilde{\mathbf{U}}_{\mathrm{LSI}} \quad \widetilde{\mathbf{U}}_{\mathrm{LSI}^{\perp}}\,]$ is orthogonal.

For convenience, when we refer to $\sigma_i[\mathbf{D}]$ for $i > \text{rank}(\mathbf{D})$, it is understood to be zero.

B.2 Proof of Theorem 3.3.1

Theorem 3.3.1 *There exist* $\dot{\mathcal{T}}_1 \geq \ldots \geq \dot{\mathcal{T}}_n \geq 0$ *such that* $\sigma_i[\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})] \stackrel{opt}{=} \dot{\mathcal{T}}_i$ *and* $\dot{\mathcal{T}}_i \stackrel{\mu}{=} \mathcal{T}_i$. *In particular, if each document is relevant to exactly one topic, then* $\mu = 0$, *so* $\sigma_i[\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})] \stackrel{opt}{=} \mathcal{T}_i$.

Proof We set $\mathbf{A}$ and $\mathbf{T}$ as in Appendix B.1, and we set $\dot{\mathcal{T}}_i = \sigma_i[\mathbf{T}]$.
By definition we have

$$\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{opt}) = \mathbf{S} - \mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})^T \mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}}) = \mathbf{T}^T\mathbf{T} - \mathbf{A}^T\mathbf{A}\,,$$

and we have

$$\sigma_i[-\mathbf{A}^T\mathbf{A}] = \sigma_i[\mathbf{A}^T\mathbf{A}] = \sigma_i[\mathbf{A}]^2\,, \qquad \sigma_i[\mathbf{T}^T\mathbf{T}] = \sigma_i[\mathbf{T}]^2 = \dot{\mathcal{T}}_i^2\,.$$

It immediately follows from Theorem A.1.1 that

$$\begin{aligned} |\sigma_i[\mathbf{A}]^2 - \dot{\mathcal{T}}_i^2| &\leq ||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{opt})||_2 = E^{max}_{\mathcal{X}_{opt}} \quad \text{for } 1 \leq i \leq n\,, \\ \sqrt{\sum_{i=1}^{n} \left(\sigma_i[\mathbf{A}]^2 - \dot{\mathcal{T}}_i^2\right)^2 / n} &\leq ||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}\mathcal{X}_{opt}||_{\mathrm{F}}/\sqrt{n} = E^{avg}_{\mathcal{X}_{opt}}\,, \end{aligned}$$

i.e. $\sigma_i[\mathbf{A}] \stackrel{opt}{=} \dot{\mathcal{T}}_i$ using our notation. Now it suffices to show how $\dot{\mathcal{T}}_i$ relates to $\mathcal{T}_i$.

Let

$$\mathbf{F} = \mathbf{T}\mathbf{T}^T - \text{diag}(\mathcal{T}_1^2, \ldots, \mathcal{T}_k^2)\,.$$

Observing that the diagonal entries of $\mathbf{F}$ are zero, we have $||\mathbf{F}||_{\mathrm{F}} = \mu$. As

$$\sigma_i[\mathbf{T}\mathbf{T}^T] = \sigma_i[\mathbf{T}]^2 = \dot{\mathcal{T}}_i^2\,, \qquad \sigma_i[\text{diag}(\mathcal{T}_1^2, \ldots, \mathcal{T}_k^2)] = \mathcal{T}_i^2\,,$$

it follows from Theorem A.1.1 that

$$\sqrt{\sum_{i=1}^{k} \left(\dot{\mathcal{T}}_i^2 - \mathcal{T}_i^2\right)^2} \leq ||\mathbf{F}||_{\mathrm{F}} = \mu\,.$$

Since $\dot{\mathcal{T}}_i = \mathcal{T}_i = 0$ for $k < i \leq n$, we obtain $\dot{\mathcal{T}}_i \stackrel{\mu}{=} \mathcal{T}_i$, which completes the proof. ∎

B.3 Proof of Theorem 3.3.2

Theorem 3.3.2 *We have* $\sigma_i[\mathbf{P}_{\mathcal{X}_{opt}^{\perp}}(\widetilde{\mathbf{A}})] \stackrel{opt}{=} \sqrt{\sigma_i[\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\mathrm{VSM}})]}$.
In particular, $\sqrt{\sigma_1[\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\mathrm{VSM}})]} = \varepsilon^{max}_{\mathrm{VSM}}$.

Proof We set $\mathbf{A}$ and $\bar{\mathbf{A}}$ as in Appendix B.1.
As $\mathbf{A}^T\bar{\mathbf{A}} = \bar{\mathbf{A}}^T\mathbf{A} = \mathbf{0}$, we have $\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}} = \mathbf{A}^T\mathbf{A} + \bar{\mathbf{A}}^T\bar{\mathbf{A}}$. It follows that

$$\begin{aligned}\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\mathrm{VSM}}) &= \mathbf{S} - \widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}} = (\mathbf{S} - \mathbf{A}^T\mathbf{A}) - \bar{\mathbf{A}}^T\bar{\mathbf{A}} \\ &= \mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{opt}) - \bar{\mathbf{A}}^T\bar{\mathbf{A}}\end{aligned}$$

That is, $\bar{\mathbf{A}}^T\bar{\mathbf{A}} - (-\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\mathrm{VSM}})) = \mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{opt})$. It follows from Theorem A.1.1 that

$$\begin{aligned}|\sigma_i[\bar{\mathbf{A}}]^2 - \sigma_i[\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\mathrm{VSM}})]| &\le ||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{opt})||_2 = E^{max}_{\mathcal{X}_{opt}} \quad \text{for } 1 \le i \le n\,, \\ \sqrt{\sum_{i=1}^{n}\Big(\sigma_i[\bar{\mathbf{A}}]^2 - \sigma_i[\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\mathrm{VSM}})]\Big)^2/n} &\le ||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{opt})||_{\mathrm{F}}/\sqrt{n} = E^{avg}_{\mathcal{X}_{opt}}\,,\end{aligned}$$

i.e. $\sigma_i[\bar{\mathbf{A}}] \overset{opt}{=} \sqrt{\sigma_i[\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\mathrm{VSM}})]}$.
In particular, $\sqrt{\sigma_1[\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\mathrm{VSM}})]} = \sqrt{||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\mathrm{VSM}})||_2} = \sqrt{E^{max}_{\mathcal{X}_{\mathrm{VSM}}}} = \varepsilon^{max}_{\mathrm{VSM}}$. ∎

B.4 Proof of Theorem 3.3.3

Theorem 3.3.3 *Let $h = \dim(\mathcal{X}_{opt})$. To facilitate presentation let $\hat{\mathcal{T}}_{max} = \sigma_1[\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})]$, $\hat{\mathcal{T}}_{min} = \sigma_h[\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})]$, and $\hat{\varepsilon}^{max}_{\mathrm{VSM}} = \sigma_1[\mathbf{P}_{\mathcal{X}^{\perp}_{opt}}(\widetilde{\mathbf{A}})]$. (Recall that $\mathcal{T}_i$ and $\varepsilon^{max}_{\mathrm{VSM}}$ are approximated by $\sigma_i[\mathbf{P}_{\mathcal{X}_{opt}}(\widetilde{\mathbf{A}})]$ and $\sigma_1[\mathbf{P}_{\mathcal{X}^{\perp}_{opt}}(\widetilde{\mathbf{A}})]$, respectively.) Let $\mathcal{X}_{LSI}$ be the h-dimensional LSI subspace associated with the first h singular values of $\widetilde{\mathbf{A}}$.*
If $\hat{\mathcal{T}}_{min} > \hat{\varepsilon}^{max}_{\mathrm{VSM}}$, then we have

$$\tan(\mathcal{X}_{\mathrm{LSI}}, \mathcal{X}_{\mathrm{opt}}) \le \frac{\hat{\mathcal{T}}_{max}}{\hat{\mathcal{T}}_{min}} \cdot \frac{\hat{\varepsilon}^{max}_{\mathrm{VSM}}/\hat{\mathcal{T}}_{min}}{1 - (\hat{\varepsilon}^{max}_{\mathrm{VSM}}/\hat{\mathcal{T}}_{min})^2}$$

where the tangent function tan *measures the distance between subspaces (see Appendix A.2 for the definition).*

Proof The main idea is to apply Theorem A.3.1 by choosing $\widetilde{\mathbf{X}}$ and $\mathbf{X}$ so that $\mathrm{range}(\widetilde{\mathbf{X}}) = \mathcal{X}_{LSI}$ and $\mathrm{range}(\mathbf{X}) = \mathcal{X}_{opt}$, and setting $\mathbf{W} = \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^T$, for which $\mathcal{X}_{LSI}$ is an invariant subspace.

We set $\mathbf{A}$ and $\bar{\mathbf{A}}$ as in Appendix B.1. Note that $\hat{\mathcal{T}}_{max} = \sigma_1[\mathbf{A}]$, $\hat{\mathcal{T}}_{min} = \sigma_h[\mathbf{A}]$, and $\hat{\varepsilon}^{max}_{\mathrm{VSM}} = \sigma_1[\bar{\mathbf{A}}]$.

First, we relate $\sigma_{h+1}[\widetilde{\mathbf{A}}]$ to $\hat{\varepsilon}^{max}_{\mathrm{VSM}}$. We have $\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}} - \mathbf{A}^T\mathbf{A} = \bar{\mathbf{A}}^T\bar{\mathbf{A}}$. It follows from Theorem A.1.1 that

$$|\sigma_{h+1}[\widetilde{\mathbf{A}}]^2 - \sigma_{h+1}[\mathbf{A}]^2| \le ||\bar{\mathbf{A}}^T\bar{\mathbf{A}}||_2\,.$$

As $\mathrm{rank}(\mathbf{A}) = h$ by assumption, we have $\sigma_{h+1}[\mathbf{A}] = 0$; therefore,

$$\sigma_{h+1}[\widetilde{\mathbf{A}}]^2 \le ||\bar{\mathbf{A}}^T\bar{\mathbf{A}}||_2 = \sigma_1[\bar{\mathbf{A}}]^2 = (\hat{\varepsilon}^{max}_{\mathrm{VSM}})^2\,. \tag{B.1}$$

We set $\mathbf{U}$, $\widetilde{\mathbf{U}}_{\mathrm{LSI}}$, and $\widetilde{\mathbf{U}}_{\mathrm{LSI}^\perp}$ as in Appendix B.1. Then, we have $\mathrm{range}(\mathbf{U}) = \mathcal{X}_{opt}$, $\mathrm{range}(\widetilde{\mathbf{U}}_{\mathrm{LSI}}) = \mathcal{X}_{LSI}$, and $[\,\widetilde{\mathbf{U}}_{\mathrm{LSI}} \quad \widetilde{\mathbf{U}}_{\mathrm{LSI}^\perp}\,]$ is orthogonal.

We apply Theorem A.3.1 by letting

$$\mathbf{W} = \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^T\,, \quad \widetilde{\mathbf{X}} = \widetilde{\mathbf{U}}_{\mathrm{LSI}}\,, \quad \widetilde{\mathbf{X}}_\perp = \widetilde{\mathbf{U}}_{\mathrm{LSI}^\perp}\,, \quad \mathbf{X} = \mathbf{U}\,.$$

We have:

$$\begin{aligned}
\widetilde{\mathbf{X}}_\perp^T\mathbf{W}\widetilde{\mathbf{X}}_\perp &= \mathrm{diag}(\sigma_{h+1}[\widetilde{\mathbf{A}}]^2, \cdots, \sigma_n[\widetilde{\mathbf{A}}]^2, 0, \cdots, 0)\,, \\
\mathbf{X}^T\mathbf{W}\mathbf{X} &= \mathbf{U}^T\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^T\mathbf{U} = \mathbf{U}^T(\mathbf{A}+\bar{\mathbf{A}})(\mathbf{A}^T+\bar{\mathbf{A}}^T)\mathbf{U} = \mathbf{U}^T\mathbf{A}\mathbf{A}^T\mathbf{U} \\
&= \mathrm{diag}(\sigma_1[\mathbf{A}]^2, \ldots, \sigma_h[\mathbf{A}]^2)\,.
\end{aligned}$$

Since the eigenvalues of $\widetilde{\mathbf{X}}_\perp^T\mathbf{W}\widetilde{\mathbf{X}}_\perp$ lie in the range $[0, \sigma_{h+1}[\widetilde{\mathbf{A}}]^2]$ while the eigenvalues of $\mathbf{X}^T\mathbf{W}\mathbf{X}$ are in $[\sigma_h[\mathbf{A}]^2, \sigma_1[\mathbf{A}]^2]$, we set $\delta = \sigma_h[\mathbf{A}]^2 - \sigma_{h+1}[\widetilde{\mathbf{A}}]^2$. By (B.1) and the assumption, we have $\delta \geq \hat{\mathcal{T}}_{min}^2 - (\hat{\varepsilon}_{\mathrm{VSM}}^{max})^2 > 0$.

We have

$$\begin{aligned}
\mathbf{R} &= \mathbf{W}\mathbf{X} - \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X}) \\
&= (\mathbf{A}+\bar{\mathbf{A}})(\mathbf{A}^T+\bar{\mathbf{A}}^T)\mathbf{U} - \mathbf{U}\,\mathrm{diag}(\sigma_1[\mathbf{A}]^2, \ldots, \sigma_h[\mathbf{A}]^2) \\
&= (\mathbf{A}+\bar{\mathbf{A}})\mathbf{A}^T\mathbf{U} - \mathbf{A}\mathbf{A}^T\mathbf{U} \\
&= \bar{\mathbf{A}}\mathbf{A}^T\mathbf{U}\,.
\end{aligned}$$

It follows that

$$||\mathbf{R}||_2 \leq ||\mathbf{A}||_2 \cdot ||\bar{\mathbf{A}}||_2 = \sigma_1[\mathbf{A}] \cdot \sigma_1[\bar{\mathbf{A}}] = \hat{\mathcal{T}}_{max} \cdot \hat{\varepsilon}_{\mathrm{VSM}}^{max}\,.$$

Hence, we obtain

$$\begin{aligned}
\tan(\mathcal{X}_{\mathrm{opt}}, \mathcal{X}_{\mathrm{LSI}}) &= \tan(\mathrm{range}(\mathbf{U}), \mathrm{range}(\widetilde{\mathbf{U}}_{\mathrm{LSI}})) \\
&\leq \frac{||\mathbf{R}||_2}{\delta} \\
&\leq \frac{\hat{\mathcal{T}}_{max} \cdot \hat{\varepsilon}_{\mathrm{VSM}}^{max}}{\hat{\mathcal{T}}_{min}^2 - (\hat{\varepsilon}_{\mathrm{VSM}}^{max})^2} \\
&= \frac{\hat{\mathcal{T}}_{max}}{\hat{\mathcal{T}}_{min}} \cdot \frac{\hat{\varepsilon}_{\mathrm{VSM}}^{max}/\hat{\mathcal{T}}_{min}}{1 - (\hat{\varepsilon}_{\mathrm{VSM}}^{max}/\hat{\mathcal{T}}_{min})^2}\,,
\end{aligned}$$

which completes the proof. ∎

B.5 Proof of Theorem 3.3.4

Theorem 3.3.4 *In the notation of Theorem 3.3.3, suppose that* $\hat{\mathcal{T}}_{min} > \hat{\varepsilon}_{\mathrm{VSM}}^{max} > 0$ *and* $\tan(\mathcal{X}_{\mathrm{LSI}}, \mathcal{X}_{\mathrm{opt}}) \neq 0$.

Then, we can construct an ω satisfying

$$\begin{aligned}\tan(\mathcal{X}_{\rm LSI},\mathcal{X}_{\rm opt}) &\geq \frac{2}{\omega+\sqrt{\omega^2+4}}\,,\\ \omega &\geq \frac{\hat{\mathcal{T}}_{min}}{\hat{\mathcal{T}}_{max}}\cdot\frac{1-(\hat{\varepsilon}^{max}_{\rm VSM}/\hat{\mathcal{T}}_{min})^2}{\hat{\varepsilon}^{max}_{\rm VSM}/\hat{\mathcal{T}}_{min}}\,.\end{aligned}$$

Observe that ω is bounded by the reciprocal of the upper bound of the tangent in Theorem 3.3.3.

Proof We use notation in Appendix B.1.
From Theorem 3.3.3 and Theorem A.4.1, there exists unique $\mathbf{P}$ such that

$$\begin{aligned}\widetilde{\mathbf{U}}_{\rm LSI} &= (\mathbf{U}+\bar{\mathbf{U}}\mathbf{P})(\mathbf{I}+\mathbf{P}^T\mathbf{P})^{-1/2}\,,\\ \widetilde{\mathbf{U}}_{\overline{\rm LSI}} &= (\bar{\mathbf{U}}-\mathbf{U}\mathbf{P}^T)(\mathbf{I}+\mathbf{P}\mathbf{P}^T)^{-1/2}\,,\\ ||\mathbf{P}||_2 &= \tan(\mathcal{X}_{\rm LSI},\mathcal{X}_{\rm opt})\,.\end{aligned}$$

Since

$$\widetilde{\mathbf{U}}^T_{\overline{\rm LSI}}\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^T\widetilde{\mathbf{U}}_{\rm LSI}=\mathbf{0}\,, \tag{B.2}$$

after some algebra, we have

$$\begin{aligned}\mathbf{0} &= \bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^T\mathbf{V}\mathbf{\Sigma}+\bar{\mathbf{\Sigma}}^2\mathbf{P}-\mathbf{P}\mathbf{\Sigma}^2-\mathbf{P}\mathbf{\Sigma}\mathbf{V}^T\bar{\mathbf{V}}\bar{\mathbf{\Sigma}}\mathbf{P}=\mathbf{0}\\ &= \bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^T\mathbf{V}\mathbf{\Sigma}+\bar{\mathbf{\Sigma}}^2\mathbf{P}-\mathbf{P}\mathbf{\Sigma}^2-\mathbf{P}(\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^T\mathbf{V}\mathbf{\Sigma})^T\mathbf{P}\,.\end{aligned} \tag{B.3}$$

Since $\mathbf{P}\neq\mathbf{0}$ and $\hat{\mathcal{T}}_{min}=\sigma_h[\mathbf{A}]>\sigma_1[\bar{\mathbf{A}}]=\hat{\varepsilon}^{max}_{\rm VSM}$ by assumption, we have $\bar{\mathbf{\Sigma}}^2\mathbf{P}-\mathbf{P}\mathbf{\Sigma}^2\neq\mathbf{0}$, and therefore, $\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^T\mathbf{V}\mathbf{\Sigma}\neq\mathbf{0}$. Let

$$\begin{aligned}\alpha &= ||\bar{\mathbf{\Sigma}}^2\mathbf{P}-\mathbf{P}\mathbf{\Sigma}^2||_2/||\mathbf{P}||_2\,,\\ \gamma &= ||\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^T\mathbf{V}\mathbf{\Sigma}||_2\,,\\ \omega &= \alpha/\gamma\,.\end{aligned}$$

From (B.3) and the definitions of γ and α, we have

$$\gamma=||\mathbf{P}\mathbf{\Sigma}^2-\bar{\mathbf{\Sigma}}^2\mathbf{P}+\mathbf{P}\mathbf{\Sigma}\mathbf{V}^T\bar{\mathbf{V}}\bar{\mathbf{\Sigma}}\mathbf{P}||_2\leq||\mathbf{P}\mathbf{\Sigma}^2-\bar{\mathbf{\Sigma}}^2\mathbf{P}||_2+\gamma||\mathbf{P}||_2^2=\alpha||\mathbf{P}||_2+\gamma||\mathbf{P}||_2^2\,.$$

Solving the inequality $\gamma\leq\alpha||\mathbf{P}||_2+\gamma||\mathbf{P}||_2^2$ for $||\mathbf{P}||_2$ (recall $||\mathbf{P}||_2>0$), we obtain

$$\tan(\mathcal{X}_{\rm LSI},\mathcal{X}_{\rm opt})=||\mathbf{P}||_2\geq\frac{-\alpha+\sqrt{\alpha^2+4\gamma^2}}{2\gamma}=\frac{-\omega+\sqrt{\omega^2+4}}{2}\,.$$

Multiplying by $(\omega+\sqrt{\omega^2+4})$ on both the numerator and denominator yields the desired result:

$$\tan(\mathcal{X}_{\rm LSI},\mathcal{X}_{\rm opt})\geq\frac{2}{\omega+\sqrt{\omega^2+4}}$$

Now we bound ω.

Since

$$||\mathbf{P}||_2 = ||(\mathbf{P}\mathbf{\Sigma}^2)(\mathbf{\Sigma}^2)^{-1}||_2 \leq ||\mathbf{P}\mathbf{\Sigma}^2||_2||(\mathbf{\Sigma}^2)^{-1}||_2 = \frac{1}{\sigma_h[\mathbf{A}]^2}||\mathbf{P}\mathbf{\Sigma}^2||_2\,,$$

we have

$$||\mathbf{P}\mathbf{\Sigma}^2||_2 \geq \sigma_h[\mathbf{A}]^2||\mathbf{P}||_2\,.$$

Hence,

$$\alpha = ||\mathbf{P}\mathbf{\Sigma}^2 - \bar{\mathbf{\Sigma}}^2\mathbf{P}||_2/||\mathbf{P}||_2 \geq (||\mathbf{P}\mathbf{\Sigma}^2||_2 - ||\bar{\mathbf{\Sigma}}^2\mathbf{P}||_2)/||\mathbf{P}||_2 \geq \sigma_h[\mathbf{A}]^2 - \sigma_1[\bar{\mathbf{A}}]^2\,.$$

As $\gamma = ||\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^T\mathbf{V}\mathbf{\Sigma}||_2 \leq \sigma_1[\mathbf{A}]\cdot\sigma_1[\bar{\mathbf{A}}]$, we obtain

$$\omega = \alpha/\gamma \geq \frac{\sigma_h[\mathbf{A}]^2 - \sigma_1[\bar{\mathbf{A}}]^2}{\sigma_1[\mathbf{A}]\cdot\sigma_1[\bar{\mathbf{A}}]} = \frac{\hat{\mathcal{T}}_{min}}{\hat{\mathcal{T}}_{max}}\cdot\frac{1-(\hat{\varepsilon}^{max}_{\mathrm{VSM}}/\hat{\mathcal{T}}_{min})^2}{\hat{\varepsilon}^{max}_{\mathrm{VSM}}/\hat{\mathcal{T}}_{min}}\,,$$

as desired. ∎

B.6 Proof of Theorem 3.3.5

Theorem 3.3.5 *In the setting of Theorem 3.3.3, if $\hat{\mathcal{T}}_{min} > \hat{\varepsilon}^{max}_{\mathrm{VSM}}$, then there exist non-negative α_1 and α_2, which are uniquely determined by $\mathcal{X}_{opt}$ and $\widetilde{\mathbf{A}}$, independently from $\mathcal{X}_{LSI}$, satisfying $E^{avg}_{\mathcal{X}_{LSI}} \leq E^{avg}_{\mathcal{X}_{opt}} + \alpha_1\tan(\mathcal{X}_{\mathrm{LSI}},\mathcal{X}_{\mathrm{opt}}) + \alpha_2\ (\tan(\mathcal{X}_{\mathrm{LSI}},\mathcal{X}_{\mathrm{opt}}))^2$.*

Proof We use notation in Appendix B.1.

As in the proof of Theorem 3.3.4, there exists unique $\mathbf{P}$ such that

$$\begin{aligned}\widetilde{\mathbf{U}}_{\mathrm{LSI}} &= (\mathbf{U}+\bar{\mathbf{U}}\mathbf{P})(\mathbf{I}+\mathbf{P}^T\mathbf{P})^{-1/2}\,,\\ ||\mathbf{P}||_2 &= \tan(\mathcal{X}_{\mathrm{LSI}},\mathcal{X}_{\mathrm{opt}})\,.\end{aligned}$$

Let $\mathbf{F} = \mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{opt}) - \mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{LSI})$. Then, since $E^{avg}_{\mathcal{X}_{LSI}} \stackrel{\mathrm{def}}{=} ||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{LSI})||_{\mathrm{F}}/\sqrt{n}$,

$$\begin{aligned}E^{avg}_{\mathcal{X}_{LSI}} &\leq \frac{||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{opt})||_{\mathrm{F}} + ||\mathbf{F}||_{\mathrm{F}}}{\sqrt{n}}\\ &= E^{avg}_{\mathcal{X}_{opt}} + \frac{||\mathbf{F}||_{\mathrm{F}}}{\sqrt{n}}\,.\end{aligned}$$

First, we bound $||\mathbf{F}||_{\mathrm{F}}$ by $||\mathbf{P}||_{\mathrm{F}}$. Let

$$\mathbf{P} = \mathbf{U}_{\mathbf{P}}\mathbf{\Sigma}_{\mathbf{P}}\mathbf{V}_{\mathbf{P}}^T$$

be the zero-padded SVD so that $\mathbf{U}_{\mathbf{P}}$ and $\mathbf{V}_{\mathbf{P}}$ are orthogonal. To simplify the notation, let

$p_i = \sigma_i[\mathbf{P}]$, and let

$$\mathbf{Q} = (\mathbf{I} + \mathbf{P}^T\mathbf{P})^{-1}\,.$$

Then,

$$\begin{aligned}
\mathbf{Q} &= (\mathbf{I} + \mathbf{P}^T\mathbf{P})^{-1} \\
&= (\mathbf{V_P}\mathbf{V_P}^T + \mathbf{V_P}\mathbf{\Sigma_P}^T\mathbf{\Sigma_P}\mathbf{V_P}^T)^{-1} \\
&= (\mathbf{V_P}(\mathbf{I} + \mathbf{\Sigma_P}^T\mathbf{\Sigma_P})\mathbf{V_P}^T)^{-1} \\
&= \mathbf{V_P}\ \mathrm{diag}\left(\frac{1}{p_1^2+1}, \ldots, \frac{1}{p_h^2+1}\right)\ \mathbf{V_P}^T,
\end{aligned}$$

and

$$||\mathbf{PQ}||_{\mathrm{F}} = \sqrt{\sum_{i=1}^{h}\left(\frac{p_i}{p_i^2+1}\right)^2} \leq \sqrt{\sum_{i=1}^{h} p_i^2} = ||\mathbf{P}||_{\mathrm{F}} \tag{B.4}$$

Similarly,

$$\begin{aligned}
\mathbf{A}^T\mathbf{U}\mathbf{Q}\mathbf{U}^T\mathbf{A} - \mathbf{A}^T\mathbf{A} &= \mathbf{V}\mathbf{\Sigma}\mathbf{V_P}\mathrm{diag}\left(\frac{1}{p_1^2+1}, \ldots, \frac{1}{p_h^2+1}\right)\mathbf{V_P}^T\mathbf{\Sigma}\mathbf{V}^T - \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T \\
&= \mathbf{V}\mathbf{\Sigma}\mathbf{V_P}(\mathrm{diag}\left(\frac{1}{p_1^2+1}, \ldots, \frac{1}{p_h^2+1}\right) - \mathbf{I})\mathbf{V_P}^T\mathbf{\Sigma}\mathbf{V}^T \\
&= -\mathbf{V}\mathbf{\Sigma}\mathbf{V_P}\ \mathrm{diag}\left(\frac{p_1^2}{p_1^2+1}, \ldots, \frac{p_h^2}{p_h^2+1}\right)\ \mathbf{V_P}^T\mathbf{\Sigma}\mathbf{V}^T
\end{aligned}$$

and therefore

$$||\mathbf{A}^T\mathbf{U}\mathbf{Q}\mathbf{U}^T\mathbf{A} - \mathbf{A}^T\mathbf{A}||_{\mathrm{F}} \leq ||\mathbf{\Sigma}||_{\mathrm{F}}^2||\mathbf{P}||_{\mathrm{F}}^2 = ||\mathbf{A}||_{\mathrm{F}}^2||\mathbf{P}||_{\mathrm{F}}^2\,. \tag{B.5}$$

We have

$$\begin{aligned}
\mathbf{F} &= (\mathbf{T}^T\mathbf{T} - \mathbf{A}^T\mathbf{A}) - (\mathbf{T}^T\mathbf{T} - \widetilde{\mathbf{A}}^T\widetilde{\mathbf{U}}_{\mathrm{LSI}}\widetilde{\mathbf{U}}_{\mathrm{LSI}}^T\widetilde{\mathbf{A}}) \\
&= \widetilde{\mathbf{A}}^T\widetilde{\mathbf{U}}_{\mathrm{LSI}}\widetilde{\mathbf{U}}_{\mathrm{LSI}}^T\widetilde{\mathbf{A}} - \mathbf{A}^T\mathbf{A} \\
&= (\mathbf{A} + \bar{\mathbf{A}})^T(\mathbf{U} + \bar{\mathbf{U}}\mathbf{P})\ \mathbf{Q}\ (\mathbf{U} + \bar{\mathbf{U}}\mathbf{P})^T(\mathbf{A} + \bar{\mathbf{A}}) - \mathbf{A}^T\mathbf{A} \\
&= (\mathbf{A}^T\mathbf{U}\mathbf{Q}\mathbf{U}^T\mathbf{A} - \mathbf{A}^T\mathbf{A}) \\
&\quad +\mathbf{A}^T\mathbf{U}\ \mathbf{Q}\mathbf{P}^T\ \bar{\mathbf{U}}^T\bar{\mathbf{A}} \\
&\quad +\bar{\mathbf{A}}^T\bar{\mathbf{U}}\ \mathbf{P}\mathbf{Q}\ \mathbf{U}^T\mathbf{A} \\
&\quad +\bar{\mathbf{A}}^T\bar{\mathbf{U}}\ \mathbf{P}\mathbf{Q}\mathbf{P}^T\ \bar{\mathbf{U}}^T\bar{\mathbf{A}}\,.
\end{aligned}$$

Using (B.4) and (B.5),

$$||\mathbf{F}||_{\mathrm{F}} \leq ||\mathbf{A}||_{\mathrm{F}}^2||\mathbf{P}||_{\mathrm{F}}^2 + 2||\mathbf{A}||_{\mathrm{F}}||\bar{\mathbf{A}}||_{\mathrm{F}}||\mathbf{P}||_{\mathrm{F}} + ||\bar{\mathbf{A}}||_{\mathrm{F}}^2||\mathbf{P}||_{\mathrm{F}}^2\,,$$

and using $||\mathbf{P}||_{\mathrm{F}} \leq \sqrt{n}||\mathbf{P}||_2$, we obtain

$$\begin{aligned} E_{\mathcal{X}_{LSI}}^{avg} - E_{\mathcal{X}_{opt}}^{avg} &\leq \frac{||\mathbf{F}||_{\mathrm{F}}}{\sqrt{n}} \\ &\leq \frac{1}{\sqrt{n}}\left(||\mathbf{A}||_{\mathrm{F}}^2||\mathbf{P}||_{\mathrm{F}}^2 + 2||\mathbf{A}||_{\mathrm{F}}||\bar{\mathbf{A}}||_{\mathrm{F}}||\mathbf{P}||_{\mathrm{F}} + ||\bar{\mathbf{A}}||_{\mathrm{F}}^2||\mathbf{P}||_{\mathrm{F}}^2\right) \\ &\leq \frac{1}{\sqrt{n}}\left(n\,||\mathbf{A}||_{\mathrm{F}}^2||\mathbf{P}||_2^2 + 2\sqrt{n}\,||\mathbf{A}||_{\mathrm{F}}||\bar{\mathbf{A}}||_{\mathrm{F}}||\mathbf{P}||_2 + n\,||\bar{\mathbf{A}}||_{\mathrm{F}}^2||\mathbf{P}||_2^2\right) \end{aligned}$$

Letting

$$\begin{aligned} \alpha_1 &= 2||\mathbf{A}||_{\mathrm{F}}||\bar{\mathbf{A}}||_{\mathrm{F}} \\ \alpha_2 &= \sqrt{n}\left(||\mathbf{A}||_{\mathrm{F}}^2 + ||\bar{\mathbf{A}}||_{\mathrm{F}}^2\right)\,, \end{aligned}$$

indeed, α_1 and α_2 are determined by $\mathcal{X}_{opt}$ and $\widetilde{\mathbf{A}}$ independently from $\mathcal{X}_{LSI}$, and $\alpha_1, \alpha_2 \geq 0$. ∎

Additionally,

$$||\mathbf{F}||_2 \leq ||\mathbf{P}||_2^2\left(||\mathbf{A}||_2^2 + ||\bar{\mathbf{A}}||_2^2\right) + 2||\mathbf{P}||_2||\mathbf{A}||_2||\bar{\mathbf{A}}||_2\,,$$

and therefore,

$$\begin{aligned} E_{\mathcal{X}_{LSI}}^{max} &\leq E_{\mathcal{X}_{opt}}^{max} + ||\mathbf{F}||_2 \\ &\leq E_{\mathcal{X}_{opt}}^{max} + \left(||\mathbf{A}||_2^2 + ||\bar{\mathbf{A}}||_2^2\right)\tan^2(\mathcal{X}_{\mathrm{LSI}}, \mathcal{X}_{\mathrm{opt}}) + 2||\mathbf{A}||_2||\bar{\mathbf{A}}||_2\tan(\mathcal{X}_{\mathrm{LSI}}, \mathcal{X}_{\mathrm{opt}})\,. \end{aligned}$$

B.7 Proof of Theorem 3.4.1

Theorem 3.4.1 *Let $\mathcal{X}_{LSI}$ be an x-dimensional LSI subspace associated with the first x non-zero singular values of $\widetilde{\mathbf{A}}$. Then, for some $\dot{\mathcal{T}}_1 \geq \cdots \geq \dot{\mathcal{T}}_n$ satisfying*

$$\dot{\mathcal{T}}_i \stackrel{\mu}{=} \mathcal{T}_i\,,$$

and for some $E_{\mathcal{X}_{\mathrm{VSM}}}^{(1)}, \ldots, E_{\mathcal{X}_{\mathrm{VSM}}}^{(n)}$ bounded by

$$E_{\mathcal{X}_{\mathrm{VSM}}}^{(i)} \leq E_{\mathcal{X}_{\mathrm{VSM}}}^{max}\,, \qquad \sqrt{\sum_{i=1}^{n}(E_{\mathcal{X}_{\mathrm{VSM}}}^{(i)})^2/n} \leq E_{\mathcal{X}_{\mathrm{VSM}}}^{avg}\,,$$

we have

$$E_{\mathcal{X}_{LSI}}^{avg} \geq \sqrt{\frac{\sum_{i=1}^{x}\left(E_{\mathcal{X}_{\mathrm{VSM}}}^{(i)}\right)^2 + \sum_{i=x+1}^{k}\dot{\mathcal{T}}_i^4}{n}} \quad \textit{for } x < k\,, \qquad (3.1)$$

$$E^{avg}_{\mathcal{X}_{LSI}} \geq \sqrt{\frac{\sum_{i=1}^{k}\left(E^{(i)}_{\mathcal{X}_{\rm VSM}}\right)^2 + \sum_{i=k+1}^{x}\sigma_i[\widetilde{\mathbf{A}}]^4}{n}} \quad \textit{for } x \geq k\,, \qquad (3.2)$$

$$|E^{avg}_{\mathcal{X}_{LSI}} - E^{avg}_{\mathcal{X}_{\rm VSM}}| \leq \sqrt{\frac{\sum_{i=x+1}^{n}\sigma_i[\widetilde{\mathbf{A}}]^4}{n}} \quad \textit{for any } x\,. \qquad (3.3)$$

Proof We use notation in Appendix B.1.
We start with inequality (3.3). By definition,

$$\begin{aligned}\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{LSI}) &= \mathbf{S} - \mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})^T\mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})\,,\\ \mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\rm VSM}) &= \mathbf{S} - \widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}\,.\end{aligned}$$

It follows that

$$\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{\rm VSM}) - \mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{LSI}) = \mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})^T\mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}}) - \widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}\,.$$

Therefore, we have

$$|E^{avg}_{\mathcal{X}_{\rm VSM}} - E^{avg}_{\mathcal{X}_{LSI}}| \leq \frac{||\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}} - \mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})^T\mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})||_{\rm F}}{\sqrt{n}}\,. \qquad (B.6)$$

To obtain inequality (3.3), it suffices to quantify $||\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}} - \mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})^T\mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})||_{\rm F}$ by the singular values of $\widetilde{\mathbf{A}}$.

By the definition of the projection operator, we have

$$\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}} - \mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})^T\mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}}) = \mathbf{P}_{\mathcal{X}^{\perp}_{LSI}}(\widetilde{\mathbf{A}})^T\mathbf{P}_{\mathcal{X}^{\perp}_{LSI}}(\widetilde{\mathbf{A}})$$

Hence,

$$||\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}} - \mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})^T\mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})||_{\rm F} = \sqrt{\sum_{i=x+1}^{n}\sigma_i[\widetilde{\mathbf{A}}]^4}\,,$$

which we assign to (B.6) and obtain inequality (3.3).

We let $\dot{\mathcal{T}}_i = \sigma_i[\mathbf{T}]$ and let $E^{(i)}_{\mathcal{X}_{\rm VSM}} = |\sigma_i[\mathbf{T}]^2 - \sigma_i[\widetilde{\mathbf{A}}]^2|$. Then, as shown in the proof of Theorem 3.3.1, $\dot{\mathcal{T}}_i \stackrel{\mu}{=} \mathcal{T}_i$. It follows from Theorem A.1.1 that

$$\begin{aligned}E^{(i)}_{\mathcal{X}_{\rm VSM}} &= |\dot{\mathcal{T}}_i^2 - \sigma_i[\widetilde{\mathbf{A}}]^2| \leq ||\mathbf{T}^T\mathbf{T} - \widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}||_2 = E^{max}_{\mathcal{X}_{\rm VSM}}\,,\\ \sqrt{\sum_{i=1}^{n}\left(E^{(i)}_{\mathcal{X}_{\rm VSM}}\right)^2/n} &= \sqrt{\sum_{i=1}^{n}|\dot{\mathcal{T}}i^2 - \sigma_i[\widetilde{\mathbf{A}}]^2|^2/n} \leq ||\mathbf{T}^T\mathbf{T} - \widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}||_{\rm F}/\sqrt{n} = E^{avg}_{\mathcal{X}_{\rm VSM}}\,.\end{aligned}$$

Let $\lambda_i = \sigma_i[\mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})^T\mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})]$. Applying Theorem A.1.1 to the definition of

$\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{LSI})$, we have

$$\sqrt{\frac{\sum_{i=1}^{n}\left(\dot{\mathcal{T}}_i^2-\lambda_i\right)^2}{n}} \leq \frac{||\mathbf{E}_{\mathbf{S},\widetilde{\mathbf{A}}}(\mathcal{X}_{LSI})||_{\mathrm{F}}}{\sqrt{n}} = E^{avg}_{\mathcal{X}_{LSI}}\,. \tag{B.7}$$

Observing that $\mathrm{rank}(\mathbf{T}^T\mathbf{T}) \leq k$ and $\mathrm{rank}(\mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})^T\mathbf{P}_{\mathcal{X}_{LSI}}(\widetilde{\mathbf{A}})) \leq x$, we have

$$\begin{aligned}\lambda_i &= \begin{cases}\sigma_i[\widetilde{\mathbf{A}}]^2 & \text{for } 1\leq i\leq x\\ 0 & \text{otherwise}\end{cases}\\ \dot{\mathcal{T}}i &\quad \begin{cases}\geq 0 & \text{for } 1\leq i\leq k\\ =0 & \text{otherwise}\end{cases}\end{aligned}$$

Combining the above and (B.7), we obtain (3.1) and (3.2). ∎

B.8 Proof of Theorem 6.1.1

Theorem 6.1.1 *Let p be a positive integer, and let*

$$f(\mathbf{x}) = \sum_{j=1}^{n}\left(\frac{\mathbf{x}^T}{||\mathbf{x}||_2}\mathbf{r}_j\right)^p\,.$$

Suppose that for some λ,

$$\lambda\mathbf{y} = \sum_{j=1}^{n}(\mathbf{y}^T\mathbf{r}_j)^{p-1}\mathbf{r}_j\,.$$

Then, $f(\mathbf{y})$ is a critical point, i.e., either a local maximum, a local minimum, or a saddle point.

Proof Let $\mathbf{R} = [\,\mathbf{r}_1 \;\cdots\; \mathbf{r}_n\,]$, and let $h = \mathrm{rank}(\mathbf{R})$. Let the columns of $\mathbf{B}$ form an orthonormal basis of range($\mathbf{R}$). Then, $\mathbf{x} \in$ range($\mathbf{R}$) can be expressed uniquely in the form of $\mathbf{x} = \sum_{j=1}^{h} s_j\mathbf{b}_j$, and f can be expressed as a function of $s_1,\cdots,s_h$. Let $g_t(s_1,\cdots,s_h) = \frac{\partial}{\partial s_t}f(s_1,\cdots,s_h)$ for $1 \leq t \leq h$. We have

$$\begin{aligned}\frac{\partial}{\partial s_t}\left(\left(\mathbf{x}^T\mathbf{x}\right)^{-p/2}\right) &= -p\cdot s_t\left(\mathbf{x}^T\mathbf{x}\right)^{-p/2-1}\,,\\ \frac{\partial}{\partial s_t}\left(\left(\mathbf{r}_i{}^T\mathbf{x}\right)^p\right) &= p(\mathbf{b}_t{}^T\mathbf{r}_i)\left(\mathbf{r}_i{}^T\mathbf{x}\right)^{p-1}\,.\end{aligned}$$

It follows that

$$g_t(s_1,\cdots,s_h) = \frac{\partial}{\partial s_t}f(s_1,\cdots,s_h)$$

$$
\begin{aligned}
&= \frac{\partial}{\partial s_t}\left(\left(\mathbf{x}^T\mathbf{x}\right)^{-p/2}\right)\sum_{i=1}^{n}\left(\mathbf{r}_i{}^T\mathbf{x}\right)^p + \left(\mathbf{x}^T\mathbf{x}\right)^{-p/2}\frac{\partial}{\partial s_t}\left(\sum_{i=1}^{n}\left(\mathbf{r}_i{}^T\mathbf{x}\right)^p\right)\\
&= -p\cdot s_t\left(\mathbf{x}^T\mathbf{x}\right)^{-p/2-1}\sum_{i=1}^{n}\left(\mathbf{r}_i{}^T\mathbf{x}\right)^p + \left(\mathbf{x}^T\mathbf{x}\right)^{-p/2}\sum_{i=1}^{n}p(\mathbf{b}_t{}^T\mathbf{r}_i)\left(\mathbf{r}_i{}^T\mathbf{x}\right)^{p-1}\\
&= \frac{p}{(\mathbf{x}^T\mathbf{x})^{p/2}}\cdot\frac{-s_t}{\mathbf{x}^T\mathbf{x}}\sum_{i=1}^{n}\left(\mathbf{r}_i{}^T\mathbf{x}\right)^p + \frac{p}{(\mathbf{x}^T\mathbf{x})^{p/2}}\mathbf{b}_t^T\sum_{i=1}^{n}\mathbf{r}_i\left(\mathbf{r}_i{}^T\mathbf{x}\right)^{p-1}\\
&= \frac{p}{(\mathbf{x}^T\mathbf{x})^{p/2}}\left(\mathbf{b}_t - \frac{s_t\mathbf{x}}{\mathbf{x}^T\mathbf{x}}\right)^T\sum_{i=1}^{n}\left(\mathbf{r}_i{}^T\mathbf{x}\right)^{p-1}\mathbf{r}_i
\end{aligned}
$$

Let $\mathbf{y} = \sum_{j=1}^{h} w_j\mathbf{b}_j$, so that $\lambda\mathbf{y} = \sum_{i=1}^{n}\left(\mathbf{r}_i{}^T\mathbf{y}\right)^{p-1}\mathbf{r}_i$. Then

$$
\begin{aligned}
g_t(w_1,\cdots,w_h) &= \frac{p}{(\mathbf{y}^T\mathbf{y})^{p/2}}\left(\mathbf{b}_t - \frac{w_t\mathbf{y}}{\mathbf{y}^T\mathbf{y}}\right)^T\sum_{i=1}^{n}\left(\mathbf{r}_i{}^T\mathbf{y}\right)^{p-1}\mathbf{r}_i\\
&= \frac{p}{(\mathbf{y}^T\mathbf{y})^{p/2}}\left(\mathbf{b}_t - \frac{w_t\mathbf{y}}{\mathbf{y}^T\mathbf{y}}\right)^T\lambda\mathbf{y}\\
&= \frac{p}{(\mathbf{y}^T\mathbf{y})^{p/2}}\lambda\left(\mathbf{b}_t^T\mathbf{y} - \frac{w_t\mathbf{y}^T\mathbf{y}}{\mathbf{y}^T\mathbf{y}}\right)\\
&= \frac{p}{(\mathbf{y}^T\mathbf{y})^{p/2}}\lambda(w_t - w_t)\\
&= 0\,.
\end{aligned}
$$

Since this holds for any t, $f(\mathbf{y})$ is a critical point. ∎

References

References

Rie Kubota Ando and Lillian Lee. 2000. Mostly-unsupervised statistical segmentation of Japanese: Applications to Kanji. In *Proceedings of NAACL'2000.*

Rie Kubota Ando and Lillian Lee. 2001. Iterative residual rescaling: An analysis and generalization of LSI. In *Proceedings of SIGIR'2001.*

Rie Kubota Ando, Branimir K. Boguraev, Roy J. Byrd, and Mary S. Neff. 2000. Multi-document summariztion by visualizing topical content. In *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization.*

Rie Kubota Ando. 2000. Latent semantic space: Iterative scaling improves inter-document similarity measurement. In *Proceedings of SIGIR '00.*

Yossi Azar, Amos Fiat, Anna Karlin, Frank McSherry, and Jared Saia. 2001. Spectral analysis of data. In *Proceedings of STOC 2001.*

Brian T. Bartell, Garrison W. Cottrell, and Richard K. Belew. 1992. Latent Semantic Indexing is an optimal special case of Multidimensional Scaling. In *Proceedings of SIGIR '92*, pages 161–167.

Brian T. Bartell, Garrison W. Cottrell, and Richard K. Belew. 1995. Representing documents using an explicit model of their similarities. *Journal of the American Society for Information Science*, 46(4):254–271.

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the ACL.*

Michael W. Berry and Ricardo D. Fierro. 1996. Low-rank orthogonal decompositions for information retrieval applications. *Numerical Linear Algebra with Applications*, 1(1):1–27.

Michael Berry, Theresa Do, Gavin O'Brien, Vijay Krishna, and Sowmini Varadhan. 1993. SVDPACKC (version 1.0) user's guide. In *Technical Report CS-93-194, University of Tennessee.*

Michael W. Berry, Susan T. Dumais, and Todd A. Letsche. 1995a. Computational methods for intelligent information access. In *Proceedings of the 1995 ACM/IEEE Supercomputing Conference*, pages 390–430.

Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. 1995b. The computational complexity of alternative updating approaches for an SVD-encoded indexing scheme. In *Proceedings of the Seventh SIAM Conference on Parallel Processing for Scientific Computing*, pages 39–44.

Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. 1995c. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595.

Michael W. Berry, Zlatko Drmac, and Elizabeth R. Jessup. 1999. Matrices, vector spaces, and information retrieval. *SIAM Review*, 41(2):335–362.

Michael W. Berry. 1992. Large scale sparse singular value computations. *International Journal of Supercomputer Applications*, 6(1):13–49.

Branimir Boguraev and Mary Neff. 2000. Discourse segmentation in aid of document summarization. In *Proceedings of Hawaii International Conference on System Sciences (*HICSS-33*), Minitrack on Digital Documents Understanding*, Maui, Hawaii. IEEE.

Jay Budzik and Kristian Hammond. 1999. Watson: Anticipating and contextualizing information needs. In *Proceedings of the Sixty-second Annual Meeting of the American Society for Information Science*.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

T. F. Chan. 1982. An improved algorithm for computing the singular value decomposition. *ACM Transaction, Mathmatics, Software*, 8:72–83.

Chandler Davis and W. M. Kahan. 1970. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, March.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Chris H.Q. Ding. 1999. A similarity-based probability model for Latent Semantic Indexing. In *Proceedings of SIGIR'99*, pages 58–65.

Susan T. Dumais and Jakob Nielsen. 1992. Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of SIGIR '92*, pages 233–244.

Susan T. Dumais, G. W. Furnas, Thomas K. Landauer, and S. Deerwester. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of CHI'88: Conference on Human Factors in Computing*, pages 281–285.

Susan T. Dumais. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2):229–236.

Susan T. Dumais. 1993. LSI meets TREC: A status report. In *The First Text REtrieval Conference (TREC1), National Institute of Standards and Technology Special Publication 500-207*, pages 137–152.

Susan T. Dumais. 1994. Latent semantic indexing (LSI) and TREC-2. In *The Second Text REtrieval Conference (TREC2), National Institute of Standards and Technology Special Publication 500-215*, pages 105–116.

Susan T. Dumais. 1995. Latent Semantic Indexing (LSI): TREC-3 report. In *The Third Text REtrieval Conference (TREC3), National Institute of Standards and Technology Special Publication 500-226*, pages 135–144.

Peter W. Foltz and Susan T. Dumais. 1992. Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM*, 35(12):51–60.

Peter W. Foltz, M. A. Britt, and C. A. Perfetti. 1996. Reasoning from multiple texts: An automatic analysis of readers' situation models. In *Proceedings of the 18th Annual Cognitive Science Conference*, pages 110–115.

Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. 1998a. The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25(2&3):285–307.

Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. 1998b. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25:285–307. Special Issue: Quantitative Approaches to Semantic Knowledge Representations.

Peter W. Foltz. 1990. Using Latent Semantic Indexing for information filtering. In *Proceedings of the Conference on Office Information Systems*, pages 40–47.

Peter W. Foltz. 1996. Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers*, 28(2):197–202.

Jan J. Gerbrands. 1981. On the relationships between SVD, KLT and PCA. *Pattern Recognition*, 14:375–381.

Gene H. Golub and C. Reinsch. 1970. Singular value decomposition and least squares solutions. *Numerical Mathmatics*, 14:403–420.

Gene H. Golub and Charles F. Van Loan. 1996. *Matrix Computations*. The Johns Hopkins University Press, third edition.

Richard L. Gorsuch. 1983. *Factor Analysis*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, second edition.

Matthias Hemmje, Clemens Kunkel, and Alexander Willett. 1994. LyberWorld – a visualization user interface supporting fulltext retrieval. In *Proceedings of SIGIR'94*, pages 249–260.

Thomas Hofmann. 1999a. Probabilistic Latent Semantic Analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*.

Thomas Hofmann. 1999b. Probabilistic Latent Semantic Indexing. In *Proceedings of SIGIR'99*.

Fan Jiang and Michael L. Littman. 2000. Approximate dimension equalization in vector-based information retrieval. In *Proceedings of International Conference on Machine Learning*.

Fan Jiang, Ravi Kannan, Michael L. Littman, and Santosh Vempala. 1999a. Efficient singular value decomposition via improved document sampling. In *Technical Report CS-99-5, Duke University, Department of Computer Science*.

J. Jiang, M. W. Berry, J. M. Donato, and G. Ostrouchov. 1999b. Mining consumer product data via Latent Semantic Indexing. *Intelligent Data Analysis*, 3(5):377–398.

Teuvo Kohonen. 1997. Exploration of large document collections by self-organizing maps. In *Proceedings of SCAI'97*, pages 5–7.

T. G. Kolda and D. P. O'Leary. 1996a. Large Latent Semantic Indexing via a semi-discrete matrix decomposition. In *Technical Report No. UMCP-CSD CS-TR-3713, Department of Computer Science, University of Maryland*.

T. G. Kolda and D. P. O'Leary. 1996b. A semi-discrete matrix decomposition for Latent Semantic Indexing in information retrieval. In *Technical Report No. UMCP-CSD CS-TR-3274, Department of Computer Science, University of Maryland*.

Tamara G. Kolda and Dianne P. O'Leary. 1998. A semidiscrete matrix decomposition for Latent Semantic Indexing in information retrieval. *ACM Transactions on Information Systems*, 16(4):322–346.

Robert R. Korfhage and Kai A. Olsen. 1995. Image organization using vibe a visual information browsing environment. In *Proceedings of SPIE*, volume 2606, pages 380–388.

Joseph B. Kruskal. 1978. Factor Analysis and Principal Components. In *International Encyclopedia of Statistics*. New York: Free Press.

Mikko Kurimo. 2000. Fast Latent Semantic Indexing of spoken documents by using self-organizing maps. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'2000*.

Krista Lagus, Timo Honkela, Samuel Kaski, and Teuvo Kohonen. 1996. Self-organizing maps of document collections: A new approach to interactive exploration. In *Proceedings of Second International Conference on Knowledge Discovery & Data Mining*, pages 238–243.

C. Lanczos. 1950. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45(4):255–282.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Thomas K. Landauer and Michael L. Littman. 1990. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31–38.

Thomas K. Landauer, D. Laham, Bob Rehder, and M. E. Schreiner. 1997. How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, pages 412–417.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998a. An introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

Thomas K. Landauer, Darrel Laham, and Peter Foltz. 1998b. Learning human-like knowledge by singular value decomposition: A progress report. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 45–51. MIT Press, Cambridge.

Daniel D. Lee and H. Sebastian Seung. 1999a. Learning in intelligent embedded systems. In *Proceedings of the Workshop on Embedded Systems.*

Daniel D. Lee and H. Sebastian Seung. 1999b. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Xia Lin. 1993. Map displays for information retrieval. *Information Processing & Management*, 29(1):69–81.

Inderjeet Mani and Eric Bloedorn. 1999. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):1–23.

Inderjeet Mani and Mark T. Maybury. 1999. *Advances in Automatic Text Summarization.* MIT press, Cambridge, Massachusetts.

D. O'Leary and S. Peleg. 1983. Digital image compression by outer product expansion. *IEEE Trans. Commun.*, 31:441–444.

Kai A. Olsen, Robert R. Korfhage, Kenneth M. Sochats, Michael B. Spring, and James G. Williams. 1993. Visualization of a document collection: The VIBE System. *Information Processing & Management*, 29(1):69–81.

Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. 2000. Latent Semantic Indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235.

Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Earl Rennison. 1994. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. *presented at the ACM Symposium on User Interface Software and Technology, Marina del Rey, CA, November*, pages 2–4.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York.

Patrick Schone and Daniel Jurafsky. 2000a. Knowledge-free induction of morphology using Latent Semantic Analysis. In *Proceedings of CoNLL-2000 and LLL-2000.*

Patrick Schone and Daniel Jurafsky. 2000b. Knowledge-free induction of inflectional morphologies. In *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL-2001).*

Hinrich Schütze and Craig Silverstein. 1997. Projections for efficient document clustering. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–81.

Sidney Siegel and N. J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw Hill, second edition.

Noam Slonim and Naftali Tishby. 2000. Document clustering using word clusters via the information bottleneck method. In *Proceedings of SIGIR '00*.

Ian M. Soboroff, Charles K. Nicholas, James M. Kukla, and David S. Ebert. 1998. Visualizing document authorship using n-grams and Latent Semantic Indexing. In *Proceedings of the workshop on New paradigms in information visualization and manipulation, Conference on Information and Knowledge Management*, pages 43–48.

G. W. Stewart and Ji-guang Sun. 1990. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press, San Diego.

G. W. Stewart. 1973. Error and perturbation bounds for subspaces asoociated with certain eigenvalue problems. *SIAM Reivew*, 15(4):727–764.

Roger E. Story. 1996. An explanation of the effectiveness of Latent Semantic Indexing by means of a Bayesian regression model. *Information Processing & Management*, 32(3):329–344.

Peter Wiemer-Hastings, Arthur Graesser, Derek Harter, and the Tutoring Research Group. 1998. The foundations and architecture of Autotutor. In *Proceedings of the 4th International Conference on Intelligent Tutoring Systems*, pages 334–343.

Peter Wiemer-Hastings, Katja Wiemer-Hastings, and Arthur C. Graesser. 1999. Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In *AI in Education 1999*, pages 535–542.

Peter Wiemer-Hastings. 1999. How latent is Latent Semantic Analysis? In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 932–937.

Peter Wiemer-Hastings. 2000. Adding syntactic information to LSA. In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, pages 989–993.

J. A. Wise, J.J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. 1995. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proceedings of Information Visualization*, pages 51–58.

Dian I. Witter and Michael W. Berry. 1998. Downdating the Latent Semantic Indexing model for conceptual information retrieval. *The Computer Journal*, 41(8):589–601.

Michael B. W. Wolfe, M. E. Schreiner, Bob Rehder, Darrell Laham, Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. 1998. Learning from text: Matching readers and text by Latent Semantic Analysis. *Discourse Processes*, 25:309–336.

Yiming Yang and Christopher G. Chute. 1993. An application of Least Squares Fit mapping to text information retrieval. In *Proceedings of 16th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 281–290.

Yiming Yang and Christopher G. Chute. 1994. An example-based mapping method for text classification and retrieval. *ACM Transactions on Information Systems (TOIS)*, 12(3):252–277.

Yiming Yang. 1995. Noise reduction in a statistical approach to text categorization. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95)*, pages 256–263.

Hongyuan Zha and Horst D. Simon. 1999. On updating problems in Latent Semantic Indexing. *SIAM Journal of Scientific Computing*, 21:782–791.

Hongyuan Zha and Zhenyue Zhang. 1999. On matrices with low-rank-plus-shift structures: Partial SVD and Latent Semantic Indexing. *SIAM Journal of Matrix Analysis and Applications*, 21:522–536.

Hongyuan Zha, Osni Marques, and Horst D. Simon. 1998. Large-scale SVD and subspace-based methods for information retrieval. In *Solving Irregularly Structured Problems in Parallel, Proceedings of 5th International Symposium (IRREGULAR'98)*, pages 29–42.